ADDIS ABABA UNIVERSITY

SCHOOL OF GRADUATE STUDIES

FACULTY OF INFORMATICS

DEPARTEMNT OF INFORMATION SCIENCE

APPLICATION OF DATA MINING TECHNOLOGY TO IDENTIFY SIGNIFICANT PATTERNS IN CENSUS OR SURVEY DATA: THE CASE OF 2001 CHILD LABOR SURVEY IN ETHIOPIA

A THESIS SUBMITTED TO THE SCHOOL OF GRADUATE STUDIES OF ADDIS ABAB UNIVERSITY IN PARTIAL FULFILMENT OF THE REQUIREMENTS FOR THE DEGREE OF MASTER OF SCIENCE IN INFORMATION SCIENCE

By

Helen Tefera Kidane

July 2003

DEDICATION

I would like to dedicate this paper to my father, Ato Tefera Kidane, who has been struggling and fighting for my education since I was a little girl. **My daddy congratulations!! Your dream comes true!!**

ACKNOWLEDGEMENT

I would like to thank my advisors Dr. Gashaw Kebede and Ato Shegaw Anagaw for their constructive and uninterrupted comments and guidance.

I would like to express my strong appreciation for Dr. Gashaw Kebede for his approach, treatment and help to his advisees at the time of difficulties.

The staffs of FDRE CSA were fully cooperative in giving me what ever needed information for the research, some advices on how the research could be done effectively, and also the actual data of the 2001 child labor survey in Ethiopia. They sacrificed their valuable working time in answering questions about the survey and generally about their data collection and management systems.

Miss Sophie DE CONINCK, the associate expert on child labor at International Labor Organization, gave me her unlimited support in any way she could.

Last, but not least, I would like to thank my students at Queens' College, who were supportive and patient when I missed some classes because of work overload.

Table of Content

LIST OF ABBREVIATIONS

FDRE CSA: Federal Democratic Republic of Ethiopia Central Statistics Authority
ILO: International Labor Organization
IPEC: International Program on the Elimination of Child Labor

LIST OF FIGURES

Abstract

Knowledge and understanding of a problem is always the first step in identifying effective solutions. Child labor is both a sign and cause of poverty that should be eliminated as soon as possible. In Ethiopia, there is no much statistical data on child labor practice. To fill this data gap, the FDRE, CSA carried out country wide child labor survey in 2001. This organization uses very simple statistical tools to show summary figures of different variables involved in 2001 child labor survey database. However traditional statistical methods are not good enough to discover complex relationships from large volume databases. The inefficiency of these tools necessitated the development of more powerful methods and techniques that can be used to study relationships and patterns through the large volumes of data collected for example for census and survey purposes. In developed world, government and non-government organizations which have access to censuses and surveys are making use of the relatively new and modern technology, data mining, to identify important patterns and relationships within the data that is accumulated in large database.

The application of data mining techniques to official data such as the 2001 child labor survey has great potential in supporting good public policy. This research focused on identifying relationships between attributes within the 2001 child labor survey database that can be used to clearly understand the nature of child labor problem in Ethiopia. So the goal of the data mining process in this research was identifying interesting patterns and relationships in the 2001 child labor database.

After the identification and understanding of the problem domain and the research objectives, the remaining stages of the research project focused on the following three major phases in data mining process. During the first phase, selection of the appropriate data mining tool which can be used to attain the defined data mining goal and the target dataset used in model building were the major tasks. The next phase, data cleaning and preparation, involved identifying and correcting mis-transmitted information, consolidating and combining records, transforming data from one form to another suitable for the selected data mining tool, handling missing attributes and selecting relevant attributes for generating meaningful association rules. As a final step for data preparation, the selected dataset was categorized into five classes using expectation maximization clustering algorithm implemented in knowledge studio version 3.0. A dataset of 2398 records with 63 attributes were used for clustering purpose.

Apriori is an association rule algorithm which is implemented in Weka software. In the third phase, model building and evaluation, the apriori algorithm was used to generate association rules from the clustered as well as non-clustered selected dataset. Different attributes were given to apriori in an effort to generate meaningful rules.

The results from this study were encouraging, which strengthened the hypothesis that interesting patterns can be generated from census and survey database by applying one of the data mining techniques: association rule mining.

Key words: Data mining, knowledge discovery, association rule, apriori algorithm.

Chapter 1

Introduction

1.1 Background

Child labor is, generally speaking, work for children that harms them or exploits them in some way (physically, mentally, or by blocking access to education). As discussed by IPEC 1994, it is difficult to give exact definition for child labor because it encompasses three concepts that are difficult to define: “child”, “work”, and “labor”. Although ILO and other organizations tried to set minimum conventions defining child labor, still its actual implementation is left to the government and public society of each country. The researcher would like to emphasize that the concept child labor can mean different things in different societies and at different times.

According to the ILO revised estimates of 2002, there are 211 million working children between the age of 5-14 in the world of which almost 120 million work full time (FDRE CSA, 2001). As may be expected given the prevailing economic conditions, the overwhelming majority of these are in developing countries like Africa, Asia, and Latin America (ILO, 2002). In most developing countries, children are engaged in activities, which are often exploitative or hazardous that affect their education, health and mental and physical development.

Child work is a complicated issue in a country like Ethiopia. In rural areas child work is perceived as an unavoidable or even necessary part of the child’s socialization process. Children are commonly involved in domestic chores. In times which require additional hands at work,

children are supposed to assist in manual labor such as weeding and harvesting. Even though there is an obvious lack of statistical figures on child labor in Ethiopia, there is no doubt that it is an enormous problem. Many start work as early as 5 and 6 years of age, devoid of any form of protection and under abusive and exploitative conditions (FDRE CSA, 1999).

The Ethiopian Government constitution has some provisions for the protection of children and the government has also signed for the convention on the rights of a child. To ensure healthy development of children, the government has mandated a special organization within the auspices of the Ministry of Labor and Social Affairs called "Child and Youth Affairs Organization". Similarly, the Labor Proclamation of Ethiopia (Proc.N0. 42/93) requires the Ministry of Labor and Social Affairs to legally prescribe lists of dangerous operations that are detrimental to the health of working children. However, MOLSA has been constrained by lack of information to carry out its mandate (FDRE CSA, 1999).

Information system that reveals the characteristics, magnitude, distribution, causes and consequences of economic activities of children in Ethiopia is required in order to design programs, strategies and final solutions for the problems of working children. Gaining enough knowledge on the problem area is the step to develop a solution for it. Having deep knowledge about children economic activity in Ethiopia helps to develop solutions in different ways. The devised solutions shall be more effective since they are supported by appropriate information.

As mentioned above, there is no even complete statistical information on the socio-economic activities of children in Ethiopia. This lack of information hinders intervention and other possible

measures as a solution. Yet, attempts are being made by FDRE CSA to fill this data gap by taking survey of children involved in domestic and productive activity.

Children labeled as 'working in domestic activity' are those who work domestic chores for their families without payment as opposed to those working in productive or economic activity. On the other hand, economic or productive activity is defined as work which involves the production of goods and/or services for sale or exchange and production of certain products for own consumption. According to the above general definition, economic activity covers production of goods and services intended for sale on the market, production of other goods and services such as government activities and, production and processing of primary products (agriculture, hunting, fishing, forestry and, logging and mining) for own consumption (FDRE CSA, 1999).

The first attempt of data collection of child labor is the 1999 National Labor Force Survey which contains information on the socio-demographic characteristics and economic participation and some other related information on children aged between 5-14 years. However, the major concern of this survey was on adult labor statistics and may not provide adequate data about children economic activity. The second attempt is the stand alone child labor survey of 2001. A stand alone survey is an independent survey taken at a time. The 2001 child labor survey resides in its own databases with out any connection to other databases of CSA. It contains full fledged information on socio-demographic, educational and economic activities of children aged 5-17 and of their guardians.

The data collected through 2001 child labor survey were from sampled areas of each *killil* in the country. Software developed specifically for census data, Integrated Microcomputer Processing System (IMPS), is used to manage data collected by CSA for different purposes at different time periods. The sample-based database of child labor survey contains a total of 180,000 records. However, as Raghavan, Deogun & Server (1998) indicated, although the capabilities to collect and store data in large computer databases has increased significantly, the relational database technology of today offers little functionality to process and explore data and establish a relationship or pattern among data elements that are hidden or previously unknown.

Although FDRE CSA has conducted a child labor survey and have a large database on this issue, due to lack of powerful data analysis tools and techniques, the collected data is not properly utilized to support a wide variety of decision making activities. The statistical analysis made by FDRE CSA is limited to computing and communicating summary figures for different variables. Further evaluation of stored data about socio-demographic and economic activities of children might lead to the discovery of trends and patterns hidden within the data that could give us clue on how to design plans and projects protecting working children, and to develop policies which help to eliminate the most intolerable forms of child employment.

Thus, this extensive amount of data gathered and stored in FDRE CSA databases require specialized data analysis tools and techniques. Given certain data analysis goal, it has been a common practice to either design a database application on on-line data or use a statistical (or an analytical) package on off-line data along with a domain expert to interpret the result. Yet, classical statistical tools need to be wielded by a trained statistician with a good or possibly

preconceived idea of what to look for. As data volumes grow dramatically, this type of manual data analysis is becoming completely impractical in many domains. Thus it is difficult to use statistical models to discover unanticipated and complex relationships, especially when the size of the data and the number of variables involved become larger and larger (Fayyad et. al., 1996).

New techniques and methods are required to evaluate, analyze, search and discover new patterns and relationships hidden in large database. As a result, the discipline of knowledge discovery or data mining in data bases, which deals with the study of such tools and techniques, has evolved into an important and active area of research. (Raghavan, et. al., 1998)

As Rea (2001) states data mining is the search for relationships and global patterns that exist in large databases but are hidden among the vast amount of data. These relationships represent valuable knowledge about the database and the objects in the database. According to Two Crows Corporation (TCC) 1999, the data mining process involves the major activities of understanding the problem domain, data collection, data preparation, model building, evaluation, and finally the deployment of results.

Han and Kamber also mentions that data mining tools perform data analysis and may uncover important data patterns, contributing greatly to business strategies, knowledge bases, and scientific, social and medical research.

1.1.1 Data Mining Applications

Data mining is increasingly popular because of the substantial contribution it can make. Data mining offers value across a broad spectrum of industries. In business, main data mining application areas includes marketing, finance, fraud detection, manufacturing and telecommunications (Fayyad et. al., 1996).

In marketing, the primary application is database marketing system, which analyzes customer databases to identify different customer groups and forecast their behavior. According to TCC (1999), telecommunications and credit card companies are two of the leaders in applying data mining to detect fraudulent use of their services. Insurance companies and stock exchanges are also interested in applying this technology to reduce fraud. Medical applications are another fruitful area. Data mining can be used to predict the effectiveness of surgical procedures, medical tests or medications. Companies active in the financial markets use data mining to determine market and industry characteristics as well as to predict individual company and stock performance. Retailers are making more use of data mining to decide which products to stock in particular stores as well as to assess the effectiveness of promotions. Pharmaceutical firms are mining large databases of chemical compounds and of genetic material to discover substances that might be candidates for development as agents for the treatments of disease (TCC, 1999).

1.1.1.1 Data Mining Applications on Official Data

The term official data denotes data collected in censuses and statistical surveys as well as administrative and registration records collected by government departments and local authorities. They are used to produce official statistics for the purpose of making policy decisions, and to facilitate the appreciation of economic, social, demographic, and other matters of interest to the governments, government departments, local authorities, businesses, and to the general public.

The application of data mining techniques on official data has great potential in supporting good public policy. Nevertheless, it is not straight forward and requires a challenging methodological research, which is still at an initial stage. Data mining techniques can be used to detect errors in data collection (outlier detection), cluster, classify, make prediction, and generate interesting association patterns out of census and survey databases.

There are some research works performed on the application of data mining technologies in different areas in Ethiopia. The first is the work of Gobena (2000) on the possible application of data mining techniques in support of flight revenue information system for Ethiopian Airlines. Then Henock (2002) attempted to investigate the application of data mining techniques to support customer relationship management at Ethiopian Airlines. Shegaw (2002) and Tesfaye (2002) also attempted on the application of data mining technology to predict child mortality patterns and the application of data mining technology to assess level of risk for insurance companies respectively. However, to the best of the researcher knowledge, there has not been any

data mining researches performed on detection of interesting patterns from census or survey data in Ethiopia.

1.2 Statement of the Problem and its Importance

The underlying problem that necessitated this research is the increased involvement of children in productive activities in Ethiopia. As pointed out by FDRE CSA (1999), the recent ILO experience indicated that no single intervention is sufficient to solve the problem of child labor due to complex and multiple causes. Therefore, action against child labor has to proceed at various levels over and above the design and implementation of a poverty-alleviating growth strategy. These would include: legislation and enforcement; education; community mobilization and awareness rising; and the development of preventive, protective and rehabilitative programs (FDRE CSA, 1999).

If we properly understand a problem area, then designing effective and efficient solutions for the problem would be much easier. Appropriate policies and strategies to overcome child labor can be drawn on the basis of enlightened knowledge about the existing realities of the issue at hand. Investigating the meaningful patterns and relationships between the different variables involved in child labor practice can give insight about the problem.

The availability of data on working children and their analysis on a continuous basis is particularly essential for establishing intervention programs and formulating policies for the eventual elimination of child labor. However, in Ethiopia there is not comprehensive and

adequate study, which shows the exact magnitude and situation of child labor in all economic sectors. Despite lack of information on the problem, there is no doubt that child labor is of critical problem to Ethiopia. Since child labor is both a sign and cause for poverty, the government stand is to completely eliminate the problem. For economically weak country, like Ethiopia, this ultimate goal may take quite a long time. But in the mean time, care should be taken to avoid the abusive forms of children employment and also to protect the working children (FDRE CSA, 1999).

As mentioned in section 1.1, FDRE CSA attempted to fill this data gap by taking survey on child labor in each region of the country. FDRE CSA is the only organization which is trying to give information on child labor in Ethiopia to different concerned parties. However the data analysis made by FDRE CSA is limited to computing aggregate figures and percentages for different variables such as percentage of children attending school, percentage of children engaged in some type of activity (productive or domestic), percentage of female children engaged in housekeeping activity and major contributing sector activities employing children in urban or rural areas.

The child labor database of FDRE CSA can be more efficiently utilized by applying the new and modern information technology, data mining. Data mining techniques can search for critical patterns or relationships that exist in the child labor database. It can find new unobserved and unsuspected relationships among existing attributes that can be used as input to develop strategies and policies of child abuse prevention. The output of this research also can be used as a ground for other further studies concerning child labor in Ethiopia.

To this end, this research, attempted to identify critical patterns in 2001 child labor survey by applying data mining technology.

1.3 Objectives of the Research

1.3.1General Objective

The general objective of this research is to identify important and interesting patterns from the 2001 child labor survey undertaken by FDRE CSA. Relevant patterns and relationships which may not be obvious for human analysts can be easily extracted from large databases by applying data mining technologies.

1.3.2 Specific Objectives

In order to achieve the above stated general objective, the following specific objectives are formulated.

- Conduct a through review of literature on the existing data mining techniques in general, and their application in identifying critical patterns in survey or census database.
- Select and extract the dataset required for analysis from the database of FDRE CSA.
- Identify an appropriate data mining algorithm and software that would do the main task of the research project: identifying meaningful patterns and relationships in the child labor survey database.

- Preparing the data for model building which includes adjusting inconsistent data encoding, accounting for missing values, combining attribute values, and separating target dataset into clusters;
- Build and train data mining models and select the best association rules by having domain expert opinion.
- Report results and make recommendations.

1.4 Research Methodology

The procedures followed in conducting the research are described below:

A. Literature Review

The researcher has conducted literature review to assess the major issues and concepts in the field of data mining. Various books, journals, articles and papers from the Internet have been consulted to assess the importance and applications of data mining technology in general and its application on census and survey data in particular.

B. Initial Data Collection

The initial target dataset for the research was selected from the 2001 child labor survey database of FDRE CSA. Series discussions were conducted with domain experts in the area of child labor and statistics before selecting the target dataset. As a result of these discussions, the researcher

selected two *killils*, *Affar* and *Gambella*, as target dataset for the research project. These regions were selected due to their high difference in the proportion of children engaged in productive and housekeeping activity.

C. Data Preparation

The process of data cleaning and preparation is highly dependent on the specific data mining algorithm and software chosen for the data mining task. The researcher attempted to prepare the data according to the requirements of the selected data mining software, Weka and selected data mining algorithm, apriori. Weka is multi-functional data mining software. The major data mining functions incorporated in the software are data preprocessing, classification, association, clustering and visualizing input and output. Apriori is the only association rule algorithm implemented in Weka.

After collecting the target dataset, the researcher checked the consistency of individual attribute values and types, and quantity and distribution of missing values. Since the selected data mining software, Weka, does not allow any inconsistency among attribute values and their definition, considerable amount of time was spent in checking this consistency. The algorithm chosen to generate patterns from the selected dataset, apriori, also does not allow much missing values for attributes. With the help of child labor and statistics experts, the researcher decided on a threshold of 90% to avoid missing values. Thus, all attributes with missing values of 90% and above were eliminated. In order to restrict possible attribute values within the required scope, the

values of the attributes individual's occupation last week and industry in which the individual worked last week were aggregated together.

The researcher made further data selection from the target dataset mainly due to time shortage to analyze all of the target dataset. The data mining tasks were carried out only on one region data, *killil* 02.

Further data preparations to optimize data quality for future model building were also performed. Expectation maximization clustering algorithm which is implemented in Knowledge Studio software was used as a tool for data preparation. This algorithm was used to categorize the selected datasets into clusters in an effort to generate relevant patterns. Since apriori algorithm can not handle numeric attributes, such attribute values were transformed to nominal. The data format also was changed into attribute relation file format (arff) because Weka accepts only comma separated arff format text files.

D. Building and Training Models

For each cluster or category identified using Expectation maximization algorithm, another data mining model was applied to identify the critical patterns and relationships.

As mentioned above, the data mining techniques, which were used to perform the data mining tasks are, Expectation maximization clustering algorithm to categorize the dataset into groups and apriori algorithm to identify interesting patterns from each cluster.

Finally, the researcher attempted to interpret the results or discovered patterns together with the domain experts, a child labor expert and statisticians who processed the child labor survey data. Generally, the performance of the model was judged according to the interestingness of the results obtained in relation to the identified problem. Some of the interestingness measures are accuracy, coverage, novelty and applicability of the obtained results (Witten and Frank, 2000).

1.5 Scope and Limitation

The scope of this research is to identify interesting and critical patterns from the child labor database of FDRE CSA. The output of the research can be used as input for child labor prevention programs. The researcher initially intended to compare the patterns to be identified in two regions, *Affar* and *Gambella*. However, the data mining task was limited only to one region, *Affar*, because of time shortage.

Another serious limitation of this research was the availability of literature on the application of data mining for census or survey data.

1.6 Organization of the Thesis

This thesis is organized into five chapters. The first chapter is an introductory part, which discusses the problem area leading to this research project, the general and specific objectives to attain in the research and the methodology to be followed.

The second chapter mainly revolves around the technology to be applied on this research project. Literature is reviewed to know and write about meaning and importance of data mining, steps involved in data mining process and about different types of data mining functionalities and algorithms. A detailed discussion of the algorithm to be utilized in attaining the goal of the data-mining task is also made.

The third chapter is devoted to give further understanding about data collection, storing and processing activities of CSA in general. These general procedures are also directly applied to the 2001 child labor survey database.

The fourth chapter provides discussions about the different data mining steps that were undertaken in this research work. This includes data collection, data selection, preparation, model building and evaluating and interpreting results obtained from apriori.

The last chapter is devoted for the final conclusions and recommendations based on the research findings.

Chapter 2

Data Mining and Association Rule Discovery

2.1 Data Mining

2.1.1 Overview

It is estimated that the amount of information in the world doubles every 20 months. High volume of digital data is enabled by a variety of cutting edge technologies. These include communication technologies (e-mail, internet, extranets and others), enhancing technologies (data warehousing, data mining, portals and others) and intelligent technologies (intelligent agents, internet search engines, and others) (Levin and Zahavi, 1999). Recent advances in communication technologies, on the one hand, and computer hardware and database technologies, on the other, have made it all easier for organizations to collect, store and manipulate massive amounts of data. Databases in today's organizations are measured by terabytes, soon petabytes, encompassing thousands, if not millions, of observations and hundreds even thousands pieces of data (features) in each record (Levin and Zahavi, 1999).

As mentioned by Deogan et. al. (2001), these large databases contain potential gold mine of valuable information, but it is beyond human ability to analyze massive amounts of data and elicit meaningful patterns. Given certain data analysis goal, it has been a common practice to either design a database application on on-line data or use a statistical (or an analytical) package on off-line data along with a domain expert to interpret the result. This traditional method of turning data into knowledge in most application areas, such as marketing, finance, retail,

insurance, science, etc., relies on manual analysis and interpretation. Moreover, it requires one or more analysts who become intimately familiar with the data and serving as an interface between the data, the users and products. This form of manual probing of a dataset is slow, expensive and highly subjective. In fact, as data volumes grow dramatically, this type of manual data analysis is becoming completely impractical in many domains (Fayyad, Piatesky-Shapiro & Smyth, 1996).

The need to scale up human analysis capabilities to handling the large number of bytes that we can collect is both economic and scientific. Businesses use data to gain competitive advantage, increase efficiency, and provide more valuable services to customers. Data we capture about our environment are the basic evidence we use to build theories and models of the universe we live in. Because computers have enabled humans to gather more data than we can digest, it is only natural to turn to computational techniques to help us unearth meaningful patterns and structures from the massive volumes of data (Fayyad, Piatesky-Shapiro & Smyth, 1996).

Up until recently, the ability to analyze and understand volume of data lagged far behind the capability to gather, store and manipulate the data. But not any more. A new generation of computerized methods is emerging in recent years to help the endeavor of interrogating and analyzing very large data sets automatically and efficiently, thereby extracting information and knowledge useful in decision making. These methods are collectively referred to as data mining (Levin and Zahavi, 1999).

Data mining is an interdisciplinary research area spanning several disciplines such as database systems, machine learning, intelligent information systems, statistics, and expert systems. Data mining has evolved into an important and active area of research because of theoretical

challenges and practical applications associated with the problem of discovering (or extracting) interesting and previously unknown knowledge from very large real-world databases. It comes from the idea that large databases can be viewed as data mines containing valuable information that can be discovered by efficient knowledge discovery techniques (Deogun, et. al., 2001).

As Han and Kamber (2001) stated, the major reason that data mining has attracted a great deal of attention in the information industry in recent years is due to the wide availability of huge amounts of data and the imminent need for turning such data into useful information and knowledge. The authors also said that data mining can be viewed as a result of the natural evolution of information technology. An evolutionary path has been witnessed in the database industry in the development of functionalities such as data collection and database creation, data management (including data storage and retrieval, and database transaction processing), and data analysis and understanding (involving data warehousing and data mining).

Data mining techniques are the result of a long process of research and product development. This evolution began when business data was first stored on computers, continued with improvements in data access, and more recently, generated technologies that allow users to navigate through their data in real time. Data mining takes this evolutionary process beyond retrospective data access and navigation to prospective and proactive information delivery. Data mining is ready for application in the different areas because it is supported by three technologies that are now sufficiently mature:

- Massive data collection
- Powerful multiprocessor computers

- Data mining algorithms

The basic steps in the evolution of data mining can be shown below in Table 2.1, which is taken form Threaling (n.d.)

Evolutionary Step	Enabling Technologies	Characteristics
Data Collection (1960s)	Computers, tapes, disks	Retrospective, static data delivery
Data Access (1980s)	Relational databases (RDBMS) Structured Query Language (SQL) Open Database Connection(ODBC)	Retrospective, dynamic data delivery at record level
Data Warehousing & Decision Support (1990s)	On-line analytical processing (OLAP) Multidimensional databases, data warehouses	Retrospective, dynamic data delivery at multiple levels
Data Mining (Emerging Today)	Advanced algorithms, multiprocessor computers, massive databases	Prospective, proactive information delivery

Table 2.1 Steps in the evolution of data mining

The core components of data mining technology have been under development for decades, in research areas such as statistics, artificial intelligence, and machine learning. Today, the maturity of these techniques, coupled with high-performance relational database engines and broad data integration efforts, make these technologies practical for current data warehouse environments (Threaling, n.d.).

Formally defined, data mining is a new generation of computerized methods for "extracting previously unknown, valid, and actionable information from large databases and then using this information to make critical decision" (Cabena et. al. 1998). This emerging technology can be put as one of the evolutionary steps in digital information processing.

Data mining allows for coping with today's large business problems by taking advantage of new hardware and software technologies, and using scalable algorithms to sift through a large amount of data and extract useful and valid information from the data relatively efficiently and inexpensively (Levin and Zahavi, 1999).

Indeed, data mining has become a new paradigm for decision making, with applications ranging from database marketing and electronic commerce to fraud detection, credit scoring, warranty management, even auditing data before storing it in a database. The fundamental reason for data mining is that there is a lot of money hidden in the data. Data mining can be used to control costs as well as contribute to revenue increases. Without data mining all we have are opinions. But what we need is information. We need to understand the data and translate it into useful information for decision making (Levin and Zahavi, 1999).

2.1.2 Data Mining and other Statistical Tools

According to TCC (1999), statistical theory and practice has been a traditional method to study and analyze data for many years. Unfortunately, these traditional methods fail when it comes to analyzing large amounts of data. When a small data set is involved with only several predictors, one can manipulate the data set manually using statistical methods to search for the combination of predictors and their transformations that best fit the data.

But with a large dataset, containing hundreds of potential features and tens of thousands of observations, the number of possible combinations of features to explore is enormous and beyond the capacity of any given individual, even to a group of statistical experts, to handle in any

reasonable amount of time. This dimensionality issue also makes it difficult to identify which features interact with one another, which features exhibit non-linear relationships, which features are redundant, which are irrelevant, which ones are noisy, etc. As a result, modelers need to experiment with a large number of combinations of predictors and try out a large number of transformations of the original features to express non-linearity and interactions, which increase the number of potential features in a model beyond a comprehensible reach (Levin and Zahavi, 1999).

Data mining takes advantage of advances in the fields of artificial intelligence and statistics, and solves the above mentioned limitations of traditional statistics methods. However, data mining does not replace traditional statistical techniques. Rather, it is an extension of statistical methods that is in part the result of a major change in statistics community (TCC, 1999).

OLAP (On-Line Analytical Processing) is also part of the spectrum of decision support tools. Trybula (1999) states that Online analytical processing is the application of traditional query-and-reporting programs to describe and extract what is in a database. It is used to answer why certain things are true. The user forms a hypothesis about a relationship and verifies it with a series of queries against the data. OLAP analysis is essentially a deductive process. But as the number of variables being analyzed is in the dozens or even hundreds, it becomes much more difficult and time-consuming to find a good hypothesis and analyze the database with OLAP to verify or disprove it. Data mining is different from OLAP because rather than verify hypothetical patterns, it uses the data itself to uncover such patterns. It is essentially an inductive process. Data mining can go further and discover a pattern the analyst did not think to try (TCC, 1999).

Data mining and OLAP can complement each other. For example, in the early stages of knowledge discovery process, OLAP can help in exploring the data by focusing attention on important variables, identifying exceptions, or finding interactions (TCC, 1999).

2.1.3 Data Mining and Knowledge Discovery in the Real World.

Historically, the notion of finding useful patterns in data has been given a variety of names, including data mining, knowledge extraction, information discovery, information harvesting, data archaeology, and data pattern processing. The term data mining has mostly been used by statisticians, data analysts, and the management information systems communities. It has also gained popularity in the database field (Fayyad, Piatesky-Shapiro & Smyth, 1996).

According to Han and Kamber (2001), the term 'Data Mining' is a misnomer. Data mining should have been more appropriately named "knowledge mining from data," which is unfortunately somewhat long.

2.1.3.1 Knowledge Discovery in Database (KDD)

Han and Kamber (2001) states that many people treat data mining as a synonym for Knowledge Discovery in Databases. Alternatively, others view data mining as simply an essential step in the process of knowledge discovery in databases. The phrase Knowledge Discovery in Databases was coined at the first KDD workshop in 1989 (Piaketsky-Shapiro 1991) to emphasize that knowledge is the end product of a data-driven discovery.

According to Fayyad et. al. (1996), KDD refers to the overall process of discovering useful knowledge from data, and Data Mining refers to a particular step in this process. Data mining is the application of specific algorithms for extracting patterns from data.

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 data sets and still run efficiently, how results can be interpreted and visualized, and how the overall man-machine interaction can usefully be modeled and supported. Han and Kamber (2001) also said that KDD process consists of an iterative sequence of steps.

The additional steps in the KDD process, such as data preparation, data selection, data cleaning, incorporation of appropriate prior knowledge, and proper interpretation of the result of mining are essential to ensure that useful knowledge is derived from the data.

KDD is the nontrivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data (Fayyad, Piatesky-Shapiro, and Symth, 1996).

The KDD process involves using the database along with any required selection, preprocessing, sub sampling and transformations of it; applying data-mining methods (algorithms) to enumerate patterns from it; and evaluating the products of data mining to identify the subset of the enumerated patterns deemed knowledge. The data mining component of the KDD process is concerned with the algorithmic means by which patterns are extracted and enumerated from data. The overall KDD process includes the evaluation and possible interpretation of the mined patterns to determine which patterns can be considered new knowledge. Thus we can say the

overall process of building and implementing a data mining solution is referred to as KDD (Fayyad, Piatesky-Shapiro, and Symth, 1996).

2.1.3.2 Data Mining

Trybula (1997), states that knowledge discovery (KD) is the process of transforming data into previously unknown or unsuspected relationships that can be employed as predictors of future action. Trybula also notes that KDD is a term that has been employed to encompass both data mining and KD. Essentially, the basic tasks of data mining and KD are to extract particular information from existing databases and convert it into understandable or sensible conclusions or knowledge. As indicated above, data mining can be viewed as a step in the KDD process that consists of applying data analysis and discovery algorithms that, under acceptable computational efficiency limitations, produce a particular enumeration of patterns (or models) over the data.

Although data mining is a step in the knowledge discovery process, it is a more popular term than the longer term of knowledge discovery in databases in industry, in media and in database research milieu (Han and Kamber, 2001). The authors also suggested adopting a broad view of data mining functionality: data mining is the process of discovering interesting knowledge from large amounts of data stored either in databases, data warehouses, or other information repositories. Trybula (1997) also mentioned that although data mining is usually considered as one step in knowledge discovery process in computer science literature, the term refers to the entire process from construction of databases through pattern identification and reporting. In this research paper, the term data mining is used in its broader view.

2.1.4 The Data Mining Process

According to the Two Crows data mining process model, the basic steps of data mining for knowledge discovery are:

1. Define business problem
2. Build data mining database
3. Explore data
4. Prepare data for modeling
5. Build model
6. Evaluate model
7. Deploy model and results

As indicated above, the basic steps for data mining to knowledge discovery start with defining business problem, which may not always be obvious. The prerequisite to knowledge discovery is to understand the problem area and the data. Without this understanding, no algorithm, regardless of sophistication, is going to provide useful result. Furthermore, we will not be able to identify the data mining tasks to be performed, prepare the data for mining, or correctly interpret the results. To make the best use of data mining we must make a clear statement of our objectives (TCC, 1999).

Trybula (1997) states that along with this step is a parallel effort of understanding the structure of the data, or data discovery. It is important to understand the formulation of the data in order to understand what is included and more importantly, what is not. We have to be aware of the potential of the data to be able to solve the identified problem. The combination of these two

activities drives the development of a goal for the work. Two Crows Corporation (1999) also states that an effective statement of the problem will include a way of measuring the results of your knowledge discovery project. It may also include a cost justification.

Once the problem area and the objective of the project are clearly defined, the next step is to create a target data. Creating target data for data mining requires collecting the data to be mined in a separate database. Based on task definitions and goals, data are selected from the data warehouse (Trybula, 1997). Databases are heterogeneous, containing a wide variety of data, not all of which may be appropriate for the analysis at hand. Incorporating all data in the analysis may make it difficult for the data mining tools to identify the most influential predictors explaining the phenomenon, as these predictors may be "diluted" by all of those irrelevant pieces of data (Levin and Zahavi, 1999).

Creating a target data set involves focusing on a subset of variables or data samples, on which discovery is to be performed (Fayyad, Piatesky-Shapiro, and Symth, 1996). One needs to extract the target data to analyze in a way that is consistent with the problem involved and the objective of the project. One can use either subjective judgment or segmentation analysis, a data mining model, to extract the relevant target set to participate in the data mining process (Levin and Zahavi, 1999).

In connection with the need for creating a target dataset for data mining task, TCC (1999) states that pp.23

> You will be better off creating a separate data mart for data mining. Mining the data will make you a very active user of the data warehouse, possibly causing resource

> allocation problems. In addition you may want to bring in data from outside your company to overlay on the data warehouse data or you may want to add new fields computed from existing fields. You may need to gather additional data through surveys. The structure of the corporate data warehouse may not easily support the kinds of exploration you need to do to understand this data. You may want to store this data in a different DBMS with a different physical design than the one you use for your corporate data warehouse especially if it can not handle the resource demands of data mining.

Once the target database is built, the data must be preprocessed before a model can be developed. This process, called data preparation and preprocessing, is often the most time consuming task of the data mining process especially if data is drawn directly from the company's operational databases rather than from a data warehouse (Levin & Zahavi, 1999). Han and Kamber (2001) explains that the data stored in a database may reflect noise, exceptional cases, or incomplete data objects. When mining data regularities, these objects may confuse the process, causing the knowledge model constructed to over fit the data. As a result, the accuracy of the discovered patterns can be poor.

In order to insure the accuracy of the data, cleaning, validation, and completion processes are performed to develop accurate database for data mining. Cleaning data refers to the process of reviewing the data to find incorrect characters or mistransmitted information (Trybula, 1997). As stated by Levin & Zahavi (1999), data preprocessing also involves other data processing tasks such as overlaying of data from other resources, consolidating and amalgamating records, summarizing fields, checking for data integrity, detecting irregularities and illegal fields, filling in for missing values, trimming outliers, cleaning noise. While being tedious and somewhat boring, data preparation and preprocessing is definitely a critical function of the knowledge

discovery process with significant impact on the quality of the modeling results (Levin & Zahavi 1999).

The next important step of data mining process is data reduction and projection. This means finding useful features to represent the data depending on the goal of the task. As TCC (1999) explains, the goal of exploring the data is to identify the most important fields in predicting an outcome, and determine which derived values may be useful. This step is the final data preparation step before building models. The author divides this task into four main parts:

a. Select variables
b. Select rows
c. Construct new variables
d. Transform variables

With dimensionality reduction methods, the effective number of variables under consideration can be reduced (Fayyad, Piatesky-Shapiro, and Symth, 1996).

As Levin and Zahavi (1999) states, the predictive power of data resides in transformation of the data, rather than in the raw data itself. These transformations are designed to account for non-linear relationships between the dependent variable and one or more independent variables (assuming all the other are constant), identifying pair-wise interaction, perhaps even higher-order interactions, between independent variables, tracking seasonal and time-related effects, even transforming data to make them compatible with the theoretical assumptions underlying the model involved. In data mining, it is common to create as many possible transformations of

variables, and then analyze them for significance. Data analysis, visualization techniques, and domain knowledge may help identify the appropriate transformations to use (Levin and Zahavi, 1999).

It is only at this point that one invokes data mining models and tools to interrogate the data and convert it into knowledge for decision making. TCC (1999) emphasizes the point that data mining model building is an iterative process. At this stage, we select a particular data-mining method that matches the goals of the data mining process defined in the first step. As mentioned earlier, the data mining component of the KDD process often involves repeated iterative application of particular data mining methods in searching for patterns of interest in a particular representational form.

By and large, data mining models belong to the following three major categories: descriptive models, predictive models and link analysis. Within each category of models there are several data mining technologies, and each technology may be solved by means of several algorithms (Levin and Zahavi 1999).

The modeling engine provides a set of outputs that need to be evaluated and interpreted to make sure the resulting model is any good, and convert the model results into useful knowledge for decision making. Taking the results of the data mining models for granted, without any evaluation process could be very risky and lead to grave consequences. In fact, the knowledge evaluation process is not a separate stage to be conducted following the data mining stage; rather, it should be integrated and interwoven in all the components of the data mining process.

Knowledge evaluation is often conducted by means of statistical measures and tools (Levin and Zahavi 1999).

Finally the discovered knowledge should be expressed in high-level languages, visual representations, or other expressive forms so that the knowledge can be easily understood and directly usable by humans. Han and Kamber (2001) emphasize this step especially if the data mining system is to be interactive. According to the authors, this requires the system to adopt expressive knowledge representation techniques, such as trees, tables, rules, graphs, and charts, cross tabs, matrices or curves.

2.1.5 Data Mining Techniques

The data mining goals are defined by the intended use of the system. The two high-level primary goals of data mining in practice tend to be prediction and description. Prediction involves using some variables or fields in the data base to predict unknown or future values of other variables of interest. In other words predictive mining tasks perform inference on the current data in order to make prediction. Descriptive mining focuses on finding human-interpretable patterns describing the data (Fayyad, Piatesky-Shapiro, and Symth, 1996).

In predictive models, the values or classes we are predicting are called the response, dependent or target variables. The values used to make the prediction are called the predictor or independent variables. Predictive models are built, or trained, using data for which the value of the response variable is already known. This kind of training is sometimes referred to as supervised learning,

because calculated or estimated values are compared with the known results. On the other hand, descriptive techniques are sometimes referred to as unsupervised learning because there is no already known result to guide the algorithms (Two Crows Corporation, 1999). As Levin and Zehavi (1999) stated, descriptive models interrogate the data base to identify patterns and relation ships in the data. AS Han and Kamber (2001) states, users may sometimes have no idea which kinds of patterns in their data may be interesting, and hence may like to search for several different kinds of patterns in parallel. Thus it is important to have a data mining system that can mine multiple kinds of patterns to accommodate different user expectations of applications.

The goals or functions of prediction and description can be achieved using a variety of particular data-mining methods (Fayyad, Piatesky-Shapiro, and Symth, 1996). Clustering algorithms, pattern recognition models, visualization methods, and link analysis are the major members of descriptive models Levin and Zehavi (1999).

2.1.5.1 Descriptive Models

2.1.5.1.1 Clustering Algorithms

Clustering is a common descriptive task where one seeks to identify a finite set of categories or clusters to describe the data. It is mapping a data item into one of several clusters which are not pre-specified but are determined from the data. Clusters are formed by finding natural groupings of data items based on similarity matrices, proximity considerations and probability measures (Levin and Zehavi, 1999). Two Crows Corporation (1999) 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.

Unlike classification, we don't know what the clusters will be when we start, or by which attributes the data will be clustered. In general, the class labels are not present in the training data simply because they are not known to begin with. Consequently, experts' knowledge is required to interpret the clusters (Two Crows Corporation, 1999).

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

2.1.5.1.3 Link Analysis

Link analysis is a descriptive approach to exploring data that can help identify relation ships among values in a database. The two most common approaches to link analysis are association discovery and sequence discovery (TCC, 1999).

2.1.5.1.3.1 Association Discovery

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 $A \Rightarrow B$, that is , (A1,----,Am $\rightarrow$ B1,---,Bn), where Ai (for i ε {1,----,m} and Bj (for jε{1,----,n}) are attribute-value pairs. The associations rule $A \Rightarrow B$ is interpreted as database tuples that satisfy the condition in A are also likely to satisfy the condition in B.

Two probability measures, called support and confidence, are introduced to assess associations in the database. The support (or prevalence) of a rule is the proportion of observations that contain the item or item set of the rule. It is also known as the coverage of the rule. As defined by Witten and Frank (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. We should be able to evaluate rules using different techniques especially because of the combinatorial explosion that results in enormous number of rules (TCC, 1999). 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.

The problem of detecting patterns was first introduced in the application domain of market basket analysis to find association between two sets of bought products. Thus initial research was concentrated on the discovery of Boolean association rules. However, more recent work is focusing on quantitative association rules (Han and Kamber 2001).

2.1.5.1.3.2 Sequence Discovery

Sequence discoveries are association rules with time dimensions. A sequential pattern is an association between sets of items, in which some temporal properties between items in each set and between sets are satisfied. In particular, items in a set have the same temporal reference (Levin and Zehavi, 1999). As Trybula (1997) states, sequential patterns are identified in a technique for predicting future activities based on observing trends over a period of time. It is based on the fact that previous activities have the potential for indicating future activities.

Two Crows Corporation (1999) point out that association or sequence rules are not really rules, but rather descriptions of relationships in a particular database. There is no formal testing of models on other data to increase the predictive power of these rules. Rather there is an implicit assumption that the past behavior will continue in the future.

2.1.5.2 Predictive Models

In predictive modeling one identifies patterns found in the data to predict future values. Predictive modeling consists of several types of models: classification models, regression models and AI-based models (Levin and Zahavi, 1999).

2.1.5.2.1 Classification

Classification problems aim to identify the characteristics that indicate the group to which each case belongs. This pattern can be used both to understand the existing data and to predict how new instances will behave. Data mining creates classification models by examining already classified data (cases) and inductively finding a predictive pattern. These existing cases may come from historical database. They may come from an experiment in which a sample of the entire database is tested in the real world and the results used to create a classifier. Sometimes an expert classifies a sample of the database, and this classification is then used to create the model which will be applied to the entire database (TCC, 1999).

2.1.5.2.2 Regression

Regression uses existing values to forecast what other values will be. This method can be used to define the boundary condition by evaluating the data and determining the boundary through mathematical analysis (Trybula, 1997). In the simplest case, regression uses standard statistical techniques such as linear regression. The linear regression method is used for modeling continuous response. Unfortunately, many real world problems are not simply linear projections of previous values. Therefore, more complex techniques, such as logistic regression, decision trees, or neural nets, may be necessary to forecast future values (TCC, 1999).

2.1.5.2.2.1 Time Series Regression

Time series forecasting predicts unknown future values based on a time-varying series of predictors. Like regression, it uses known results to guide its predictions. Models must take into account the distinctive properties of time (TCC, 1999).

2.1.5.2.4 AI Based Models

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

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

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

Consequently, we as model developer will need a variety of tools and technologies in order to find the best possible model.

2.2 Association Rule Discovery

2.2.1 Overview

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

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

Since its introduction in 1993, the task of association rule mining has received a great deal of attention. Today the mining of such rules is still one of the most popular pattern discovery methods in knowledge discovery (Hipp et. al., 2000).

Association rules are widely used in data mining to find patterns in data. The patterns reveal combinations of events that occur at the same time. Once identified, these combinations, which are also known as "group associations" can be used to improve decision-making in a wide variety of applications. The association rule components can help one to

- Management of 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 (SPSS Inc. 2002).

Association rules are similar with classification rules. However, in case of association rules any attribute might occur on the right hand side with any possible value, and a single association rule

can often predict the value of more than one attribute (Witten and Frank, 2000). These authors also explain that in order to generate association rules, rule induction procedure would have to be executed for every possible combination of attributes, with every possible combination of values, on the right hand side. The result would be an enormous number of association rules, which would then have to be pruned down on the basis of their support and confidence. As mentioned earlier, support of a rule means the number of instances or observations they predict correctly and confidence of a rule is when support figure is expressed as a proportion of all instances that the rule applies to. However, these standard criteria are often not sufficient to restrict the set of rules to the interesting ones. Therefore, efficient algorithms are needed that restrict the search space and check only a subset of all rules, but, if possible, without missing important rules. One such algorithm is the apriori algorithm, which was developed by Agrawal et al. (Borgelt, 2002).

2.2.2 How Do We Extract Association Rules from Datasets

As mentioned earlier, an association rule is an expression $X \Rightarrow Y$, where X and Y are sets of items. The meaning of such rule is quite intuitive: Given a database D of transactions-where each transaction $T \in D$ is a set of items, $X \Rightarrow Y$ express that whenever a transaction T contains X then T probably contains Y also. The probability or rule confidence is defined as the percentage of transactions containing Y in addition to X with regard to the overall number transactions containing X. That is, the rule confidence can be understood as the conditional probability $P(Y \subseteq T / X \subseteq T)$. The idea of mining association rules originates from the analysis of market-basket data where rules like "A customer who buys products X_1 and X_2 will also buy product y with probability c%." are found. Their direct applicability to business problems together with

their inherent understandability- even for non-data mining experts- made association rules a popular mining method. Moreover it became clear that association rules are not restricted to dependency analysis in the context of retail applications, but are successfully applicable to a wide range of business problems (Hipp et. al., 2000).

2.2.3 Basic Principles

2.2.3.1 Formal Problem Description

As Hipp et. al., (2000) put it, the association rule discovery problem can be expressed mathematically as follows:

> Let L= $\{x_1. \ . \ . \ x_n\}$ be a set of distinct literals, called items. A set $X \subseteq L$ with k=$|X|$ is called a k-itemset or simply an itemset. Let a database D be multi-set subsets of L. Each T$\in$ D is called a transaction. We say that a transaction T $\in$ D supports an itemset $X \subseteq L$ if $X \subseteq T$ holds. An association rule is an expression $X \Rightarrow Y$, where X, Y are itemsets and $X \cap Y = \emptyset$ holds. The fraction of transactions T supporting an itemset X with respect to database D is called the support of X, supp(X) = T$\in$D $|\{X \subseteq T\}| / |D|$. The support of a rule $X \Rightarrow Y$ is defined as supp $(X \Rightarrow Y) = p(X \cup Y)$.

The main challenge when mining association rules is the immense number of rules that theoretically must be considered. In fact the number of rules grows exponentially with $|L|$. Since it is neither practical nor desirable to mine such a huge set of rules, the rule sets are typically restricted by minimal thresholds for the quality measures support and confidence, **minsupp** and **minconf** respectively. This restriction allows us to split the problem into two separate parts: An itemset X is frequent if **supp** (X) $\geq$ **min-supp**. Once, F=$\{X \subseteq L / X$ frequent$\}$, the set of all frequent itemsets together with their support values is known, deriving the desired association rules is straight forward: For every X $\in$ F check the confidence of all values

$X/Y \Rightarrow Y, Y \subseteq X, \emptyset \neq Y \neq X$ and drop those that do not achieve minconf. According to its definition above, it suffices to know all support values of the subsets of X to determine the confidence of each rule. The knowledge about the support values of all subsets of X is ensured by the downward closure property of itemset support: All subsets of a frequent itemset must also be frequent (Hipp et. al., 2000).

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

1. Find all itemsets (sets of items appearing together in a transaction) whose support is greater than the specified threshold. Itemsets with minimum support are called frequent itemsets.
2. Generate association rules from the frequent itemsets. To do this, consider all partitioning of the itemset into rule left-hand and right-hand sides. Confidence of a candidate rule $X \rightarrow Y$ is calculated as support (XY) / support (X). All rules that meet the confidence threshold are reported as discoveries of the algorithm.

2.2.3.2Traveling the Search Space

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

The basic principle of most association rule algorithms is that if the parent class E' of a class E does not contain at least two frequent itemsets then E must also not contain any frequent itemset. If we encounter such a class E' on our way down the tree, then we have reached the border separating the infrequent from the frequent itemsets. We do not need to go behind this border so

we prune E and all descendants of E from the search space. This procedure allows us to efficiently restrict the number of itemsets to investigate. We simply determine the support values only of those itemsets that we "visit" on our search for the border between frequent and infrequent itemsets. Today's common approaches for search employ either breadth-first search (BFS) or depth-first search (DFS). With BFS the support values of all (k-1) itemsets are determined before counting the support values of the k-itemsets. In contrast, DFS recursively descends the tree structure defined for itemsets (Hipp et. al., 2000).

2.2.3.3 Determine Itemset Supports

One common approach to determine the support value of an itemset is to directly count its occurrences in the database. Then all transactions are scanned and whenever one of the candidates is recognized as a subset of a transaction, its counter is incremented. Typically subset generation and candidate lookup is integrated and implemented on a hash tree or a similar data structure. Not all subsets of each transaction are generated but only those that are contained in the candidates or those which have a prefix in common with at least one of the candidates (Hipp et. al., 2000).

Another approach is to determine the support values of candidates by set intersections. A tid is a unique transaction identifier. For a single item the tidlist is the set of identifiers that correspond to the transactions containing this item. Accordingly tidlists also exist for every itemset X and are denoted by X.tidlist. The tidlist of a candidate $C= X \cup Y$ is obtained by $C.tidlist=X.tidlist \cap Y.tidlist$. The tidlists are sorted in ascending order to allow efficient intersections. By buffering the tidlists of frequent candidates as intermediate results, we

remarkably speedup the generation of the tidlists of the following candidates. Finally the actual support of a candidate is obtained by determining │C.tidlist│ (Hipp et. al., 2000).

2.2.4 Apriori Algorithm

When mining association rules there are mainly two problems to deal with: First of all there is the algorithmic complexity. The number of rules grows exponentially with the number of items. Fortunately today's algorithms are able to efficiently prune this immense search space based on minimal thresholds for quality measures on the rules. Second, interesting rules must be picked from the set of generated rules. This might be quite costly because the generated rule sets normally are quite large and in contrast the percentage of useful rules is typically only a very small fraction. The work concerning the second problem mainly focuses on supporting the user when browsing the rule set and the development of further useful quality measures on the rules (Hipp et. al., 2000).

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

Thus, the process or the algorithm starts by scanning all transactions in the database and computing the frequent items. Next, a set of potentially frequent candidate 2-itemsets is formed from the frequent items. Another database scan obtains their supports. The frequent 2-itemsets

are retained for the next pass, and the process is repeated until all frequent itemsets have been enumerated. The algorithm has three main steps:

1. Generate candidates of length k from the frequent $(k-1)$ length itemsets, by a self-join on F_k–1. For example, for $F_2 = \{AC, AT, AW, CD, CT, CW, DW, TW\}$, we get $C3 = \{ACT, ACW, ATW, CDT, CDW, CTW\}$.
2. Prune any candidate that has at least one infrequent subset. For example, *CDT* will be pruned because *DT* is not frequent.
3. Scan all transactions to obtain candidate supports.

Apriori stores the candidates in a hash tree for fast support counting. In a hash tree, itemsets are stored in the leaves; internal nodes contain hash tables (hashed by items) to direct the search for a candidate (Hipp et. al., 2000). Apriori uses BFS and counts occurrences of itemsets. It prunes those candidates that have an infrequent subset before counting their supports. This optimization becomes possible because BFS ensures that the support values of all subsets of a candidate are known in advance.

Apriori counts all candidates of cardinality together in one scan over the database. The critical part is looking up the candidates in each of the transactions. For this purpose introduces a hash tree structure. The items in each transaction are used to descend in the hash tree. Whenever we reach one of its leafs, we find a set of candidates having a common prefix that is contained in the transaction. Then these candidates are searched in the transaction. In the case of success the counter of the candidate in the tree is incremented. AprioriTID is an extension of the basic Apriori approach. Instead of relying on the raw database AprioriTID internally represents each

transaction by the current candidates it contains. With AprioriHybrid both approaches are combined (Hipp et. al., 2000).

Chapter Three

Analysis of Child Labor Survey at Central Statistics Authority

3.1. Introduction

Child labor remains to be a serious problem in the world today. The International Labor Organization (ILO) statistical data proves that many children in the world are exposed to dangerous and hazardous activities. According to the ILO, the number of working children between the ages of 5 and 14 is about 211 million. The overwhelming majority of these are in the developing countries, in the sub-Saharan Africa 23 percent, Asia and Pacific 60 percent and Latin America and the Caribbean 8 percent (ILO, 2002).

Similar to other developing countries, child labor is also a problem in Ethiopia. Children are engaged in economic and non-economic activities not compatible with their age in both the urban and rural areas. In rural areas, child work is perceived as un avoidable or even necessary part of children's socialization process. Children are commonly involved in domestic chores, and are supposed to assist in manual labor in the agriculture sector such as attending domestic animals, seeding and harvesting. In urban areas, children are often forced into labor due to a situation of persisting poverty, which requires all family members to contribute to the household income.

Unacceptable forms of exploitation of children at work exist and persist, but they are particularly difficult to research due to their hidden, sometimes illegal or even criminal nature. However, the perception of child labor by society further complicates the problem. For some, it is widely

accepted as a natural order of bringing up children to be responsible future adults consequently child labor is often equated with child work, with the argument that work is good for the socialization of children and a means of helping families.

The international community through the ILO makes a distinction between child work and child labor. The former refers to any work for pay or unpaid family (domestic) work, which is part of socialization process. Child work may sometimes include hazardous work. On the other hand, child labor refers to situations where children are actually doing work either in industries or occupations where the child is below the established minimum age. This contravenes the ILO Conventions No. 138 on Minimum age of employment and Convention No.182 on the worst forms of child labor. It also includes, children who try to earn their living either through paid employment or engage in small business activities opened by the children themselves or working for the benefit of adults who exploit them. These include those children working in hazardous work environment, in exploitative condition, work for long hours or work in activities that require intense physical effort, and work in servitude.

Heavy work at an early age has a direct deterring effect on children's physical and mental development. Physically, children are not fit to long hours of strenuous and monotonous work. Moreover, children are especially vulnerable to accidents because they have neither the awareness of the danger nor the knowledge of the precaution to be taken at work. In general child labor is considered as an aspect of child exploitation and child abuse.

Ethiopia has ratified the UN convention on the Rights of the Child and included provisions in her constitution on basic rights and privileges of children. Recommendation accompanying the

convention concerning the prohibition and immediate action for the elimination of the worst forms of child labor, states that

> …detailed information and statistical data on the nature and extent of child labor should be compiled and kept up to date to serve as a basis for determining priorities for national action for the abolition of child labor, in particular for the prohibition and elimination of its worst forms, as a matter of urgency.

3.2 Child Labor Survey in Ethiopia

The availability of data on working children and their analysis on a continuous basis is particularly essential for establishing intervention programs and formulating policies for the eventual elimination of child labor. As fully discussed in the introductory part, there is shortage of data and also information on child labor issue especially in developing countries. Experts in the area make decisions based on best opinions. Sometimes responses from parents and children may be considered. Expert best opinions and parents and children responses should be further verified using objective analysis method. To address this problem of data gap, the Government of Ethiopia, through the Ministry of Labor and Social Affairs (MOLSA) and the Central Statistical Authority (CSA), with the technical and financial support of the ILO, has launched a National Child Labor Survey in March 2001. ILO principally funded the survey as part of its statistical information and monitoring program on child labor. Initial attempt, before the stand-alone child labor survey of 2001, was made by FDRE CSA to gather information on socio-demographic and economic activities of children aged 5-14. Even though the 1999 national labor force survey includes some information on child labor, it is just a module attached to a survey. The major concern of the survey is on adult labor statistics.

3.3 Scope and Coverage of the Survey

The Child Labor Survey was a household-based survey, where only conventional households were the sampling units. Hence, children who don not live in the households such as street children and children in institutions were excluded. The survey results did not cover the situation of street children and may not show the situation of working children at specific work places, like in plantations, industries, etc.

3.4 Objectives of the Survey

The 2001 Ethiopia Child Labor Survey was designed to provide statistical data on children's activities focusing on the status of schooling, non-economic and economic activities. Specifically, the survey was aimed at to provide statistical data that will help to:

- Learn the demographic and socio-economic characteristics of Children: age, sex, literacy status levels of education and training, occupations, skill-levels, hours of work, earnings and other working and living conditions;
- Assess the working situation of children and the influence on their education, health, physical and mental development;
- Examine the characteristics of the sectors that employ most children;
- Study the movement of children between households;
- Identify where and how long the children have been working and the factors that lead children to work or families to put children to work;
- Assess the health and welfare status of working children;
- Generate data on child affairs for intervention and policy formulation.

All of these goals are to be met by CSA using simple statistical tools on the collected data. The results of CSA statistical analysis are simple tabulations and summaries extracted directly from the database without any further processing. Since the collected data represent responses of parents and children, it is highly possible that the responses are biased. Thus, although identifying the factors that lead children to work is mentioned as one of CSA's objectives, it is highly advantageous to generate meaningful and useful associations from the database. These associations can objectively indicate relationships between attributes and some of them can be interpreted as major reasons of children to work or parents to send their children to work. Comparisons can also be made between the reasons mentioned by parents and children and the rules generated by data mining algorithm.

3.5 Data Collection Methods

Questionnaire is the major field data collection instrument used by CSA to collect different types of data at different time periods. Nothing was different to collect data about child labor. The questionnaire used for the child labor survey consisted of three Forms. Form I was used to obtain information on the socio-economic and demographic composition of household members and specific questions about households and housing particulars. Form II was used to obtain information on children aged 5-17 years on their schooling and non-schooling activities, including working conditions and related matters. The Form III of the survey questionnaire which is similar in its content to the second part but which refers to children aged 10-17 years was addressed to the children themselves.

In all the three forms of the survey questionnaires, most questions were designed with pre-coded answers. The 2001 National Child Labor Survey of Ethiopia covered 11 killils, both rural and urban areas. However, it has not covered non-sedentary areas of two zones of the Afar Region and six zones of the Somali Region. Residents of collective quarters, homeless and foreigners were not covered in the survey. For the purpose of the survey, the population of the country was divided into three major categories namely, rural, major urban centers and other urban centers.

In addition to the above domains of study, the survey results were also reported at regional and country levels by aggregating the survey results from the corresponding domains. All in all 48 basic survey domains (reporting levels) including urban part of each regional state, total (urban+rural) part of each region, country level urban, country level rural and country level total were defined for the survey.

From the total child labor database, the researcher selects two specific reporting levels or regions: Affar and Gambella. This selection is purely based on subjective judgment. The judgment was made with domain experts, statisticians as well as child labor experts. The two regions are selected due to their significant difference in their respective rate of children involvement in economic activities, housekeeping activities and in both economic activities and housekeeping activities. Therefore, the researcher believes that meaningful explanation can be extracted from the database for the difference mentioned.

3.5 CSA Database

The CSA child labor database consists of

1. Area identification of the selected household and member's demographic, economic and social activities. Each selected household has an identification particular which is 18 character long. Each individual within a household is also given a serial number starting from head of the household. Households can thus be characterized by a unique identity number within the entire child labor database. In this section, basic demographic, social, economic and housing conditions are recorded for each household and specifically to children aged 5018 years.
2. Information on children aged 5-17 years such as children migration status, educational status and working activities. This information is obtained by addressing the appropriate questions to parents or guardians or responsible proxies in the household where the child usually resides.
3. The same information mention in number 2 above was also obtained from children aged 10-17 and stored in the database.

Since the 2001 child labor survey is one time study, the original database resides only in one file.

3.6 Data Processing

The methodology section of CSA is responsible for selecting EAs, designing sample frames and limiting sample sizes for any type of study conducted. The same is true to the 2001 stand-alone child labor survey. This section also assigns appropriate weight for variables under study.

3.7 Data Quality Assurance

Data quality assurance mechanisms start at the early stage of the child labor survey. Before starting the actual field survey, the field staffs were given training program in two stages. The first-stage trainees are composed of statisticians from Head Office, and Branch Statistical Offices, and some selected senior field supervisors. Many of the personnel trained in the first-stage, conducted similar training for field supervisors and enumerators for about two weeks in the 22 Branch Statistical Offices that are located all over the country. During this second-stage training, the field staff are given detailed classroom instruction on the objectives and uses of the survey, concepts and definitions of terms used, interviewing procedures, how to fill questionnaires, … etc. The enumerators' training also included a field practice to reinforce the classroom training.

Although CSA collects and processes a wide variety of data, it has standard methods, techniques and tools for processing its data and then presenting its outputs. The following are the general procedures followed by CSA in processing it data. The same procedures were also applied to process the 2001 stand-alone child labor survey data.

3.8 Manual Data Editing and Coding

The filled-in questionnaires that were received from the field were first subjected to manual editing and coding. Instruction manuals used in editing and coding were prepared by departments that originally developed the questionnaires. Training is given to editing and coding employees

about the specific questionnaire. Then data is coded and edited manually by the employees of the section. Manual coding at this stage is required for those answers which were not coded in the field data collection. All the edited and coded questionnaires are again fully verified and checked for consistency before they are submitted to the data entry section. Different forms are used to control the flow of questionnaires to avoid misplacement and loss of questionnaires.

3.9 Data Entry

Data entry section in turn received coded data from data editing and coding section, and enters it in CSA database. Data entry programs needed to enter the data for the survey is prepared using the Integrated Microcomputer Processing System (IMPS). IMPS is a software for entry, editing, tabulation and management of census and survey data. This software is specifically designed for large census and surreys. CSA could not use other statistical application software such as access and SPSS because of their limited data processing capacity. The volume and range of data collected by CSA at different times is really very large and various. So it needs larger capacity software to accommodate all of its huge volume of various types of collected data. As the researcher understands, CSA has not yet obtained another equivalent or better system to manage its census data. This software provides a user interface that imposes checks and restrictions on how individual records are handled. It also provides a mechanism for receiving data from the field, checking it, and raising queries as needed when data are incomplete or inconsistent. The software operates under DOS operating system and uses a menu-based approach.

The major modules of IMPS are

A. Data Dictionary

B. Data Entry (CENTRY)

C. Data Edit and Imputation (CONCOR)

D. Publication Tabulation (CENTS)

E. Quick Tabulation (QUICKTAB)

F. Table Retrieval (TRS)

G. Variance Calculation (CENVAR)

H. Data Entry Control (CENTRACK)

I. Utilities

J. Exit

Here data is entered in coded format. The data is again verified or cleaned using the computer after it has been entered into IMPS database. Using edit specifications rules prepared by the subject matter specialists, the entered data are checked for consistencies and then computer editing or data cleaning is made. Frequent references are normally made to the original document to check values which seem to be out of the acceptable range of values. This takes very long time. This is an important part of data processing operation done by CSA in attaining the required level of data quality.

The cleaning and computer editing for the survey was done using CONCOR and SPSS for Windows. CONCOR is the editing module of IMPS. It is used to identify and edit invalid or inconsistent data. Missing Enumeration Areas (EAs), or inclusion of EAs were checked using the master list. The number of Households in each EA and the number of eligible children in each household were checked. Codes for each item were carefully checked and verified. Errors or inconsistencies in data were checked and corrected.

3.10 Merging and Tabulation

For the Child Labor Survey, the cleaned data for the three forms were merged. For the purpose of tabulation different new variables were created. Dummy tables were prepared by the subject-matter specialists and handed to computer programmers. Computer programs used to produce statistical tables were developed using the CENTS software package. CENTS is a module of IMPS, which tabulates, summarizes, and displays statistical tables. CENTS has the facility of producing a single table at different reporting levels required (National, Regional, Zone, etc) including urban and rural reporting levels. Consistency checks and rechecks were also made based on tabulation results. This is done by senior programmers using IMPS software in collaboration with relevant senior staff of the CSA. Coefficient of variation (CV) was generated for selected statistical tables. CV's are produced using the CENVAR component of IMPS. CENVAR calculates the reliability (precision) measures for a sample design (estimates, standard error, CV, confidence interval, design effect) Statistical tables generated were converted to WordPerfect format for publication purposes.

The final data files of the survey are kept in the following formats: ASCII, SPSS and CSPro (Ibid).

Chapter 4

Experimentation

4.1 Overview

In this section, the researcher presents how each step of data mining process is applied on the real world dataset, child labor data. As mentioned in TCC (1999), the basic steps of data mining process are:

- Defining the target problem and the goal of data mining task.
- Identifying source of data.
- Selecting and storing target dataset in a separate database.
- Data preparation.
- Model building
- Model testing or evaluation.
- Model deployment.

4.2 Data Mining Goals

The first important step in the whole data mining process is to understand the need to do data mining, i.e. understanding the problem we have to solve. This is the objective of the data mining effort. According to TCC (1999), identifying the goal of the data mining process is a prerequisite to discover knowledge from the database. The goal of the data mining process depends on the type of problem to be solved using data mining technology. So, before starting the actual data

mining task, we should be able to clearly define our problem and also have a good understanding of our data to be used for the data mining task.

As mentioned in chapter 1 section 1.3, the main objective and output of this research project is to find interesting and meaningful patterns and relationships in child labor survey database at FDERE CSA. Provided that meaningful relationship among attributes were to be established, prevention programs for abusive form of child labor could have a better understanding of the nature of child labor in Ethiopia and thus could develop strategic solution to avoid the most intolerable form of child labor and protect working children. After we define the goal of our data mining task, we should be able to select an appropriate data mining tool which can perform the expected functions.

4.2.1 Data Mining Tool Selection

Data mining tool selection is normally initiated after the definition of problem to be solved and the related data mining goals. However, more appropriate tools and techniques can also be selected at the model selection and building phase. Selection of appropriate data mining tools and techniques depends on the main task of the data mining process. As Han and Kamber (2001) states, the following factors can be used to asses the usefulness of data mining tools or software to the intended data mining task:

- The goal of the data mining task in the research project. The selected software should be able to provide the required data mining functions and methodologies. The data mining

functions that were to be carried out in this research project are clustering, to prepare the data for next algorithm, and association rule mining, to find meaningful relationship in the dataset. The data mining software selected for this research are Knowledge Studio (to categorize the selected dataset into clusters) and Weka (to find interesting patterns in the selected dataset). In addition, the methodologies used by data mining software to perform each of the data mining functions are also an important factor to consider. This means the algorithms supported by the software should be known. The researcher selected two algorithms, expectation maximization for clustering and apriori for association. Expectation maximization algorithm is chosen to categorize the selected dataset into clusters before applying an association rule algorithm. It is used as a means of data preparation for model building. Apriori algorithm is used to identify interesting patterns and relationships out of the selected clustered and non-clustered dataset.

- Architecture and operating system. The computer architecture and the operating system on which the software runs should be first studied. Some data mining software operate on specific types of architecture and operating systems. In the research project, both Knowledge studio and Weka operated of stand alone and MS Windows operating system.
- Data sources- Specific data format on which the data mining software will operate is also another important factor to consider. The suitable data format for Knowledge studio and Weka data mining software are MS Access or MS Excel and arff formats respectively.
- Scalability-Maximum number of columns and rows the software can efficiently handle. In the original target dataset there are about 365 columns and 5000 records. However, in the selected data set, the number of columns and the number of records were reduced to 147

and 2398. Both of the selected data mining software support the volume of data in the selected dataset.

- Visualization capabilities- The variety, quality, and flexibility of visualization tools may strongly influence the usability, interpretability, and attractiveness of a data mining system. Both Knowledge studio and Weka have a facility to visualize their outputs.

Review of data mining and knowledge discovery software tools by Goebel and Gruenwald (1999) was used as a base for further investigation and analysis. The researcher selected Knowledge Studio version 3.0 of Angoss Software Corporation for clustering and Weka for association rule mining. The researcher chose to use these data mining softwares mainly due to easy and quick access. Weka provides a number of data mining functionalities such as classification, clustering, association, attribute selection and visualization. Familiarity was also another reason to select Weka data mining software.

However, the researcher preferred to use Knowledge studio rather than Weka for clustering datasets because the output of Knowledge studio is more interpretable than that of Weka. Knowledge studio provides tree data and the tree itself for the its clustered dataset. Thus, instances categorized in each cluster can be easily observed. Even though Weka has the facility for clustering, the clusters cannot be easily interpreted without high involvement of child labor and statistical experts. Within the time limit given for this research, the researcher preferred to use relatively easily understandable clustering software.

Weka is developed at the University of Waikato in New Zealand. “Weka” stands for the Waikato Environment of Knowledge Analysis. The system is written in Java, an object-oriented programming language that is widely available for all major computer platforms, and Weka has been tested under Linux, Windows, and Macintosh operating systems. Java allows us to provide 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 expects the data to be fed into to be in ARFF format. It is necessary to have information about each attribute which can not be automatically deduced form the attribute values (Witten and Frank, 2000).

Weka includes a variety of tools for preprocessing a dataset, such as attribute selection, attribute filtering and attribute transformation, feeding into a learning scheme, and analyze the resulting classifier and its performance. Weka is organized in packages that correspond to a directory hierarchy. The important packages of Weka are association, attribute selection, classifiers, clusterers, estimators, and filters packages. The association package has only one association rule mining algorithm, apriori (Witten and Frank, 2000).

4.3 Data Understanding

After setting up the problem and a rough plan for its solution, the researcher proceeded with the central item in data mining process - data. There are several things to be learned about the data before the actual application of data mining techniques.

4.3.1 Initial Data Collection

Using the right data for data mining task is one of the primary keys for successful data mining (TCC, 1999). Taking all of the databases on 2001 child labor survey would be too much for this research project. The researcher selected two killils, Affar and Gambella, based on judgmental sampling. In order to come up with this selection, series discussions were conducted with Associate child labor expert at International Labor Organization. As mentioned in chapter 1, section 1.4, these two sections were selected due to their higher difference in the percentage of children engaged in productive and housekeeping activity.

CSA use software called Integrated Microcomputer Processing System (IMPS) to manage its census and survey data of various kinds. So the child labor survey data initially resides on IMPS database organized into rows and columns. Rows represent records whereas columns represent attributes. Selecting the target dataset from the child labor database was done using utilities option provided on IMPS.

4.3.1.1 Description of the data collected

After the initial data collection, new database was created both in Ms Excel and Ms Access formats. Thus two tables, one for data from each region, appeared in Ms Excel and Ms Access. Ms Access and Ms Excel were used for preparing the dataset into a form acceptable by the selected data mining software, and Knowledge studio Weka. The first table is for child labor data of Affar region with 361 columns and 2397 rows. The second table is for child labor data of Gambella region with 361 columns and 2016 rows. The 253 attributes are taken from the three

different forms of questionnaire, and the remaining 108 variables are added by the data analysts to facilitate their data analysis process. Some of these added variables are created by combining attribute values within the child labor database and the others new attributes created by data processing department employees for their own analysis purpose.

4.4 Data preparation

The main goal of this activity was the production of the dataset (datasets) used for modeling by the selected data mining software. The activities during this phase included data cleaning, data selection, attribute or feature selection, transformation and aggregation, integration and formatting.

4.4.1 Data quality assessment and data cleaning

A data quality assessment identifies characteristics of the data that will affect the model quality. Data cleaning is the process of examining data and determining the existence of incorrect characters and mis-transmitted information (TCC, 1999). In this phase the researcher attempted to ensure not only the correctness and consistency of values but also that all the data are measured in consistent way. One of the data mining software used in this research project, Weka, forces the use of clean data. Weka would not open a data file unless it is clean and in required format. This data mining software has a facility for data analysis. An option named 'Analysis' helps the user to clean the data by pointing out the type and the position of error in the dataset. Weka considers any attribute value which is not compatible with its requirements as error. The researcher used this option to clean the entire target data selected from the child labor database.

In this phase the dataset of each killil was categorized into two groups in order to facilitate the data cleaning, data preparation and model building process. The second category with higher proportion of missing values was ignored for each killil. Basically the separation was required in order to reduce missing values in the database and the number of attributes handled at a time.

4.4.2 Data Selection

The whole target dataset may not be taken for data mining task. Irrelevant or unneeded data are usually eliminated from the data mining database before starting the actual data mining function. Other criteria for excluding data may include resource constraints, cost, restrictions on data use, or quality problems (TCC, 1999). As mentioned above, the target dataset was divided into four groups to facilitate the data mining process. Because of time constraint, the researcher could not apply the intended data mining task on all of the target dataset. Thus only the first category, killil 02, was taken for clustering and rule generation functions. The selected dataset is attached in appendix 2.

4.4.3 Feature Selection

The ideal practice for variable selection is to take all the variables in the database, feed them to the data mining tool and let it find those which are the best predictors. But in practice this practice doesn't 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. Although in principle some data mining algorithms will automatically ignore irrelevant variables and properly account for related (covariant) columns, in practice it is

wise to avoid depending solely on the tool. Often knowledge of the problem domain helps to make these selections correctly (TCC, 1999).

After consulting statistics experts at FDRE CSA about the meaning of the attributes added by data analysts for easy data analysis purpose, the researcher totally eliminated the variables from the target dataset. These attributes were eliminated because they are already represented by other attributes in the database or they are redundant. This reduces the number of attributes from 365 to 253. As mentioned above, the target dataset taken from each killil was divided into two parts and only the first half of one *killil, Affar,* was taken for analysis. Thus, out of the 253 attributes in the original target dataset, the researcher included only 147 in the selected dataset. Attributes with no variation in their value through out the dataset and attributes which serve for assigning sequence number for the records were all eliminated. Attributes which have missing value for more than 90% of instances are also cancelled. Out of the 147 attributes in the selected dataset, a total of 58 attributes are eliminated because they have missing values in more than 90% of records. And the other three attribute again are eliminated since their value is constant throughout the database or they are used simply to assign sequences to the records. At this point, the number of attributes is diminished to 86. Since association rule mining algorithms generate association rules only from frequent itemsets, missing values considerably reduce their performance. When the number of missing values becomes higher, the rules to be generated by the association rule algorithm reduces continuously. Sometimes the association rule algorithm even may not produce any rules. In addition to this, some attributes which were classified as irrelevant to the problem domain by the domain experts are also eliminated. After this further attribute elimination, the number of attributes taken for analysis was 63. Domain expert help was

obtained in analyzing the importance of the attributes to the data mining goal and fixing the threshold for missing values. The list of attributes selected for the first and second experiment of association rule is shown in appendix 3 and 4.

4.4.4 Data Transformation and Aggregation

This task includes constructive data preparation operations such as the production of derived attributes, creating new records or transformed values for existing attributes, consolidating and amalgamating records and summarizing fields.

The first task performed in this case was the aggregation of the occupation and industry in which the child is working for or had worked for. Since the possible values for the two attributes are more detailed than required for this research purpose, the researcher decided to represent them using general figures. As a result, the number of possible values for these two attributes reduced from 103 to 11 for occupation and from 150 to 20 for industry in which the child is working for. Since apriori algorithm of association rule mining accepts only nominal attributes, the researcher had to redefine numeric attributes as nominal by listing down their possible values.

Since the data mining software used to generate association rules accepts data only in arff format, the researcher first converted the data on Ms Excel file into comma separated text format and then to arff format. Data in arff format is then given to Weka software, apriori algorithm for association rule mining. However, Knowledge studio software used to prepare the data for the Weka does not require arff format data representation.

4.5 Model Building

In the initial experiment the researcher took 86 attributes for association rule model building purpose. The selection of attribute is made using subjective judgment. But in arriving at these attributes, the researcher took input from child labor domain experts and statisticians. To build the association rule model, the arff format of the selected dataset was given to Weka, apriori algorithm. The following is the first ten rules generated in the first attempt:

Experiment 1

=== Run information ===

Scheme: weka.associations.Apriori -N 10 -T 0 -C 0.9 -D 0.05 -U 1.0 -M 0.1 -S -1.0

Relation: *killil* 02 part 1

Instances: 2398

Attributes: 86 List of attributes used appeared in appendix 3

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

Apriori
=======

Minimum support: 0.95

Minimum metric <confidence>: 0.9

Number of cycles performed: 1

Generated sets of large itemsets:

Size of set of large itemsets 1: Frequency of large itemset 6

Size of set of large itemsets 2: Frequency of large itemset 15

Size of set of large itemsets 3: Frequency of large itemset 15

Size of set of large itemsets 4: Frequency of large itemset 6

Best rules found:

1. Telephone=2 2343 ==> Electric Mitad=2 2343 conf:(1)

2. Telephone=2 Car=2 2335 ==> Electric Mitad=2 2335 conf:(1)

3. Telephone=2 Bofe and Sofa=2 2314 ==> Electric Mitad=2 2314 conf:(1)

4. Telephone=2 Bofe and Sofa=2 Car=2 2306 ==> Electric Mitad=2 2306 conf:(1)

5. Telephone=2 Refrigerator=2 2300 ==> Electric Mitad=2 2300 conf:(1)

6. Type of fuel=5 Telephone=2 2300 ==> Electric Mitad=2 2300 conf:(1)

7. Telephone=2 Refrigerator=2 Car=2 2292 ==> Electric Mitad=2 2292 conf:(1)

8. Type of fuel=5 Telephone=2 Car=2 2292 ==> Electric Mitad=2 2292 conf:(1)

9. Telephone=2 Refrigerator=2 Bofe and Sofa=2 2283 ==> Electric Mitad=2 2283 conf:(1)

10. Bofe and Sofa=2 2344 ==> Electric Mitad=2 2342 conf:(1)

Meaning of the parameters mentioned above

-N(required number of rules output)	10
-T(metric type by which to rank rules)	0 (confidence)
-C (the minimum confidence of a rule)	0.9
-D (delta at which the minimum support is decreased at each iteration)	0.05
-U (upper bound for minimum support)	1.0
-M (the lower bound for the minimum support)	0.1
-S (significance of a rule at a given level)[1]	-1.0

[1] The letter –S indicates significance test for each rule. In apriori algorithm we use only confidence of a rule to measure its significance or accuracy.

As Han and Kamber (2001) states, the occurrence frequency of an itemset is the number of transactions that contain the itemset. This is also known as the frequency or support count of the itemset. If an itemset satisfies minimum support count, then it is a frequent or large itemset.

Apriori algorithm generates strong association rules from these frequent or large itemset. According to apriori property, which is the base for apriori association rule algorithm, all non empty subset of a frequent itemset must also be frequent. Thus, a number of interrelated rules can be generated from large or frequent itemsets.

It is also true that if the subset is not a frequent itemset then the superset also is not in the frequent itemset collection and ignored in the process of rule generation. The support (or prevalence) of a rule is the proportion of observations that contain the item or itemset of the rule. It is also known as the coverage of the rule. The support of a rule $A \Rightarrow B$ is the percentage of transaction that contain $A \cup B$. Its confidence is measured as the percentage of transactions containing A that also contain B. Confidence of a rule represents the number of instances correctly predicted out of the total instances that the rule applies to. The ten best rules satisfying minimum support of 90% and minimum confidence of 95% threshold are listed above. Rules satisfying minimum requirement of support and confidence threshold are strong rules. However confidence and support are not the exhaustive measures to evaluate the importance of the rules generated. The researcher used additional interestingness measure, experts' opinion, to further evaluate rule importance.

In selecting rules for discussion, the researcher focused on the rules generated from the superset of frequent or large itemset consisting of the highest size of large itemset. The following are rules selected for discussion from experiment 1:

1. Telephone=2 2343 ==> Electric *Mitad*=2 2343 confidence :(1)

The meaning of this rule is if a household does not have telephone, then this household will not also have electric *mitad.* This rule is an example of two itemsets rule. The support for this rule can be computed by dividing the figure on the right-hand-side of the rule 2345 by the total number of instances considered in generating association rules, 2398. This rule has a support of 98%. The number 2345 on the right-hand-side of the rule indicates the number of items covered by its antecedent. The confidence is also computed by dividing the figure on the left-hand-side of the rule by the figure on the right-hand-side of the rule. Following the rule is the number of those items for which the rule's consequent holds as well.

7. Telephone=2 Refrigerator=2 Car=2 2292 ==> Electric *Mitad*=2 2292 confidence: (1)

As in the previous case rule 7 can be interpreted as, if a household does not have telephone, refrigerator and car, it will not have electric *mitad*. So the possession of electric *mitad* is associated with other three types of household properties; telephone, refrigerator and car.

This rule has a confidence of 100% and a support of 96%.

8. Type of fuel=5 Telephone=2 Car=2 2292 ==> Electric *Mitad*=2 2292 confidence: (1)

If the type of fuel used by a household is firewood or charcoal or dung, and does not have telephone and car, this household will not also have electric *mitad*. This rule has a confidence of 100% and a support of 96%.

9. Telephone=2 Refrigerator=2 Bofe and Sofa=2 2283 ==> Electric *Mitad*=2 2283 confidence :(1)

If a household does not have telephone, refrigerator, *bofe* and *sofa* then the household will not also have electric *mitad.* This rule has a confidence of 100% and support of 95%.

It is easier to observe that the above rules apply to the real world and thus are meaningful. The researcher categorized these rules as meaningful because they are logical and supported by real life experience especially in Ethiopia. So, the rules are confirming facts and realities. However, these set of rules concentrated on lower level concept that is relationship between properties owned by a household. When compared to the problem of child labor, details of household properties were judged to be lower level concepts by child labor experts.

The above result indicates that the attributes still needed further selection. In the second round 63 attributes for 2398 were selected by the researcher with the help of domain expert as relevant attributes to the problem at hand. The following result was obtained:

Experiment 2

=== Run information ===

Scheme: weka.associations.Apriori -N 10 -T 0 -C 0.9 -D 0.05 -U 1.0 -M 0.1 -S -1.0

Relation: k021

Instances: 2398

Attributes: 63 (List of attributes appeared in appendix 2)

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

Apriori
=======

Minimum support: 0.8

Minimum metric <confidence>: 0.9

Number of cycles performed: 4

Generated sets of large itemsets:

Size of set of large itemsets L(1): Frequency of large itemset 5

Size of set of large itemsets L(2): Frequency of large itemset 6

Size of set of large itemsets L(3): Frequency of large itemset 1

Best rules found:

1. Children living outside family=2 Has the child been Living with parents since birth=1 2017 ==> Living outside current town/rural=2 1937 conf:(0.96)

2. Has the child been Living with parents since birth=1 2170 ==> Living outside current town/rural=2 2076 conf:(0.96)

3. Children living outside family=2 2199 ==> Living outside current town/rural=2 2090 conf:(0.95)

4. Religion=4 2135 ==> Living outside current town/rural=2 2028 conf:(0.95)

5. Living outside current town/rural=2 Has the child been Living with parents since birth=1 2076 ==> children living outside family=2 1937 conf:(0.93)

6. Has the child been Living with parents since birth=1 2170 ==> children living outside family=2 2017 conf:(0.93)

7. Living outside current town/rural=2 children living outside family=2 2090 ==> Has the child been Living with parents since birth=1 1937 conf:(0.93)

8. Living outside current town/rural=2 2259 ==> children living outside family=2 2090 conf:(0.93)

9. Religion=4 2135 ==> children living outside family=2 1966 conf:(0.92)

10. Living outside current town/rural=2 2259 ==> Has the child been Living with parents since birth=1 2076 conf:(0.92)

The researcher selected the following rules for discussion:

1. Children living outside family=2 Has the child been Living with parents since birth=1 2017 ==> Living outside current town/rural=2 1937 confidence :(0.96)

If a child is not currently living apart from his or her family and has been living with his or her family since birth, then 96% of the household members considered under this experiment have never been living outside the current town or rural it is currently residing. The support of this rule is 85%.

4. Religion=4 2135 ==> Living outside current town/rural=2 2028 confidence :(0.95)

If the religion of one of the household members is Muslim, then 95% of the household members considered under this experiment have never lived outside the current town or rural it is currently residing. The support of this rule is 89%. According to statistics experts Muslim is the common religion of the people residing in Affar *killil* from where the selected dataset for this data mining task is taken. That is why it appeared as frequent itemset and included in the rule.

9. Religion=4 2135 ==> children living outside family=2 1966 confidence :(0.92)

If the religion of one of the household members is Muslim, then 92% of the children considered in this experiment are not currently living apart from their family. The support of the rule is 89%.

10. Living outside current town/rural=2 2259 ==> Has the child been Living with parents since birth=1 2076 confidence :(0.92)

If a household has never been living outside the current town or rural it is currently residing, then 92% the children considered in this experiment also have been living with his or her family since birth. The support of this rule is 94%.

As indicated by the above rules, there is strong relationship between household migration and the possibility of children living with their parents. According to the explanation of child labor experts, if a household frequently migrates from one place to another, then most of the time it will not be able to keep the children. Such type of family is obligated to give the children to near relatives or to non-relatives with whom good relationship is established. According to child labor and statistics experts, the above rules tell about the living condition of the household living around the specific region selected, Affar.

From the generated rules one can easily understand that there is no much household migration and also most of the time children in this area are living with their families. As the statisticians of FDRE CSA further elaborated, this result in part may be due to scope and coverage of the child labor survey. The survey did not include non-sedentary areas of two zones of Affar region. If data on these two zones were collected and considered for this data mining task, the results would have been much different.

Even though the rules generated in second experiment are evaluated as meaningful and interesting, the researcher believed that more of such rules could be generated and thus proceeded

with another experiment. The researcher attempted to generate more rules using the same attribute and see if there was any change. The following was the result:

Experiment 3

=== Run information ===

Scheme: weka.associations.Apriori -N 15 -T 0 -C 0.9 -D 0.05 -U 1.0 -M 0.1 -S -1.0

Relation: k021

Instances: 2398

Attributes: 63

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

Apriori
=======

Minimum support: 0.75

Minimum metric <confidence>: 0.9

Number of cycles performed: 5

Generated sets of large itemsets:

Size of set of large itemsets L(1): Frequency of large itemset 5

Size of set of large itemsets L(2): Frequency of large itemset 8

Size of set of large itemsets L(3): Frequency of large itemset 4

Best rules found:

1. Religion=4 Has the child been Living with parents since birth=1 1957 ==> Living outside current town/rural=2 1882 conf:(0.96)

2. children living outside family=2 Has the child been Living with parents since birth=1 2017 ==> Living outside current town/rural=2 1937 conf:(0.96)

3. Religion=4 children living outside family=2 1966 ==> Living outside current town/rural=2 1884 conf :(0.96)

4. Has the child been Living with parents since birth=1 2170 ==> Living outside current town/rural=2 2076 conf:(0.96)

5. Children living outside family=2 2199 ==> Living outside current town/rural=2 2090 conf:(0.95)

6. Religion=4 2135 ==> Living outside current town/rural=2 2028 conf:(0.95)

7. Injuries at work place=2 1988 ==> Living outside current town/rural=2 1878 conf:(0.94)

8. Living outside current town/rural=2 Has the child been Living with parents since birth=1 2076 ==> children living outside family=2 1937 conf:(0.93)

9. Religion=4 Has the child been Living with parents since birth=1 1957 ==> children living outside family=2 1823 conf:(0.93)

10. Has the child been Living with parents since birth=1 2170 ==> children living outside family=2 2017 conf:(0.93)

11. Religion=4 Living outside current town/rural=2 2028 ==> children living outside family=2 1884 conf:(0.93)

12. Religion=4 Living outside current town/rural=2 2028 ==> Has the child been Living with parents since birth=1 1882 conf:(0.93)

13. Religion=4 children living outside family=2 1966 ==> Has the child been Living with parents since birth=1 1823 conf:(0.93)

14. Living outside current town/rural=2 children living outside family=2 2090 ==> Has the child been Living with parents since birth=1 1937 conf:(0.93)

15. Living outside current town/rural=2 2259 ==> children living outside family=2 2090 conf:(0.93)

As can be seen from the above rules, there is only one different rule generated in this experiment:

7. Injuries at work place=2 1988 ==> Living outside current town/rural=2 1878 conf:(0.94)

This rule indicates relationship between children injury at work place and migration status of household. If a child has never been injured before, then 94% of the parents considered in this experiment also has never been outside their original place of residence. This rule also implies that the families of children who are injured at work place, migrates from one living place to another for different reasons. This rule is classified as important and interesting by the child labor experts.

In order to get more powerful rules, the researcher applied a clustering algorithm as a tool for data preparation. As Han and Kamber (2001) states cluster analysis can be used as a stand-alone tool to gain insight into the distribution of data to observe the characteristics of each cluster, and to focus on a particular set of clusters for further analysis. Alternatively, it may serve as a preprocessing step for other algorithms, such as association and classification, which would then operate on the detected clusters.

The first step in the clustering process was to import the dataset from Excel into Knowledge studio clustering software to get categories of similar instances. Knowledge studio supports two types of clustering algorithms; K-means and Expectation Maximization (EM). EM algorithm assigns each object to a cluster according to a weight representing the probability of membership. It is best suitable for dataset containing significant amount of missing values (Bishop, 1995). Even though the researcher attempted to eliminate attributes with greater than 90% missing values from the dataset, the number of missing data still left was not insignificant.

Considering the number of attributes involved, the researcher attempted to separate the dataset into two, three, four and five clusters in series of experiments. The researcher attempted to assign possible interpretation to the different clusters generated with the help of domain experts. After having through discussion on possible meanings of the clusters with statisticians and child labor experts, the researcher decided to take the fourth clustering experiment which divided the selected dataset into five clusters. And again from the four clusters, the one with larger number of instances was taken for analysis purpose.

Clustering Experiment

The parameters set for the cluster run during the four consecutive experiments are as follows:

Table 4.1 Summary of clustering input parameters for the four cluster runs

Cluster run	No. of variables	No. of records	No. of clusters	No. of iterations
1	63	2398	2	1000
2	63	2398	3	1000

3	63	2398	4	1000
4	63	2398	5	1000

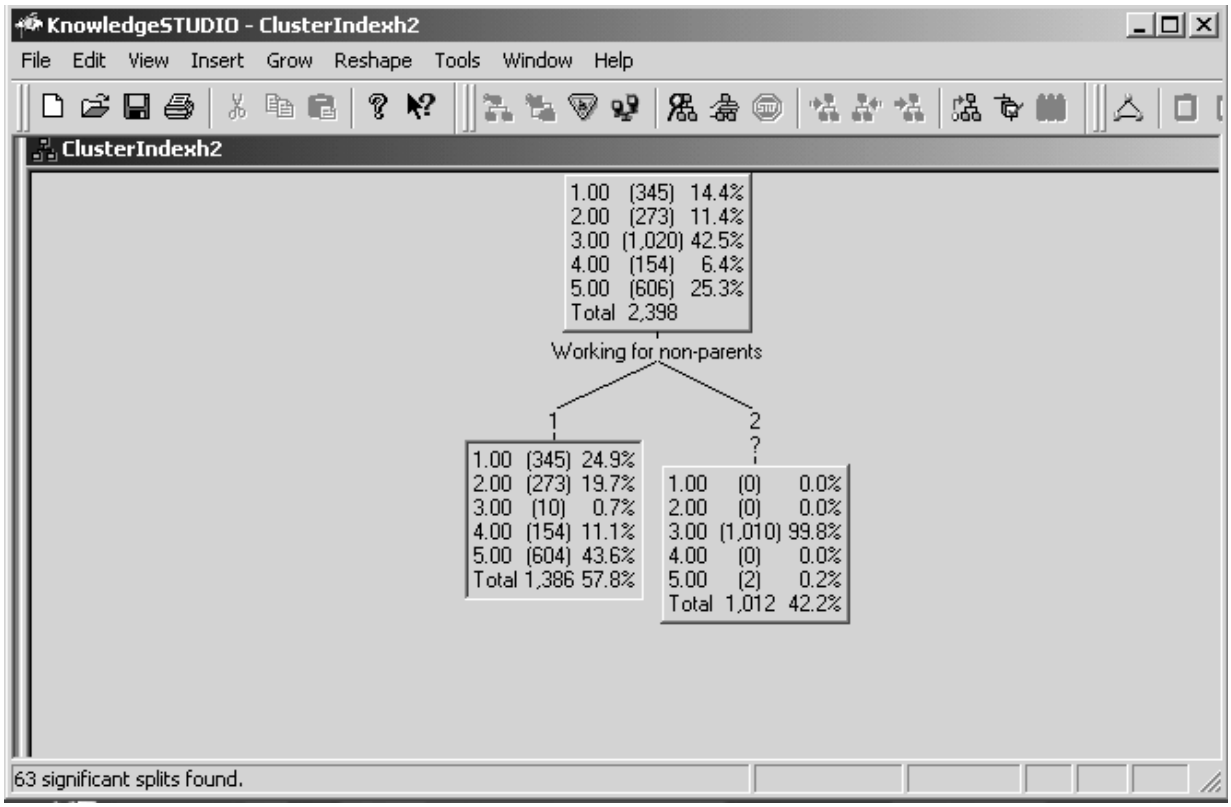
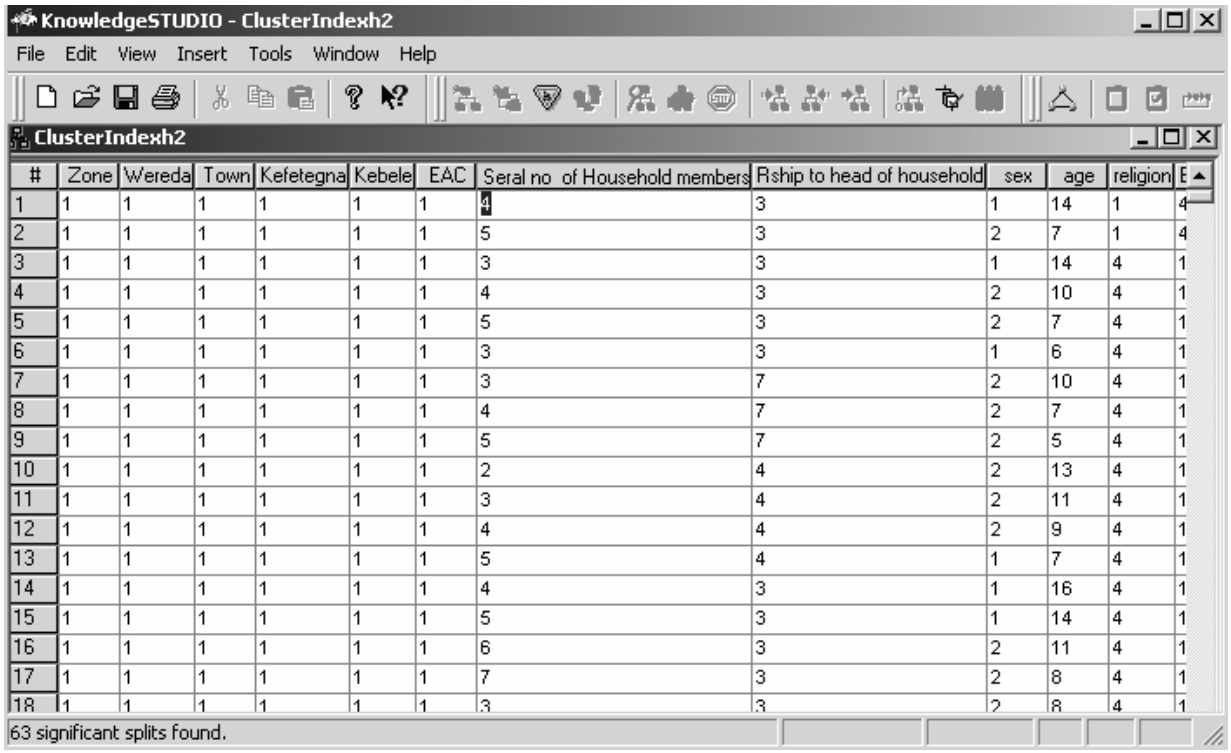


Figure 4.1 Output of the fourth cluster run (cluster tree)

The above figure indicates the three nodes of tree generated by EM clustering algorithm. The first node represents the five clusters the selected dataset is categorized into. As can be seen above, the third cluster consists of higher proportion of the records, 42.5%. These clusters further can be split using a variable working for non parents. The node no.1 shown below the root node above in Figure 4.1 represents data about children working for non-parents while the node no.2 represents data about children not working for non-parents and those children for whom this

question is not applicable. A question mark in the target dataset always represents inapplicable question to the respondent.

The tree data shown in figure 4.2 gives information about which record fall into which cluster. So it is possible to further investigate the behavior of the records or data which fall into one category. After separating the selected datasets into clusters, the researcher then exported the tree data of the cluster run into Ms Access. Microsoft Access has the facility for grouping all one cluster records consecutively. Thus, the exported data were organized into major clusters or classes. As mentioned above, the third cluster was chosen due to its higher number of records involved.



#	Zone	Wereda	Town	Kefetegna	Kebele	EAC	Seral no of Household members	Rship to head of household	sex	age	religion	E
1	1	1	1	1	1	1	4	3	1	14	1	4
2	1	1	1	1	1	1	5	3	2	7	1	4
3	1	1	1	1	1	1	3	3	1	14	4	1
4	1	1	1	1	1	1	4	3	2	10	4	1
5	1	1	1	1	1	1	5	3	2	7	4	1
6	1	1	1	1	1	1	3	3	1	6	4	1
7	1	1	1	1	1	1	3	7	2	10	4	1
8	1	1	1	1	1	1	4	7	2	7	4	1
9	1	1	1	1	1	1	5	7	2	5	4	1
10	1	1	1	1	1	1	2	4	2	13	4	1
11	1	1	1	1	1	1	3	4	2	11	4	1
12	1	1	1	1	1	1	4	4	2	9	4	1
13	1	1	1	1	1	1	5	4	1	7	4	1
14	1	1	1	1	1	1	4	3	1	16	4	1
15	1	1	1	1	1	1	5	3	1	14	4	1
16	1	1	1	1	1	1	6	3	2	11	4	1
17	1	1	1	1	1	1	7	3	2	8	4	1
18	1	1	1	1	1	1	3	3	2	8	4	1

Figure 4.2 Output of the fourth cluster run (Tree Data)

The next step performed by the researcher at this stage was to export the data from Ms Access to Ms Excel and then convert the data format to comma separated text document and finally to arff format. Then the arff format of cluster number 3, with 63 attributes and 1020 records was given to Weka, apriori for rule generation. The following result was generated after the first run:

Experiment 4

=== Run information ===

Scheme: weka.associations.Apriori -N 10 -T 0 -C 0.9 -D 0.05 -U 1.0 -M 0.1 -S -1.0

Relation: k021q

Instances: 1020

Attributes: 63

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

Apriori

=======

Minimum support: 0.95

Minimum metric <confidence>: 0.9

Number of cycles performed: 1

Generated sets of large itemsets:

Size of set of large itemsets L(1): Frequency of large itemset 6

Size of set of large itemsets L(2): Frequency of large itemset 15

Size of set of large itemsets L(3): Frequency of large itemset 20

Size of set of large itemsets L(4): Frequency of large itemset 15

Size of set of large itemsets L(5): Frequency of large itemset 6

Size of set of large itemsets L(6): Frequency of large itemset 1

Best rules found:

1. Did you have job last week=4 1009==>Engagement in productive activities last week=2 1009 conf:(1)

2. Engagement in productive activities last week=2 1009 ==> Did you have job last week=4 1009 conf:(1)

3. Productive work for non-parents=2 1008 ==> Productive work for family=2 1008 conf:(1)

4. Productive work for family=2 1008 ==> Productive work for non-parents=2 1008 conf:(1)

5. Productive work for non-parents=2 Did you have job last week=4 1007 ==> Productive work for family=2 Engagement in productive activities last week=2 1007 conf:(1)

6. Productive work for non-parents=2 Engagement in productive activities last week=2 1007 ==> Productive work for family=2 Did you have job last week=4 1007 conf:(1)

7. Productive work for family=2 Did you have job last week=4 1007 ==> Productive work for non-parents=2 Engagement in productive activities last week=2 1007 conf:(1)

8. Productive work for family=2 Engagement in productive activities last week=2 1007 ==> Productive work for non-parents=2 Did you have job last week=4 1007 conf:(1)

9. Productive work for non-parents=2 Productive work for family=2 Did you have job last week=4 1007 ==> Engagement in productive activities last week=2 1007 conf:(1)

10. Productive work for non-parents=2 Productive work for family=2 Engagement in productive activities last week=2 1007 ==> Did you have job last week=4 1007 conf:(1)

From the above 10 best rules generated by apriori, the researcher selected one general rule for discussion:

5. Productive work for non-parents=2 Did you have job last week=4 1007 ==> Productive work for family=2 Engagement in productive activities last week=2 1007 confidence: (1)

If a child does not work for non-parents and did not have job last week, then he or she also does not work for his or her family and also this child did not engage in any productive activity last week. The support of this rule is 99%. This rule indicates strong relationship between the attributes work for non-parents, having job (last week), engagement in productive activity (last week), and productive work for family. The child labor experts find this rule to be interesting because it tells important point about child labor: Children who are exposed to domestic work are most of the time exposed to productive activity.

The remaining rules are different combinations of the same attributes described in rule number 5 above so they do not convey difference in meanings.

As can be seen from the outputs listed above, the rules generated out of clustered dataset revolve around child work, domestic or productive. Clustering datasets helps the researcher to limit or direct the relationship and association towards the most demanding class or attribute. Since the associations are to be generated between similar attributes, there exists high support and confidence for the rules generated using clustering algorithm as a means of data preparation.

From the result of the experiments undertaken, the researcher together with the child labor experts concluded that, the attributes selected for association models using clustering as tool for data preparation are more relevant to the problem area of child labor.

The researcher attempted to further investigate the possibility of generating more interesting rule by increasing the number of rules generated from 10 to 20. The results obtained were as follows:

Experiment 5

=== Run information ===

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

Relation: k021q

Instances: 1020

Attributes: 63

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

Apriori

=======

Minimum support: 0.95

Minimum metric <confidence>: 0.9

Number of cycles performed: 1

Generated sets of large itemsets:

Size of set of large itemsets L(1): Frequency of large itemset 6

Size of set of large itemsets L(2): Frequency of large itemset 15

Size of set of large itemsets L(3): Frequency of large itemset 20

Size of set of large itemsets L(4): Frequency of large itemset 15

Size of set of large itemsets L(5): Frequency of large itemset 6

Size of set of large itemsets L(6): Frequency of large itemset 1

Best rules found:

1. Did you have job last week=4 1009 ==> Engagement in productive activities last week=2 1009 conf:(1)

2. Engagement in productive activities last week=2 1009 ==> Did you have job last week=4 1009 conf:(1)

3. Productive work for non-parents=2 1008 ==> Productive work for family=2 1008 conf:(1)

4. Productive work for family=2 1008 ==> Productive work for non-parents=2 1008 conf:(1)

5. Productive work for non-parents=2 Did you have job last week=4 1007 ==> Productive work for family=2 Engagement in productive activities last week=2 1007 conf:(1)

6. Productive work for non-parents=2 Engagement in productive activities last week=2 1007 ==> Productive work for family=2 Did you have job last week=4 1007 conf:(1)

7. Productive work for family=2 Did you have job last week=4 1007 ==> Productive work for non-parents=2 Engagement in productive activities last week=2 1007 conf:(1)

8. Productive work for family=2 Engagement in productive activities last week=2 1007 ==> Productive work for non-parents=2 Did you have job last week=4 1007 conf:(1)

9. Productive work for non-parents=2 Productive work for family=2 Did you have job last week=4 1007 ==> Engagement in productive activities last week=2 1007 conf:(1)

10. Productive work for non-parents=2 Productive work for family=2 Engagement in productive activities last week=2 1007 ==> Did you have job last week=4 1007 conf:(1)

11. Productive work for non-parents=2 Did you have job last week=4 Engagement in productive activities last week=2 1007 ==> Productive work for family=2 1007 conf:(1)

12. Productive work for family=2 Did you have job last week=4 Engagement in productive activities last week=2 1007 ==> Productive work for non-parents=2 1007 conf:(1)

13. Productive work for family=2 Did you have job last week=4 1007 ==> Engagement in productive activities last week=2 1007 conf:(1)

14. Productive work for family=2 Engagement in productive activities last week=2 1007 ==> Did you have job last week=4 1007 conf:(1)

15. Productive work for non-parents=2 Did you have job last week=4 1007 ==> Engagement in productive activities last week=2 1007 conf:(1)

16. Productive work for non-parents=2 Engagement in productive activities last week=2 1007 ==> Did you have job last week=4 1007 conf:(1)

17. Productive work for non-parents=2 Engagement in productive activities last week=2 1007 ==> Productive work for family=2 1007 conf:(1)

18. Productive work for family=2 Engagement in productive activities last week=2 1007 ==> Productive work for non-parents=2 1007 conf:(1)

19. Productive work for non-parents=2 Did you have job last week=4 1007 ==> Productive work for family=2 1007 conf:(1)

20. Productive work for family=2 Did you have job last week=4 1007 ==> Productive work for non-parents=2 1007 conf:(1)

Even though the number of rules generated was doubled, there was not any difference between the attributes used for rule generation in the experiment 4 and experiment 5. So, increasing the number of rules generated did not improve the importance of the rules generated.

The researcher further attempted to generate more relevant set of rules which can give insight about the problem area of child labor. Careful attribute selection is one important step in data mining process to get a better result. Attributes which can give more information on the specific problem area, child labor should be incorporated in the rule mining process. The researcher

attempted to evaluate attribute relevance to the problem of child labor using one of Weka's facilities, attribute selection. The researcher together with the child labor experts selected the attribute engagement in productive activity as an important attribute which represents the child labor problem. The 62 remaining attributes used in experiment 2 up to experiment 5 were evaluated based on their information gain value towards the selected attribute, engagement in productive activities.

Attribute Selection Scheme

=== Run information ===

Evaluator: weka.attributeSelection.InfoGainAttributeEval -M

Search: weka.attributeSelection.Ranker -T -1.7976931348623157E308 -N -1

Relation: k021

Instances: 2398

Attributes: 63

Evaluation mode: evaluate on all training data

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

Search Method: Attribute ranking

Attribute Evaluator (supervised, Class (nominal): 63 Engagement in productive activities):

Information Gain Ranking Filter

Missing values treated as separate

Ranked attributes:

1.0867 61 Engaged in non-productive activity or attending school

0.98452 60 Engagement in productive activities last week

0.97817 24 Employment Status last week

0.97165 21 Did you have job last week

0.96989 20 Below or Above 4 Hours of work last week

0.96591 25 Working Shift

0.96491 19 Total Number of hours worked last week

0.95863 17 Productive work for non-parents

0.95432 18 Productive work for family

0.94445 26 Usual hours of work

0.58476 28 Looking for additional productive activity during last week

0.58476 29 Effort to change your job last week

0.58476 27 Additional productive Activity during last week

0.56424 62 Injuries at work place

0.50566 33 Engaged in productive activities in last year

0.39026 54 Gathering firewood and dung cake

0.35386 46 Engaged in housekeeping activities without payment

0.34787 34 For how many months did you engage in productive activities

0.34319 32 Preparedness for work in coming month

0.34107 30 Effort to find job last 3 months

0.33997 45 Main reason for not attending formal education

0.33496 31 Reason for not seeking job

0.3302 48 Housekeeping

0.32797 53 Shopping

0.31497 56 Message

0.30885 49 Cleaning of the household dwelling

0.30632 57 Fetching water (not for sale)

0.29566 55 Caring for infants

0.29558 59 Didn't work

0.29543 52 Mending washing and pressing clothes

0.2944 50 Preparing meals

0.29385 58 Other

0.29367 51 Serving meals

0.26003 38 Estimated average yearly income rural

0.25348 37 Estimated Average monthly Income Urban

0.24796 3 Town

0.24519 4 Kefetegna

0.23206 41 Type of education or training the child is attending in current academic year

0.21967 42 Grade level in current academic year

0.21331 5 Kebele

0.21309 44 Did the child attend formal education before this academic year

0.21078 43 Did the child attend education or training during last week

0.20783 23 Industry you worked last week

0.18244 22 Your Occupation last week

0.1741 47 Average Number of hours spent in housekeeping activities per day

0.16626 14 Highest Grade

0.16028 13 Read or Write

0.15437 6 Enumeration Area Code

0.15228 12 Ethnic Group

0.11317 1 Zone

0.0929 10 Age

0.08538 2 Wereda

0.0598 16 Martial Status

0.05833 15 Have Training

0.03627 11 Religion

0.0316 8 Relationship to head of household

0.02339 36 Estimated average monthly Consumption

0.02041 7 Serial number selected family

0.01582 40 Has the child been Living with parents since birth

0.01367 9 Sex

0.01307 35 Living outside current town/rural

0.00612 39 children living outside family

In the survey under study, 2001 child labor survey, there were a number of identical questions addressed for the child and for the guardian separately. An example of such questions is engagement of a child in productive activities last week. Both the child and the guardian were asked whether the child has been involved in productive activities last week. And thus, the attribute engagement in productive activities last week with an information gain value of 0.98452 listed above is identical with the attribute selected to represent the problem of child labor. According to the statisticians of FDRE CSA and child labor experts, most of the responses from the children and the guardians in this particular survey are similar. Thus the researcher did not make differentiation between such types of identical attributes.

Based on the relationship of the 62 attributes to the child labor problem, the experts in the area suggested the first 22 attributes as useful. Thus, after transforming the format of these selected attributes together with their values into arff, the researcher fed it to the apriori association algorithm of Weka. The following was the result of the association rule mining:

Experiment 6

=== Run information ===

Scheme: weka.associations.Apriori -N 10 -T 0 -C 0.9 -D 0.05 -U 1.0 -M 0.1 -S -1.0

Relation: k021w

Instances: 2398

Attributes: 22

Productive work for non-parents

Engagement in productive activities last week

Productive work for family

Total Number of hours worked last week

Below or above 4 Hours of work last week

Did you have job last week

Employment Status last week

Working Shift

Usual hours of work

Additional productive Activity during last week

Looking for additional productive activity during last week

Effort to change your job last week

Effort to find job last 3 months

Reason for not seeking job

Preparedness for work in coming month

Engaged in productive activities in last year

For how many months did you engage in productive activities

Main reason for not attending formal education

Engaged in housekeeping activities without payment

Gathering firewood and dung cake

Engaged in non-productive activity or attending school

Injuries at work place

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

Apriori

=======

Minimum support: 0.55

Minimum metric <confidence>: 0.9

Number of cycles performed: 9

Generated sets of large itemsets:

Size of set of large itemsets L(1): Frequency of large itemset 6

Size of set of large itemsets L(2): Frequency of large itemset 9

Size of set of large itemsets L(3): Frequency of large itemset 7

Size of set of large itemsets L(4): Frequency of large itemset 2

Best rules found:

1. Productive work for non-parents=1 Engagement in productive activities last week=1 1383 ==> Below or Above 4 Hours of work last week=1 1383 conf:(1)

2. Productive work for non-parents=1 Working Shift=1 1346 ==> Below or Above 4 Hours of work last week=1 1346 conf:(1)

3. Productive work for non-parents=1 Working Shift=1 Engagement in productive activities last week=1 1345 ==> Below or Above 4 Hours of work last week=1 1345 conf:(1)

4. Employment Status last week=5 1323 ==> Below or Above 4 Hours of work last week=1 1323 conf:(1)

5. Employment Status last week=5 Engagement in productive activities last week=1 1322 ==> Below or Above 4 Hours of work last week=1 1322 conf:(1)

6. Productive work for non-parents=1 Employment Status last week=5 1322 ==> Below or Above 4 Hours of work last week=1 1322 conf:(1)

7. Productive work for non-parents=1 Employment Status last week=5 Engagement in productive activities last week=1 1321 ==> Below or Above 4 Hours of work last week=1 1321 conf:(1)

8. Below or Above 4 Hours of work last week=1 1385 ==> Engagement in productive activities last week=1 1384 conf:(1)

9. Below or Above 4 Hours of work last week=1 1385 ==> Productive work for non-parents=1 1384 conf:(1)

10. Productive work for non-parents=1 Below or Above 4 Hours of work last week=1 1384 ==> Engagement in productive activities last week=1 1383 conf:(1)

Among the ten rules generated above, the following rules were selected for discussion:

3. Productive work for non-parents=1 Working Shift=1 Engagement in productive activities last week=1 1345 ==> Below or Above 4 Hours of work last week=1 1345 confidence: (1)

If a child works for non-parents in the day time working shift and has been engaged in productive activities last week, then the total number of hours he or she worked last week exceeded 4 hours. The support of this rule is 56%.

5. Employment Status last week=5 Engagement in productive activities last week=1 1322 ==> Below or Above 4 Hours of work last week=1 1322 confidence: (1)

If a child has been working as unpaid family worker last week and also engaged in productive activities, then he or she worked for a total working hours of greater than 4 in all kinds of jobs. The support of this rule is 55%.

7. Productive work for non-parents=1 Employment Status last week=5 Engagement in productive activities last week=1 1321 ==> Below or Above 4 Hours of work last week=1 1321 confidence: (1)

If a child has been engaged in productive activity for non-parents and also has been working as unpaid family worker, then the total working hours for this child during last week exceeded 4 hours. The support of this rule is 55%.

10. Productive work for non-parents=1 Below or Above 4 Hours of work last week=1 1384 ==> Engagement in productive activities last week=1 1383 confidence: (1)

If a child has been engaged in productive activity for non-parents and the total working hours of the child last week is greater than 4, then it is true that the child generally involved in productive activities last week. The attribute engagement in productive activities last week includes both productive work for non-parents and productive work for parents performed by a child during last week. The support of this rule is 58%.

The rules generated in this experiment revolve around working children: both domestic and productive work. The strong relationship indicated by these rules is among the attributes engagement in productive activities last week, productive work for non-parents, working shift, below or above 4 hours of work last week and employment status last week.

According to the generated rules, most of the children working as unpaid family workers also are engaged in productive activities for non-parents and they work in the daytime working shift. As the minimum support level tells us the generated rules are applied for more than half of the children in the area. The child labor experts found these rules to be useful and interesting. They

indicated that according to these generated rules, there is indeed high rate of child labor in *Affar killil*.

The researcher believes that several alternatives should be exhaustively considered to improve this result. Because of serious time shortage, the researcher attempted only the next alternative, increase the number of rules to be generated from 10 to 20. The result obtained was as follows:

Experiment 7

=== Run information ===

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

Relation: k021w

Instances: 2398

Attributes: 22 (The same list of attributes as experiment 6)

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

Apriori

=======

Minimum support: 0.55

Minimum metric <confidence>: 0.9

Number of cycles performed: 9

Generated sets of large itemsets:

Size of set of large itemsets L(1): Frequency of large itemset 6

Size of set of large itemsets L(2): Frequency of large itemset 9

Size of set of large itemsets L(3): Frequency of large itemset 7

Size of set of large itemsets L(4): Frequency of large itemset 2

Best rules found:

1. Productive work for non-parents=1 Engagement in productive activities last week=1 1383 ==> Below or Above 4 Hours of work last week=1 1383 conf:(1)

2. Productive work for non-parents=1 Working Shift=1 1346 ==> Below or Above 4 Hours of work last week=1 1346 conf:(1)

3. Productive work for non-parents=1 Working Shift=1 Engagement in productive activities last week=1 1345 ==> Below or Above 4 Hours of work last week=1 1345 conf:(1)

4. Employment Status last week=5 1323 ==> Below or Above 4 Hours of work last week=1 1323 conf:(1)

5. Employment Status last week=5 Engagement in productive activities last week=1 1322 ==> Below or Above 4 Hours of work last week=1 1322 conf:(1)

6. Productive work for non-parents=1 Employment Status last week=5 1322 ==> Below or Above 4 Hours of work last week=1 1322 conf:(1)

7. Productive work for non-parents=1 Employment Status last week=5 Engagement in productive activities last week=1 1321 ==> Below or Above 4 Hours of work last week=1 1321 conf:(1)

8. Below or Above 4 Hours of work last week=1 1385 ==> Engagement in productive activities last week=1 1384 conf:(1)

9. Below or Above 4 Hours of work last week=1 1385 ==> Productive work for non-parents=1 1384 conf:(1)

10. Productive work for non-parents=1 Below or Above 4 Hours of work last week=1 1384 ==> Engagement in productive activities last week=1 1383 conf:(1)

11. Below or Above 4 Hours of work last week=1 Engagement in productive activities last week=1 1384 ==> Productive work for non-parents=1 1383 conf:(1)

12. Working Shift=1 1349 ==> Engagement in productive activities last week=1 1348 conf:(1)

13. Below or Above 4 Hours of work last week=1 Working Shift=1 1347 ==> Engagement in productive activities last week=1 1346 conf:(1)

14. Below or Above 4 Hours of work last week=1 Working Shift=1 1347 ==> Productive work for non-parents=1 1346 conf:(1)

15. Productive work for non-parents=1 Working Shift=1 1346 ==> Below or Above 4 Hours of work last week=1 Engagement in productive activities last week=1 1345 conf:(1)

16. Productive work for non-parents=1 Below or Above 4 Hours of work last week=1 Working Shift=1 1346 ==> Engagement in productive activities last week=1 1345 conf:(1)

17. Below or Above 4 Hours of work last week=1 Working Shift=1 Engagement in productive activities last week=1 1346 ==> Productive work for non-parents=1 1345 conf:(1)

18. Productive work for non-parents=1 Working Shift=1 1346 ==> Engagement in productive activities last week=1 1345 conf:(1)

19. Employment Status last week=5 1323 ==> Below or Above 4 Hours of work last week=1 Engagement in productive activities last week=1 1322 conf:(1)

20. Below or Above 4 Hours of work last week=1 Employment Status last week=5 1323 ==> Engagement in productive activities last week=1 1322 conf:(1)

Even though the number of attributes was increased from 10 to 20, there was not change in the types of rules generated.

Based on the evaluation of the child labor experts, the rules generated above using different attributes have practical relevance to the problem of child labor because they did give some additional insight about the problem of child labor. A problem area should be first fully known before rushing to develop solutions. Therefore the result of this research can be used as input for developing programs and strategies for prevention of intolerable form of child labor and protection of working children.

Chapter 5

Conclusion and Recommendation

5.1 Conclusion

This research attempted to study the application of association rule mining to find relationship and interesting patterns between attributes of census or survey data. The particular survey data taken for the research is 2001 stand-alone child labor survey. The study was conducted based on the data mining steps or process discussed in chapter 2: defining the data mining goal, data collection, data quality verification, data selection, data cleaning and preparation for model building, model building and evaluation. However, since a data mining task is an iterative process, these steps were not followed strictly in linear order.

As TCC (1999) discusses there are two keys to success in data mining. First is coming up with a precise formulation of the problem to be solved by the data mining technology. As the authors say, “A focused statement usually results in the best payoff”. The second key is using the right data.

Data collection, selection and cleaning were major tasks which took most of the experimental time of the research. This is due to higher volume or size of the data residing in CSA database in general and also of the target dataset. EM clustering algorithm was used to segment the selected dataset into groups of similar records. The total number of attributes in the original target dataset is 361, and it is hardly possible to analyze all of these attributes especially with such short period of time given for the research. The researcher attempted to select relevant attributes for the data

mining task. Statistical and child labor experts' advice was used to identify such relevant attributes. In addition, the characteristic of the algorithm and the software was also considered in selecting attributes.

A dataset totaling 2398 records was used in generating association rules. Initially, all of the records were given with 86 attributes to apriori, and the first 10 best rules were generated. These rules have a minimum coverage or support of 90% and minimum accuracy of confidence of 95%. In the second round the number of selected attributes was reduced to 63 to generate the first 10 best rules. At this time the minimum support level was 80% and minimum confidence level was 90%.

After evaluating the rules together with domain experts, the researcher applied clustering algorithm on the dataset in order to further refine the result. The clustering algorithm, expectation maximization, was run using the 63 attributes used in experiment 2. A total of four clustering models were built by varying the number of clusters from 2 up to 5. The cluster model, which according to the domain experts made good sense about child labor, segmented the records into five clusters. Among these five clusters, the third cluster which contains 42.5% of the selected dataset was chosen and given to the apriori algorithm of Weka. Cluster number 3 was selected because it was recommended by domain expert and it has higher number of instances than the remaining clusters. The association rule algorithm, apriori, generated its 10 best association rules with minimum coverage of 95% and minimum accuracy of 90%. Based on the evaluation given by domain experts on these rules, it was found out that the application of clustering algorithm for

data preparation can significantly improve the relevance of the rules to be generated for the defined problem area.

The researcher further attempted to improve the result by using attribute selection scheme of Weka software. The attribute selector did select attributes with higher information gain with respect to the defined class. The minimum support level was 55% and minimum confidence was 90%. This means the generated rules were applicable more than half of the instances involved.

In general, the results from this study were encouraging. It was possible to segment child labor survey data using data mining techniques that made good meanings to domain experts. Besides, these clusters significantly helped to generate relevant rules for the problem domain. It is the researcher's belief that a more thorough study using data mining techniques can help to understand more about child labor problem in Ethiopia.

5.2 Recommendations

This research work is conducted mainly for academic purpose. However, the researcher highly believes that the findings of this research project can be used by concerned organizations to further investigate the nature of child labor problem in Ethiopia.

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

There were a number of missing values in the database of 2001 child labor survey because a number of questions on the questionnaire were not applicable to respondents. The researcher

simply eliminates attributes with 90% and above missing value. It is possible that the trend of these missing values may indicate important pattern in the database. The researcher could not attempt this alternative mainly due to shortage of time.

As indicated in chapter 4, the result of the initial experiment was rated as poor by the domain experts. It is the researcher's belief that the main reason for having such types of poor rules is low or primitive level of abstraction used for data representation. All those property types making up the rules can be generalized into a higher level concept or abstraction, household living standard. As Han and Kamber (2001) put it strong associations discovered at high concept levels may represent common sense knowledge. However, what may represent common sense to one user may seem novel to another. Therefore, data mining systems should provide capabilities to mine association rules at multiple levels of abstraction and traverse easily among different abstraction spaces. So, future research should consider mining association rules at a level of abstraction which is appropriate for a specific problem being addressed by the data mining process. Here also availability of time was the constraint for not attempting the suggested alternative.

The researcher feels that the number of experiments undertaken in this research project is not enough to have a complete conclusion about the application of association rule mining on census data such as the child labor survey of this research project. Future research work on this area should consider attempting a number of alternatives to generate meaningful associations between the attributes of the database. The associations to be generated also should help in eliminating the

worst form of child labor and protecting working children in Ethiopia. In the long run the rules also should help in completely eliminating child labor problem from Ethiopia.

Another future work is to test the applicability of other association rule mining algorithms and software for mining rules from census or survey data and compare the results. One weakness of the apriori association rule algorithm is inability to handle numeric data. The researcher transformed numeric attributes into nominal by listing their possible values. So, other algorithms which can perform more efficiently and effectively than apriori algorithm and Weka software should be investigated and applied.

There are a number of techniques used to enhance apriori algorithm or association rule algorithms in general. In this research project the apriori algorithm was applied directly as it is implemented in Weka, without any adjustment to improve its performance. Thus, it is important to investigate techniques of improving apriori efficiency in future research work.

As Han and Kamber (2001) mentions, association rules can also be applied for classification purpose. It is recommended that classification based on concepts for association rule be investigated in future research works.

References

Berry, Michael J. A. and Linoff, Gordon. 1997. Data Mining Techniques: for Marketing, Sales and Customer support. New York: John Wiley & Sons, Inc.

Bigus, Joseph P. 1996. Data Mining with Neural Networks: Solving Business Problems from Application Development to Decision Support. McGraw-Hill: New York.

Cabena, P., et. al. 1998. Discovering Data Mining – From concept to Implementation, Prentice Hall, New Jersey.

Collier, Ken and et. al. 1999. A Methodology for Evaluating and Selecting Data Mining Software. Available at: http://www.computer.org/proceedings/hicss/0001/00016/000160.09ab8.htm

Deogun, Jitender S. 2001. Data Mining: research Trends, Challenges, and Applications. Available URL:http://citeseer.nj.nec.com/deogun97data.html

Fayyad, Usma, Piatetsky-shapiro, G. and Smyth, Padharic. 1996. From Data Mining to Knowledge Discovery in Databases. Available URL: http://citeseer.nj.nec.com.fayyad96from.html

FDRE Central Statistical Authority. 2001. Ethiopia Child Labor Survey Report. Addis Ababa, Ethiopia.

FDRE CSA. 1999. Statistical Report on the 1999 National Labor Force Survey. Addis Ababa, Ethiopia.

Han, Jiawei and Kamber, Micheline. 2001. Data Mining: Concepts and Techniques and Applications. San Fransisico; Morgan Kufman Publishers.

Han, Eui-Hung and et. al., n.d. Scalable Parallel Data Mining for Association Rules. Internet Source.

Hipp, Jochen and et. al., 2000. Algorithms for Association Rule Mining: A General Survey and Comparison. Available at: http://www.cs.sfu.ca/coursecentral/884/G2/2002-3/references/high00.pdf

Hipp, Jochen and et. al., n.d. Integrating Association Rule Mining Algorithms with Relational Database Systems. Available at: http://www-db.informatik.uni-tuebingen.de/forschung/papers/iceisol.pdf

J.Zaki, Mohammed and Ogihara, Mitsunuri. N.d. Theoretical Foundations of Association Rules. Available at:

http://fano.ics.uci.edu/cites/documents/theoretical-foundations-of association-rules.html

J.Zaki, Mohammed. 1999. Parallel and Distributed Association Mining: A Survey Available at: http://www.cs.rpi.edu/~Zaki/ps/concurrency.pdf

Levin, Nissan and Zahavi, Jacob, 1999. Data Mining. Available URL: www.urbanscience.com/Data Mining.pdf

LIS-Rudjet Boskovic Institute. 2001. Data Mining Methodology. Available at: http://dms.irg.hr/tutorial/tut.intro.php

Malerba, Donato and et. al., n.d. Mining Spatial Association Rules in census data: A Relational Approach. Available at: http://www.colorado.edu/geography/babs/geog-6180-002-s03/bibliography/Malerba.pdf

Plate, Tony et. al. 1997. A comparison between neural networks and other statistical techniques for modeling the relationship between tobacco and alcohol and cancer. http://citeseer.jn.nec.com/plate96comparison.html

Raghavan, Bijay, Deogun, Jitender S. and Sover Mayri, 2002. Data Mining: Trends and Issues. Available URL:http://citeseer.nj.nec.com/138316.html

Raghavan, Vijav V., et al. 1998. A Perspective on Data Mining. Journal of the American Society for Information Science; 49(5): 397-402.

Rea, Allan. 2001. Data Mining: an introduction Student Notes. Available URL: http://www.pcc.qub.ac.uk/tec/courses/datamining/stu_notes/dm_book_1.html

Statsoft, Inc. 2003. Data Mining Techniques. Available at: http://www.statsoftinc.com/textbook/stdatmin.html#concepts

Srikant, Ramakrishnan and et. al. Mining Association Rules with Item Constraints. Available at: http://citeseer.nj.nec.com/761.html

Srikant, Ramakrishnan and Agrawal, Rakesh. n.d. Mining Generalized Association Rules. Available at: http://citeseer.nj.nec.com/srikan95mining.html

Trybula, Walter J. 1997. Data Mining and Knowledge Discovery: Annual review of Information Science and Technology (ARIST); (32): 197 – 229.

Two Crows Corporation. 1999. Introduction to Data Mining and Knowledge Discovery. Available URL: http://www.twocrows.com

Witten, Ian H. and Frank, Eibe. 2000. Practical Machine Learning Tools and Techniques with Java Implementations. USA: Academic Press.

Appendices

Appendix One

Original list of attributes

Sequence number of attributes	Attribute Name	Data Type	Description
1	Survey No.	Numeric	Serves as identification number for the survey
2	Killil	Nominal	Current address of the person responding the specific question
3	Zone	Nominal	Current address of the person responding the specific question
4	Wereda	Numeric	Current address of the person responding the specific question
5	Town	Nominal	Current address of the person responding the specific question
6	Kefetegna	Numeric	Current address of the person responding the specific question
7	Kebele	Numeric	Current address of the person responding the specific question
8	Enumeration Area code	Numeric	Code given to the area from which the data is collected
9	Selection Number of selected family	Numeric	Number assigned for each selected family for identification purpose
10	Selection Number of Respondent	Numeric	Number assigned for each respondent

11	**Serial Number of Household Members**	**Numeric**	**Number assigned for each selected family member**
12	**Relationship to head of household**	**Nominal**	**Relationship of the household member to the head of the household**
13	**Sex**	**Nominal**	**Sex of respondent**
14	**Age**	**Numeric**	**Age of respondent**
15	**Religion**	**Nominal**	**Religion of respondent**
16	**Ethnic Group**	**Nominal**	**Ethnic of respondent**
17	**Can you read/write**	**Nominal**	**Educational status of the respondent (addressed for household members whose age is greater than or equal to five)**
18	**The highest grade completed**	**Nominal**	**Educational status of the respondent (addressed for household members whose age is greater than or equal to five)**
19	**Have training?**	**Nominal**	**Training condition (addressed for household members whose age is greater than or equal to ten)**
20	**Type of training**	**Nominal**	**Training condition (addressed for household members whose age is greater than or equal to ten)**
21	**Marital Status**	**Nominal**	**(addressed for household members whose age is greater than or equal to ten)**
22	**Productive work for non-parents last week?**	**Nominal**	**Addressed for household members whose age is greater than or equal to five.**
23	**Productive work for**	**Nominal**	**Addressed for**

	family last week?		**household members whose age is greater than or equal to five**
24	**Total number of hours worked last week**	**Numeric**	**Addressed for household members whose age is greater than or equal to five**
25	**Is the total working hour during last week greater or less than 4 hours?**		
26	**Did you have another job last week?**	**Nominal**	**Addressed for household members whose age is greater than or equal to five**
27	**Occupation you have been working during last week?**	**Nominal**	**Addressed for household members whose age is greater than or equal to five**
28	**Industry you have been working during last week?**	**Nominal**	**Addressed for household members whose age is greater than or equal to five**
29	**Employment status in major occupation last week**	**Nominal**	**Addressed for household members whose age is greater than or equal to five**
30	**Payment period last week**	**Nominal**	**Addressed for household members whose age is greater than or equal to five**
31	**Payment type (cash or kind)**	**Nominal**	**Addressed for household members whose age is greater than or equal to five**
32	**Recent cash payment amount**	**Numeric**	**Addressed for household members whose age is greater than or equal to five**
33	**Recent paid amount in kind**	**Numeric**	**Addressed for household members whose age is greater than or equal to five**
34	**Recent total payment amount**	**Numeric**	**Addressed for household members whose age is greater**

			than or equal to five
35	**Working hour shift**	**Nominal**	**Addressed for household members whose age is greater than or equal to five**
36	**Usual hours of work**	**Numeric**	**Addressed for household members whose age is greater than or equal to five**
37	**Additional productive activity to your major activity last week**	**Nominal**	**Addressed for household members whose age is greater than or equal to ten**
38	**Looking for additional productive activity last week**	**Nominal**	**Addressed for household members whose age is greater than or equal to ten**
39	**Attempt to change your job last week**	**Nominal**	**Addressed for household members whose age is greater than or equal to ten**
40	**Three months attempt to find job**	**Nominal**	**Addressed for household members whose age is greater than or equal to ten**
41	**Reason for unemployment**	**Nominal**	**Addressed for household members whose age is greater than or equal to ten**
42	**Preparedness for work in coming one month**	**Nominal**	**Addressed for household members whose age is greater than or equal to ten**
43	**Productive activities for the last 12 months**	**Nominal**	**Addressed for household members whose age is greater than or equal to ten**
44	**For how long did you engage in productive activities for last year (in months)?**	**numeric**	**Addressed for household members whose age is greater than or equal to ten**
45	**Major occupation you worked for last year?**	**Nominal**	**Addressed for household members whose age is greater than or equal to**

			eighteen
46	**Major industry you worked for last year?**	**Nominal**	**Addressed for household members whose age is greater than or equal to eighteen**
47	**Reason for unemployment during last year?**	**Nominal**	**Addressed for household members whose age is greater than or equal to eighteen**
48	**Is the child eligible for the study?**	**Nominal**	**Eligible child is a child whose age is between 5 and 17**
49	**Has the household been living outside the current town/rural?**	**Nominal**	**Addressed to the household head**
50	**Previous killil in which the household been living?**	**Nominal**	**Addressed to the household head**
51	**Previous zone in which the household been living?**	**Nominal**	**Addressed to the household head**
52	**Was the previous household residence rural of town?**	**Nominal**	**Addressed to the household head**
53	**Reason for changing to the present place of resident**	**Nominal**	**Addressed to the household head**
54	**How long had this household been living in the present resident?**	**Nominal**	**Addressed to the household head**
55	**Ownership status of the household dwelling**	**Nominal**	**Addressed to the household head**
56	**Monthly amount of rent in birr**	**Numeric**	**Addressed to the household head**
57	**Major materials of house wall construction**	**Nominal**	**Addressed to the household head**
58	**Major materials of house roof construction**	**Nominal**	**Addressed to the household head**
59	**Number of rooms in**	**Nominal**	**Addressed to the**

	the house		**household head**
60	**Toilet facility**	**Nominal**	**Addressed to the household head**
61	**Type of Kitchen**	**Nominal**	**Addressed to the household head**
62	**Source of drinking water**	**Nominal**	**Addressed to the household head**
63	**Type of fuel**	**Nominal**	**Addressed to the household head**
64	**Properties owned by household-own a house?**	**Nominal**	**Addressed to the household head**
65	**Have radio?**	**Nominal**	**Addressed to the household head**
66	**Have television?**	**Nominal**	**Addressed to the household head**
67	**Have telephone?**	**Nominal**	**Addressed to the household head**
68	**Have electric mitad?**	**Nominal**	**Addressed to the household head**
69	**Have refrigerator?**	**Nominal**	**Addressed to the household head**
70	**Have bofe and sofa?**	**Nominal**	**Addressed to the household head**
71	**Have car?**	**Nominal**	**Addressed to the household head**
72	**Have table and chair?**	**Nominal**	**Addressed to the household head**
73	**Have cropland?**	**Nominal**	**Addressed to the household head**
74	**Have cattle?**	**Nominal**	**Addressed to the household head**
75	**Have horse/mule/donkey?**	**Nominal**	**Addressed to the household head**
76	**Have camel?**	**Nominal**	**Addressed to the household head**
77	**Have sheep/goats?**	**Nominal**	**Addressed to the household head**
78	**Grow cash crops?**	**Nominal**	**Addressed to the household head**
79	**Estimated average monthly household consumption in cash**	**Nominal**	**Addressed to the household head**
80	**Estimated average monthly income of household(urban)**	**Nominal**	**Addressed to the household head**

81	**Estimated average yearly income of household (rural)**	**Nominal**	**Addressed to the household head**
82	**Are there any children who live other place than their family?**	**Nominal**	**Addressed to the child's parents or guardians**
83	**How many male live outside family?**	**Numeric**	**Addressed to the child's parents or guardians**
84	**How many female live outside family?**		
85	**Age of elder child living outside family**	**Numeric**	**Addressed to the child's parents or guardians**
86	**Age of next elder child living outside family**	**numeric**	**Addressed to the child's parents or guardians**
87	**Age of younger child living outside family**	**Nominal**	**Addressed to the child's parents or guardians**
88	**Sex of elder child living outside family**	**Nominal**	**Addressed to the child's parents or guardians**
89	**Sex of next elder child living outside family**	**Nominal**	**Addressed to the child's parents or guardians**
90	**Sex of younger child living outside family**	**Nominal**	**Addressed to the child's parents or guardians**
91	**With whom is the elder child living?**	**Nominal**	**Addressed to the child's parents or guardians**
92	**With whom is the next elder child living?**	**Nominal**	**Addressed to the child's parents or guardians**
93	**With whom is the younger child living?**	**Nominal**	**Addressed to the child's parents or guardians**
94	**Living condition of the elder child living apart from family**	**Nominal**	**Addressed to the child's parents or guardians**
95	**Living condition of the next elder child living apart from**	**Nominal**	**Addressed to the child's parents or guardians**

	family		
96	**Living condition of the younger child living apart from family**	**Nominal**	**Addressed to the child's parents or guardians**
97	**Has the child been living with present household since birth?**	**Nominal**	**Addressed to the child's parents or guardians**
98	**Last place of usual child residence before coming here? *Killil***	**Nominal**	**Addressed to the child's parents or guardians**
99	**Previous zone**	**Nominal**	**Addressed to the child's parents or guardians**
100	**Was the last place of residence rural or urban?**	**Nominal**	**Addressed to the child's parents or guardians**
101	**Child's activity before coming to this household**	**Nominal**	**Addressed to the child's parents or guardians**
102	**Main reason for coming to present household?**	**Nominal**	**Addressed to the child's parents or guardians**
103	**How long has the child been living with the present household?**	**Nominal**	**Addressed to the child's parents or guardians**
104	**What type of education of training institution does the child attend in this academic year?**	**Nominal**	**Addressed to the child's parents or guardians**
105	**The grade level the child is currently attending**	**Nominal**	**Addressed to the child's parents or guardians**
106	**Did the child attend school or training institution during last week?**	**Nominal**	**Addressed to the child's parents or guardians**
107	**Did the child attend formal education before this academic**	**Nominal**	**Addressed to the child's parents or guardians**

	year?		
108	**Main reason for not attending formal education?**	**Nominal**	**Addressed to the child's parents or guardians**
109	**Engagement in housekeeping activities without payment**	**Nominal**	**Addressed to the child's parents or guardians**
110	**Average number of working hours per day**	**Numeric**	**Addressed to the child's parents or guardians**
111	**Main housekeeping activities last week-Housekeeping**	**Nominal**	**Addressed to the child's parents or guardians**
112	**Cleaning of the household dwelling**	**Nominal**	**Addressed to the child's parents or guardians**
113	**Preparing meals**	**Nominal**	**Addressed to the child's parents or guardians**
114	**Serving meals**	**Nominal**	**Addressed to the child's parents or guardians**
115	**Mending, washing and pressing clothes**	**Nominal**	**Addressed to the child's parents or guardians**
116	**Shopping**	**Nominal**	**Addressed to the child's parents or guardians**
117	**Gathering firewood and dong cake**	**Nominal**	**Addressed to the child's parents or guardians**
118	**Caring for infants**	**Nominal**	**Addressed to the child's parents or guardians**
119	**Message**	**Nominal**	**Addressed to the child's parents or guardians**
120	**Fetching water**	**Nominal**	**Addressed to the child's parents or guardians**
121	**Other**	**Nominal**	**Addressed to the child's parents or guardians**
122	**Didn't work**	**Nominal**	**Addressed to the child's parents or**

			guardians
123	**Productive activity of child last week (guardians)**	**Nominal**	**Addressed to the child's parents or guardians**
124	**Did the child do non productive activities or attend school last week?**	**Nominal**	**Addressed to the child's parents or guardians**
125	**Main reason for not attending school and not doing any domestic work**	**Nominal**	**Addressed to the child's parents or guardians**
126	**Injury or illness at work place**	**Nominal**	**Addressed to the child's parents or guardians**
127	**How often was the child hurt**	**Nominal**	**Addressed to the child's parents or guardians**
128	**Industry in which serious injury occurred**	**Nominal**	**Addressed to the child's parents or guardians**
129	**Occupations in which serious damage occurred**	**Nominal**	**Addressed to the child's parents or guardians**
130	**Types of injuries/illness-Type1**	**Nominal**	**Addressed to the child's parents or guardians**
131	**Type2**	**Nominal**	**Addressed to the child's parents or guardians**
132	**Type3**	**Nominal**	**Addressed to the child's parents or guardians**
133	**Seriousness of the accident**	**Nominal**	**Addressed to the child's parents or guardians**
134	**Type of treatment received**	**Nominal**	**Addressed to the child's parents or guardians**
135	**Number of days hospitalized**	**Numeric**	**Addressed to the child's parents or guardians**
136	**Where did the child consult health personnel? Work place**	**Nominal**	**Addressed to the child's parents or guardians**

137	**Hospital/health station**	**Nominal**	**Addressed to the child's parents or guardians**
138	**Pharmacy**	**Nominal**	**Addressed to the child's parents or guardians**
139	**Clinic**	**Nominal**	**Addressed to the child's parents or guardians**
140	**Other place**	**Nominal**	**Addressed to the child's parents or guardians**
141	**Who paid most part of medical treatment cost? Parents**	**Nominal**	**Addressed to the child's parents or guardians**
142	**Employer**	**Nominal**	**Addressed to the child's parents or guardians**
143	**The child**	**Nominal**	**Addressed to the child's parents or guardians**
144	**Free medical service**	**Nominal**	**Addressed to the child's parents or guardians**
145	**No need for payment**	**Nominal**	**Addressed to the child's parents or guardians**
146	**Other**	**Nominal**	**Addressed to the child's parents or guardians**
147	**Did the child do productive work last week(guardians)**	**Nominal**	**Addressed to the child's parents or guardians**
148	**Does the child use protective wears while working-Glasses**	**Nominal**	**Addressed to the child's parents or guardians**
149	**Helmet**	**Nominal**	**Addressed to the child's parents or guardians**
150	**Earplugs**	**Nominal**	**Addressed to the child's parents or guardians**
151	**Special shoes**	**Nominal**	**Addressed to the child's parents or guardians**

152	**Glove**	**Nominal**	**Addressed to the child's parents or guardians**
153	**Other**	**Nominal**	**Addressed to the child's parents or guardians**
154	**Don't use**	**Nominal**	**Addressed to the child's parents or guardians**
155	**Do other people doing the same work use protective wear while working?**	**Nominal**	**Addressed to the child's parents or guardians**
156	**Type of protective wears they use-Glasses**	**Nominal**	**Addressed to the child's parents or guardians**
157	**Helmet**	**Nominal**	**Addressed to the child's parents or guardians**
158	**Earplugs**	**Nominal**	**Addressed to the child's parents or guardians**
159	**Special shoes**	**Nominal**	**Addressed to the child's parents or guardians**
160	**Gloves**	**Nominal**	**Addressed to the child's parents or guardians**
161	**Other**	**Nominal**	**Addressed to the child's parents or guardians**
162	**Awareness of health problem in connection with work**	**Nominal**	**Addressed to the child's parents or guardians**
163	**For whom is the child currently working?**	**Nominal**	**Addressed to the child's parents or guardians**
164	**Relationship with employer**	**Nominal**	**Addressed to the child's parents or guardians**
165	**Reasons for bad relationship with employer. Work overload**	**Nominal**	**Addressed to the child's parents or guardians**
166	**Long hours of work**	**Nominal**	**Addressed to the**

			child's parents or guardians
167	**Small amount of payment**	**Nominal**	**Addressed to the child's parents or guardians**
168	**Payment is not made on time**	**Nominal**	**Addressed to the child's parents or guardians**
169	**Physical abuse**	**Nominal**	**Addressed to the child's parents or guardians**
170	**Verbal abuse**	**Nominal**	**Addressed to the child's parents or guardians**
171	**Other**	**Nominal**	**Addressed to the child's parents or guardians**
172	**Benefits given by employer-Paid holidays or sick leave**	**Nominal**	**Addressed to the child's parents or guardians**
173	**Social security insurance**	**Nominal**	**Addressed to the child's parents or guardians**
174	**Regular bonus**	**Nominal**	**Addressed to the child's parents or guardians**
175	**Free or subsidized uniform**	**Nominal**	**Addressed to the child's parents or guardians**
176	**Free meals**	**Nominal**	**Addressed to the child's parents or guardians**
177	**Subsidized meals**	**Nominal**	**Addressed to the child's parents or guardians**
178	**Free or subsidized transport**	**Nominal**	**Addressed to the child's parents or guardians**
179	**Free or subsidized lodging**	**Nominal**	**Addressed to the child's parents or guardians**
180	**Others**	**Nominal**	**Addressed to the child's parents or guardians**
181	**Don't give any**	**Nominal**	**Addressed to the**

	benefit		**child's parents or guardians**
182	**Don't know**	**Nominal**	**Addressed to the child's parents or guardians**
183	**Consequence of quitting job to the family**	**Nominal**	**Addressed to the child's parents or guardians**
184	**What would the family prefer the child to do currently?**	**Nominal**	**Addressed to the child's parents or guardians**
185	**What would the family prefer the child to do in the future?**	**Nominal**	**Addressed to the child's parents or guardians**
186	**What does the child do for fun? Playing with friends**	**Nominal**	**Addressed to the child's parents or guardians**
187	**Watching TV/Video/Listening radio/reading**	**Nominal**	**Addressed to the child's parents or guardians**
188	**Studying**	**Nominal**	**Addressed to the child's parents or guardians**
189	**Asking relatives**	**Nominal**	**Addressed to the child's parents or guardians**
190	**Others**	**Nominal**	**Addressed to the child's parents or guardians**
191	**Main reason forced a child to work**	**Nominal**	**Addressed to the child's parents or guardians**
192	**Age when starting work for first time**	**Numeric**	**Addressed to the child's parents or guardians**
193	**Serial number of person answering form II**	**Numeric**	**Addressed to the child's parents or guardians**
194	**Are you (the child) attending school or training institution in the current academic year?**	**Nominal**	**Addressed to eligible child for the study (whose age is between 5 to 17)**
195	**What type of school**	**Nominal**	**Addressed to**

	or training institution are you (the child) attending in the current academic year?		**eligible child for the study (whose age is between 5 to 17)**
196	**The grade level you (the child) is attending currently**	**Nominal**	**Addressed to eligible child for the study (whose age is between 5 to 17)**
197	**Education attendance during last week**	**Nominal**	**Addressed to eligible child for the study (whose age is between 5 to 17)**
198	**Did the child attend vocational/training institution during last week**	**Nominal**	**Addressed to eligible child for the study (whose age is between 5 to 17)**
199	**Did the child attend formal education or vocational or training before this academic year?**	**Nominal**	**Addressed to eligible child for the study (whose age is between 5 to 17)**
200	**Reason for not attending formal education or training institution**	**Nominal**	**Addressed to eligible child for the study (whose age is between 5 to 17)**
201	**Reason for not attending school during last week**	**Nominal**	**Addressed to eligible child for the study (whose age is between 5 to 17)**
202	**Were you engaged in economic or non-economic activity during last week**	**Nominal**	**Addressed to eligible child for the study (whose age is between 5 to 17)**
203	**Number of hours worked yesterday**	**Numeric**	**Addressed to eligible child for the study (whose age is between 5 to 17)**
204	**Did the child attend formal or vocational education or training last week?(the child is asked)**	**Nominal**	**Addressed to eligible child for the study (whose age is between 5 to 17)**
205	**Effect of work on regular attendance**	**Nominal**	**Addressed to eligible child for the**

	of education or study		**study (whose age is between 5 to 17)**
206	**For whom is the child currently working? (the child is asked)**	**Nominal**	**Addressed to eligible child for the study (whose age is between 5 to 17)**
207	**Is there overtime time work and payment for the work?**	**Nominal**	**Addressed to eligible child for the study (whose age is between 5 to 17)**
208	**Relationship with your (the child) employer**	**Nominal**	**Addressed to eligible child for the study (whose age is between 5 to 17)**
209	**Reason for bad relationship with employer (the child is asked) Wants too much work**	**Nominal**	**Addressed to eligible child for the study (whose age is between 5 to 17)**
210	**Wants work done for long hours**	**Nominal**	**Addressed to eligible child for the study (whose age is between 5 to 17)**
211	**Pays poorly**	**Nominal**	**Addressed to eligible child for the study (whose age is between 5 to 17)**
212	**Does not pay on time**	**Nominal**	**Addressed to eligible child for the study (whose age is between 5 to 17)**
213	**Abuses physically**	**Nominal**	**Addressed to eligible child for the study (whose age is between 5 to 17)**
214	**Abuses verbally**	**Nominal**	**Addressed to eligible child for the study (whose age is between 5 to 17)**
215	**Other**	**Nominal**	**Addressed to eligible child for the study (whose age is between 5 to 17)**
216	**If you (the child) are currently in paid employment, what is**	**Nominal**	**Addressed to eligible child for the study (whose age is**

	the term of payment?		**between 5 to 17)**
217	**Did you (the child) receive fair payment?**	**Nominal**	**Addressed to eligible child for the study (whose age is between 5 to 17)**
218	**Amount of your (the child) last salary payment in cash**	**Numeric**	**Addressed to eligible child for the study (whose age is between 5 to 17)**
219	**Amount of your (the child) last salary payment in kind**	**Numeric**	**Addressed to eligible child for the study (whose age is between 5 to 17)**
220	**Last total earning amount in cash and in kind**	**Numeric**	**Addressed to eligible child for the study (whose age is between 5 to 17)**
221	**Contribution to parents or others you (the child) resides with**	**Nominal**	**Addressed to eligible child for the study (whose age is between 5 to 17)**
222	**Saving**	**Nominal**	**Addressed to eligible child for the study (whose age is between 5 to 17)**
223	**How do you (the child) save**	**Nominal**	**Addressed to eligible child for the study (whose age is between 5 to 17)**
224	**Reason for saving**	**Nominal**	**Addressed to eligible child for the study (whose age is between 5 to 17)**
225	**Present job satisfaction**	**Nominal**	**Addressed to eligible child for the study (whose age is between 5 to 17)**
226	**Reason for present job dissatisfaction**	**Nominal**	**Addressed to eligible child for the study (whose age is between 5 to 17)**
227	**Have you (the child) been injured at work place?**	**Nominal**	**Addressed to eligible child for the study (whose age is between 5 to 17)**

228	**Type of injury**	**Nominal**	**Addressed to eligible child for the study (whose age is between 5 to 17)**
229	**How serious was the injury**	**Nominal**	**Addressed to eligible child for the study (whose age is between 5 to 17)**
230	**What type of treatment did you (the child) get?**	**Nominal**	**Addressed to eligible child for the study (whose age is between 5 to 17)**
231	**Number of days hospitalized**	**Numeric**	**Addressed to eligible child for the study (whose age is between 5 to 17)**
232	**Were you (the child) engaged in economic or non-economic activity last week**	**Nominal**	**Addressed to eligible child for the study (whose age is between 5 to 17)**
233	**Are you (the child) required to operate any tools or equipment?**	**Nominal**	**Addressed to eligible child for the study (whose age is between 5 to 17)**
234	**Use of protective wears while working-glasses**	**Nominal**	**Addressed to eligible child for the study (whose age is between 5 to 17)**
235	**Helmet**	**Nominal**	**Addressed to eligible child for the study (whose age is between 5 to 17)**
236	**Earplugs**	**Nominal**	**Addressed to eligible child for the study (whose age is between 5 to 17)**
237	**Special shoes**	**Nominal**	**Addressed to eligible child for the study (whose age is between 5 to 17)**
238	**Gloves**	**Nominal**	**Addressed to eligible child for the study (whose age is between 5 to 17)**
239	**Other**	**Nominal**	**Addressed to eligible child for the**

			study (whose age is between 5 to 17)
240	**Don't use**	**Nominal**	**Addressed to eligible child for the study (whose age is between 5 to 17)**
241	**Do other people doing the same work use protective wears?**	**Nominal**	**Addressed to eligible child for the study (whose age is between 5 to 17)**
242	**What type of protective wears did they use?- Glasses**	**Nominal**	**Addressed to eligible child for the study (whose age is between 5 to 17)**
243	**Helmet**	**Nominal**	**Addressed to eligible child for the study (whose age is between 5 to 17)**
244	**Ear plugs**	**Nominal**	**Addressed to eligible child for the study (whose age is between 5 to 17)**
245	**Special shoes**	**Nominal**	**Addressed to eligible child for the 245study (whose age is b246etween 5 to 17)**
246	**Gloves**	**Nominal**	**Addressed to eligible child for the study (whose age is between 5 to 17)**
247	**Other**	**Nominal**	**Addressed to eligible child for the study (whose age is between 5 to 17)**
248	**Were you (the child) aware of any health problems in connection with you work?**	**Nominal**	**Addressed to eligible child for the study (whose age is between 5 to 17)**
249	**Is there any problem with the present job?**	**Nominal**	**Addressed to eligible child for the study (whose age is between 5 to 17)**
250	**Did you participate in any economic or**	**Nominal**	**Addressed to eligible child for the**

	non-economic activity or attend education or training		**study (whose age is between 5 to 17)**
251	**Main reason for complete idleness**	**Nominal**	**Addressed to eligible child for the study (whose age is between 5 to 17)**
252	**Given a choice, what would you (the child) like to do now?**	**Nominal**	**Addressed to eligible child for the study (whose age is between 5 to 17)**
253	**Given a choice what would you (the child) like to do in the future?**	**Nominal**	

Appendix 2

List of attributes used in the first experiment (147 attributes)

Sequence number of attributes	Attribute Name	Data Type	Description
1	Survey No.	Numeric	Serves as identification number for the survey
2	*Killil*	Nominal	Current address of the person responding the specific question
3	Zone	Nominal	Current address of the person responding the specific question
4	*Wereda*	Numeric	Current address of the person responding the specific question
5	Town	Nominal	Current address of the person responding the specific question
6	*Kefetegna*	Numeric	Current address of the person responding the specific question
7	*Kebele*	Numeric	Current address of the person responding the specific question
8	Enumeration Area code	Numeric	Code given to the area from which the data is collected
9	Selection Number of selected family	Numeric	Number assigned for each selected family for identification purpose
10	Selection Number of Respondent	Numeric	Number assigned for each respondent
11	Serial Number of Household Members	Numeric	Number assigned for each selected

			family member
12	**Relationship to head of household**	**Nominal**	**Relationship of the household member to the head of the household**
13	**Sex**	**Nominal**	**Sex of respondent**
14	**Age**	**Numeric**	**Age of respondent**
15	**Religion**	**Nominal**	**Religion of respondent**
16	**Ethnic Group**	**Nominal**	**Ethnic of respondent**
17	**Can you read/write**	**Nominal**	**Educational status of the respondent (addressed for household members whose age is greater than or equal to five)**
18	**The highest grade completed**	**Nominal**	**Educational status of the respondent (addressed for household members whose age is greater than or equal to five)**
19	**Have training?**	**Nominal**	**Training condition (addressed for household members whose age is greater than or equal to ten)**
20	**Type of training**	**Nominal**	**Training condition (addressed for household members whose age is greater than or equal to ten)**
21	**Marital Status**	**Nominal**	**(addressed for household members whose age is greater than or equal to ten)**
22	**Productive work for non-parents last week?**	**Nominal**	**Addressed for household members whose age is greater than or equal to five.**
23	**Productive work for family last week?**	**Nominal**	**Addressed for household members whose age is greater**

			than or equal to five
24	**Total number of hours worked last week**	**Numeric**	**Addressed for household members whose age is greater than or equal to five**
25	**Is the total working hour during last week greater or less than 4 hours?**		
26	**Did you have another job last week?**	**Nominal**	**Addressed for household members whose age is greater than or equal to five**
27	**Occupation you have been working during last week?**	**Nominal**	**Addressed for household members whose age is greater than or equal to five**
28	**Industry you have been working during last week?**	**Nominal**	**Addressed for household members whose age is greater than or equal to five**
29	**Employment status in major occupation last week**	**Nominal**	**Addressed for household members whose age is greater than or equal to five**
30	**Payment period last week**	**Nominal**	**Addressed for household members whose age is greater than or equal to five**
31	**Payment type (cash or kind)**	**Nominal**	**Addressed for household members whose age is greater than or equal to five**
32	**Recent cash payment amount**	**Numeric**	**Addressed for household members whose age is greater than or equal to five**
33	**Recent paid amount in kind**	**Numeric**	**Addressed for household members whose age is greater than or equal to five**
34	**Recent total payment amount**	**Numeric**	**Addressed for household members whose age is greater than or equal to five**
35	**Working hour shift**	**Nominal**	**Addressed for**

			household members whose age is greater than or equal to five
36	**Usual hours of work**	**Numeric**	**Addressed for household members whose age is greater than or equal to five**
37	**Additional productive activity to your major activity last week**	**Nominal**	**Addressed for household members whose age is greater than or equal to ten**
38	**Looking for additional productive activity last week**	**Nominal**	**Addressed for household members whose age is greater than or equal to ten**
39	**Attempt to change your job last week**	**Nominal**	**Addressed for household members whose age is greater than or equal to ten**
40	**Three months attempt to find job**	**Nominal**	**Addressed for household members whose age is greater than or equal to ten**
41	**Reason for unemployment**	**Nominal**	**Addressed for household members whose age is greater than or equal to ten**
42	**Preparedness for work in coming one month**	**Nominal**	**Addressed for household members whose age is greater than or equal to ten**
43	**Productive activities for the last 12 months**	**Nominal**	**Addressed for household members whose age is greater than or equal to ten**
44	**For how long did you engage in productive activities for last year (in months)?**	**numeric**	**Addressed for household members whose age is greater than or equal to ten**
45	**Major occupation you worked for last year?**	**Nominal**	**Addressed for household members whose age is greater than or equal to eighteen**
46	**Major industry you**	**Nominal**	**Addressed for**

	worked for last year?		**household members whose age is greater than or equal to eighteen**
47	**Reason for unemployment during last year?**	**Nominal**	**Addressed for household members whose age is greater than or equal to eighteen**
48	**Is the child eligible for the study?**	**Nominal**	**Eligible child is a child whose age is between 5 and 17**
49	**Has the household been living outside the current town/rural?**	**Nominal**	**Addressed to the household head**
50	**Previous *killil* in which the household been living?**	**Nominal**	**Addressed to the household head**
51	**Previous zone in which the household been living?**	**Nominal**	**Addressed to the household head**
52	**Was the previous household residence rural of town?**	**Nominal**	**Addressed to the household head**
53	**Reason for changing to the present place of resident**	**Nominal**	**Addressed to the household head**
54	**How long had this household been living in the present resident?**	**Nominal**	**Addressed to the household head**
55	**Ownership status of the household dwelling**	**Nominal**	**Addressed to the household head**
56	**Monthly amount of rent in birr**	**Numeric**	**Addressed to the household head**
57	**Major materials of house wall construction**	**Nominal**	**Addressed to the household head**
58	**Major materials of house roof construction**	**Nominal**	**Addressed to the household head**
59	**Number of rooms in the house**	**Nominal**	**Addressed to the household head**
60	**Toilet facility**	**Nominal**	**Addressed to the**

			household head
61	**Type of Kitchen**	**Nominal**	**Addressed to the household head**
62	**Source of drinking water**	**Nominal**	**Addressed to the household head**
63	**Type of fuel**	**Nominal**	**Addressed to the household head**
64	**Properties owned by household-own a house?**	**Nominal**	**Addressed to the household head**
65	**Have radio?**	**Nominal**	**Addressed to the household head**
66	**Have television?**	**Nominal**	**Addressed to the household head**
67	**Have telephone?**	**Nominal**	**Addressed to the household head**
68	**Have electric *mitad*?**	**Nominal**	**Addressed to the household head**
69	**Have refrigerator?**	**Nominal**	**Addressed to the household head**
70	**Have *bofe* and sofa?**	**Nominal**	**Addressed to the household head**
71	**Have car?**	**Nominal**	**Addressed to the household head**
72	**Have table and chair?**	**Nominal**	**Addressed to the household head**
73	**Have cropland?**	**Nominal**	**Addressed to the household head**
74	**Have cattle?**	**Nominal**	**Addressed to the household head**
75	**Have horse/mule/donkey?**	**Nominal**	**Addressed to the household head**
76	**Have camel?**	**Nominal**	**Addressed to the household head**
77	**Have sheep/goats?**	**Nominal**	**Addressed to the household head**
78	**Grow cash crops?**	**Nominal**	**Addressed to the household head**
79	**Estimated average monthly household consumption in cash**	**Nominal**	**Addressed to the household head**
80	**Estimated average monthly income of household(urban)**	**Nominal**	**Addressed to the household head**
81	**Estimated average yearly income of**	**Nominal**	**Addressed to the household head**

	household (rural)		
82	**Are there any children who live other place than their family?**	**Nominal**	**Addressed to the child's parents or guardians**
83	**How many male live outside family?**	**Numeric**	**Addressed to the child's parents or guardians**
84	**How many female live outside family?**		
85	**Age of elder child living outside family**	**Numeric**	**Addressed to the child's parents or guardians**
86	**Age of next elder child living outside family**	**numeric**	**Addressed to the child's parents or guardians**
87	**Age of younger child living outside family**	**Nominal**	**Addressed to the child's parents or guardians**
88	**Sex of elder child living outside family**	**Nominal**	**Addressed to the child's parents or guardians**
89	**Sex of next elder child living outside family**	**Nominal**	**Addressed to the child's parents or guardians**
90	**Sex of younger child living outside family**	**Nominal**	**Addressed to the child's parents or guardians**
91	**With whom is the elder child living?**	**Nominal**	**Addressed to the child's parents or guardians**
92	**With whom is the next elder child living?**	**Nominal**	**Addressed to the child's parents or guardians**
93	**With whom is the younger child living?**	**Nominal**	**Addressed to the child's parents or guardians**
94	**Living condition of the elder child living apart from family**	**Nominal**	**Addressed to the child's parents or guardians**
95	**Living condition of the next elder child living apart from family**	**Nominal**	**Addressed to the child's parents or guardians**
96	**Living condition of the younger child**	**Nominal**	**Addressed to the child's parents or**

	living apart from family		**guardians**
97	**Has the child been living with present household since birth?**	**Nominal**	**Addressed to the child's parents or guardians**
98	**Last place of usual child residence before coming here? *Killil***	**Nominal**	**Addressed to the child's parents or guardians**
99	**Previous zone**	**Nominal**	**Addressed to the child's parents or guardians**
100	**Was the last place of residence rural or urban?**	**Nominal**	**Addressed to the child's parents or guardians**
101	**Child's activity before coming to this household**	**Nominal**	**Addressed to the child's parents or guardians**
102	**Main reason for coming to present household?**	**Nominal**	**Addressed to the child's parents or guardians**
103	**How long has the child been living with the present household?**	**Nominal**	**Addressed to the child's parents or guardians**
104	**What type of education of training institution does the child attend in this academic year?**	**Nominal**	**Addressed to the child's parents or guardians**
105	**The grade level the child is currently attending**	**Nominal**	**Addressed to the child's parents or guardians**
106	**Did the child attend school or training institution during last week?**	**Nominal**	**Addressed to the child's parents or guardians**
107	**Did the child attend formal education before this academic year?**	**Nominal**	**Addressed to the child's parents or guardians**
108	**Main reason for not attending formal education?**	**Nominal**	**Addressed to the child's parents or guardians**

109	**Engagement in housekeeping activities without payment**	**Nominal**	**Addressed to the child's parents or guardians**
110	**Average number of working hours per day**	**Numeric**	**Addressed to the child's parents or guardians**
111	**Main housekeeping activities last week-Housekeeping**	**Nominal**	**Addressed to the child's parents or guardians**
112	**Cleaning of the household dwelling**	**Nominal**	**Addressed to the child's parents or guardians**
113	**Preparing meals**	**Nominal**	**Addressed to the child's parents or guardians**
114	**Serving meals**	**Nominal**	**Addressed to the child's parents or guardians**
115	**Mending, washing and pressing clothes**	**Nominal**	**Addressed to the child's parents or guardians**
116	**Shopping**	**Nominal**	**Addressed to the child's parents or guardians**
117	**Gathering firewood and dong cake**	**Nominal**	**Addressed to the child's parents or guardians**
118	**Caring for infants**	**Nominal**	**Addressed to the child's parents or guardians**
119	**Message**	**Nominal**	**Addressed to the child's parents or guardians**
120	**Fetching water**	**Nominal**	**Addressed to the child's parents or guardians**
121	**Other**	**Nominal**	**Addressed to the child's parents or guardians**
122	**Didn't work**	**Nominal**	**Addressed to the child's parents or guardians**
123	**Productive activity of child last week (guardians)**	**Nominal**	**Addressed to the child's parents or guardians**

124	**Did the child do non productive activities or attend school last week?**	**Nominal**	**Addressed to the child's parents or guardians**
125	**Main reason for not attending school and not doing any domestic work**	**Nominal**	**Addressed to the child's parents or guardians**
126	**Injury or illness at work place**	**Nominal**	**Addressed to the child's parents or guardians**
127	**How often was the child hurt**	**Nominal**	**Addressed to the child's parents or guardians**
128	**Industry in which serious injury occurred**	**Nominal**	**Addressed to the child's parents or guardians**
129	**Occupations in which serious damage occurred**	**Nominal**	**Addressed to the child's parents or guardians**
130	**Types of injuries/illness-Type1**	**Nominal**	**Addressed to the child's parents or guardians**
131	**Type2**	**Nominal**	**Addressed to the child's parents or guardians**
132	**Type3**	**Nominal**	**Addressed to the child's parents or guardians**
133	**Seriousness of the accident**	**Nominal**	**Addressed to the child's parents or guardians**
134	**Type of treatment received**	**Nominal**	**Addressed to the child's parents or guardians**
135	**Number of days hospitalized**	**Numeric**	**Addressed to the child's parents or guardians**
136	**Where did the child consult health personnel? Work place**	**Nominal**	**Addressed to the child's parents or guardians**
137	**Hospital/health station**	**Nominal**	**Addressed to the child's parents or guardians**
138	**Pharmacy**	**Nominal**	**Addressed to the**

			child's parents or guardians
139	**Clinic**	**Nominal**	**Addressed to the child's parents or guardians**
140	**Other place**	**Nominal**	**Addressed to the child's parents or guardians**
141	**Who paid most part of medical treatment cost? Parents**	**Nominal**	**Addressed to the child's parents or guardians**
142	**Employer**	**Nominal**	**Addressed to the child's parents or guardians**
143	**The child**	**Nominal**	**Addressed to the child's parents or guardians**
144	**Free medical service**	**Nominal**	**Addressed to the child's parents or guardians**
145	**No need for payment**	**Nominal**	**Addressed to the child's parents or guardians**
146	**Other**	**Nominal**	**Addressed to the child's parents or guardians**
147	**Did the child do productive work last week(guardians)**	**Nominal**	**Addressed to the child's parents or guardians**

Appendix 3

List of 86 attributes

Sequence number of attributes	Attribute Name	Data Type	Description
1	**Zone**	**Nominal**	**Current address of the person responding the specific question**
2	***Wereda***	**Numeric**	**Current address of the person responding the specific question**
3	**Town**	**Nominal**	**Current address of the person responding the specific question**
4	***Kefetegna***	**Numeric**	**Current address of the person responding the specific question**
5	***Kebele***	**Numeric**	**Current address of the person responding the specific question**
6	**Enumeration Area code**	**Numeric**	**Code given to the area from which the data is collected**
7	**Relationship to head of household**	**Nominal**	**Relationship of the household member to the head of the household**
8	**Sex**	**Nominal**	**Sex of respondent**
9	**Age**	**Numeric**	**Age of respondent**
10	**Religion**	**Nominal**	**Religion of**

			respondent
11	**Ethnic Group**	**Nominal**	**Ethnic of respondent**
12	**Can you read/write**	**Nominal**	**Educational status of the respondent (addressed for household members whose age is greater than or equal to five)**
13	**The highest grade completed**	**Nominal**	**Educational status of the respondent (addressed for household members whose age is greater than or equal to five)**
14	**Have training?**	**Nominal**	**Training condition (addressed for household members whose age is greater than or equal to ten)**
15	**Marital Status**	**Nominal**	**(addressed for household members whose age is greater than or equal to ten)**
16	**Productive work for non-parents last week?**	**Nominal**	**Addressed for household members whose age is greater than or equal to five.**
17	**Productive work for family last week?**	**Nominal**	**Addressed for household members whose age is greater than or equal to five**
18	**Total number of**	**Numeric**	**Addressed for**

	hours worked last week		**household members whose age is greater than or equal to five**
19	**Is the total working hour during last week greater or less than 4 hours?**	**Nominal**	**Addressed for household members whose age is greater than or equal to five**
20	**Did you have another job last week?**	**Nominal**	**Addressed for household members whose age is greater than or equal to five**
21	**Occupation you have been working during last week?**	**Nominal**	**Addressed for household members whose age is greater than or equal to five**
22	**Industry you have been working during last week?**	**Nominal**	**Addressed for household members whose age is greater than or equal to five**
23	**Employment status in major occupation last week**	**Nominal**	**Addressed for household members whose age is greater than or equal to five**
24	**Working hour shift**	**Nominal**	**Addressed for household members whose age is greater than or equal to five**
25	**Usual hours of work**	**Numeric**	**Addressed for household members whose age is greater than or equal to five**
26	**Additional productive activity to your major activity last week**	**Nominal**	**Addressed for household members whose age is greater than or equal to ten**

27	**Looking for additional productive activity last week**	**Nominal**	**Addressed for household members whose age is greater than or equal to ten**
28	**Attempt to change your job last week**	**Nominal**	**Addressed for household members whose age is greater than or equal to ten**
29	**Three months attempt to find job**	**Nominal**	**Addressed for household members whose age is greater than or equal to ten**
30	**Reason for unemployment**	**Nominal**	**Addressed for household members whose age is greater than or equal to ten**
31	**Preparedness for work in coming one month**	**Nominal**	**Addressed for household members whose age is greater than or equal to ten**
32	**Productive activities for the last 12 months**	**Nominal**	**Addressed for household members whose age is greater than or equal to ten**
33	**For how long did you engage in productive activities for last year (in months)?**	**numeric**	**Addressed for household members whose age is greater than or equal to ten**
34	**Is the child eligible for the study?**	**Nominal**	**Eligible child is a child whose age is between 5 and 17**
35	**Has the household been living outside the current**	**Nominal**	**Addressed to the household head**

	town/rural?		
36	**Ownership status of the household dwelling**	**Nominal**	**Addressed to the household head**
37	**Major materials of house wall construction**	**Nominal**	**Addressed to the household head**
38	**Major materials of house roof construction**	**Nominal**	**Addressed to the household head**
39	**Number of rooms in the house**	**Nominal**	**Addressed to the household head**
40	**Toilet facility**	**Nominal**	**Addressed to the household head**
41	**Type of Kitchen**	**Nominal**	**Addressed to the household head**
42	**Source of drinking water**	**Nominal**	**Addressed to the household head**
43	**Type of fuel**	**Nominal**	**Addressed to the household head**
44	**Properties owned by household-own a house?**	**Nominal**	**Addressed to the household head**
45	**Have radio?**	**Nominal**	**Addressed to the household head**
46	**Have television?**	**Nominal**	**Addressed to the household head**
47	**Have telephone?**	**Nominal**	**Addressed to the household head**
48	**Have electric mitad?**	**Nominal**	**Addressed to the household head**
49	**Have refrigerator?**	**Nominal**	**Addressed to the household head**
50	**Have bofe and sofa?**	**Nominal**	**Addressed to the household head**
51	**Have car?**	**Nominal**	**Addressed to the**

			household head
52	**Have table and chair?**	**Nominal**	**Addressed to the household head**
53	**Have cropland?**	**Nominal**	**Addressed to the household head**
54	**Have cattle?**	**Nominal**	**Addressed to the household head**
55	**Have horse/mule/donkey?**	**Nominal**	**Addressed to the household head**
56	**Have camel?**	**Nominal**	**Addressed to the household head**
57	**Have sheep/goats?**	**Nominal**	**Addressed to the household head**
58	**Grow cash crops?**	**Nominal**	**Addressed to the household head**
59	**Estimated average monthly household consumption in cash**	**Nominal**	**Addressed to the household head**
60	**Estimated average monthly income of household(urban)**	**Nominal**	**Addressed to the household head**
61	**Estimated average yearly income of household (rural)**	**Nominal**	**Addressed to the household head**
62	**Are there any children who live other place than their family?**	**Nominal**	**Addressed to the child's parents or guardians**
63	**Has the child been living with present household since birth?**	**Nominal**	**Addressed to the child's parents or guardians**
64	**What type of education of training institution**	**Nominal**	**Addressed to the child's parents or guardians**

	does the child attend in this academic year?		
65	**The grade level the child is currently attending**	**Nominal**	**Addressed to the child's parents or guardians**
66	**Did the child attend school or training institution during last week?**	**Nominal**	**Addressed to the child's parents or guardians**
67	**Did the child attend formal education before this academic year?**	**Nominal**	**Addressed to the child's parents or guardians**
68	**Main reason for not attending formal education?**	**Nominal**	**Addressed to the child's parents or guardians**
69	**Engagement in housekeeping activities without payment**	**Nominal**	**Addressed to the child's parents or guardians**
70	**Average number of working hours per day**	**Numeric**	**Addressed to the child's parents or guardians**
71	**Main housekeeping activities last week-Housekeeping**	**Nominal**	**Addressed to the child's parents or guardians**
72	**Cleaning of the household dwelling**	**Nominal**	**Addressed to the child's parents or guardians**
73	**Preparing meals**	**Nominal**	**Addressed to the child's parents or guardians**
74	**Serving meals**	**Nominal**	**Addressed to the child's parents or**

			guardians
75	**Mending, washing and pressing clothes**	**Nominal**	**Addressed to the child's parents or guardians**
76	**Shopping**	**Nominal**	**Addressed to the child's parents or guardians**
77	**Gathering firewood and dong cake**	**Nominal**	**Addressed to the child's parents or guardians**
78	**Caring for infants**	**Nominal**	**Addressed to the child's parents or guardians**
79	**Message**	**Nominal**	**Addressed to the child's parents or guardians**
80	**Fetching water**	**Nominal**	**Addressed to the child's parents or guardians**
81	**Other**	**Nominal**	**Addressed to the child's parents or guardians**
82	**Didn't work**	**Nominal**	**Addressed to the child's parents or guardians**
83	**Productive activity of child last week (guardians)**	**Nominal**	**Addressed to the child's parents or guardians**
84	**Did the child do non productive activities or attend school last week?**	**Nominal**	**Addressed to the child's parents or guardians**
85	**Injury or illness at work place**	**Nominal**	**Addressed to the child's parents or guardians**

86	**Did the child do productive work last week(guardians)**	**Nominal**	**Addressed to the child's parents or guardians**

Appendix 4

List of 63 attributes

Sequence number of attributes	Attribute Name	Data Type	Description
1	Zone	Nominal	Current address of the person responding the specific question
2	Wereda	Numeric	Current address of the person responding the specific question
3	Town	Nominal	Current address of the person responding the specific question
4	Kefetegna	Numeric	Current address of the person responding the specific question
5	Kebele	Numeric	Current address of the person responding the specific question
6	Enumeration Area code	Numeric	Code given to the area from which the data is collected
7	Serial Number of selected family	Numeric	Sequence number assigned for each selected family member
8	Relationship to head of household	Nominal	Relationship of the household member to the head of the

			household
9	**Sex**	**Nominal**	**Sex of respondent**
10	**Age**	**Numeric**	**Age of respondent**
11	**Religion**	**Nominal**	**Religion of respondent**
12	**Ethnic Group**	**Nominal**	**Ethnic of respondent**
13	**Can you read/write**	**Nominal**	**Educational status of the respondent (addressed for household members whose age is greater than or equal to five)**
14	**The highest grade completed**	**Nominal**	**Educational status of the respondent (addressed for household members whose age is greater than or equal to five)**
15	**Have training?**	**Nominal**	**Training condition (addressed for household members whose age is greater than or equal to ten)**
16	**Marital Status**	**Nominal**	**(addressed for household members whose age is greater than or equal to ten)**
17	**Productive work for non-parents last week?**	**Nominal**	**Addressed for household members whose age is greater than or equal to five.**
18	**Productive work for**	**Nominal**	**Addressed for**

	family last week?		**household members whose age is greater than or equal to five**
19	**Total number of hours worked last week**	**Numeric**	**Addressed for household members whose age is greater than or equal to five**
20	**Is the total working hour during last week greater or less than 4 hours?**	**Nominal**	**Addressed for household members whose age is greater than or equal to five**
21	**Did you have another job last week?**	**Nominal**	**Addressed for household members whose age is greater than or equal to five**
22	**Occupation you have been working during last week?**	**Nominal**	**Addressed for household members whose age is greater than or equal to five**
23	**Industry you have been working during last week?**	**Nominal**	**Addressed for household members whose age is greater than or equal to five**
24	**Employment status in major occupation last week**	**Nominal**	**Addressed for household members whose age is greater than or equal to five**
25	**Working hour shift**	**Nominal**	**Addressed for household members whose age is greater than or equal to five**
26	**Usual hours of work**	**Numeric**	**Addressed for household members whose age is greater than or equal to five**

27	**Additional productive activity to your major activity last week**	**Nominal**	**Addressed for household members whose age is greater than or equal to ten**
28	**Looking for additional productive activity last week**	**Nominal**	**Addressed for household members whose age is greater than or equal to ten**
29	**Attempt to change your job last week**	**Nominal**	**Addressed for household members whose age is greater than or equal to ten**
30	**Three months attempt to find job**	**Nominal**	**Addressed for household members whose age is greater than or equal to ten**
31	**Reason for unemployment**	**Nominal**	**Addressed for household members whose age is greater than or equal to ten**
32	**Preparedness for work in coming one month**	**Nominal**	**Addressed for household members whose age is greater than or equal to ten**
33	**Productive activities for the last 12 months**	**Nominal**	**Addressed for household members whose age is greater than or equal to ten**
34	**For how long did you engage in productive activities for last year (in months)?**	**numeric**	**Addressed for household members whose age is greater than or equal to ten**
35	**Has the household been living outside**	**Nominal**	**Addressed to the household head**

	the current town/rural?		
36	**Estimated average monthly household consumption in cash**	**Nominal**	**Addressed to the household head**
37	**Estimated average monthly income of household(urban)**	**Nominal**	**Addressed to the household head**
38	**Estimated average yearly income of household (rural)**	**Nominal**	**Addressed to the household head**
39	**Are there any children who live other place than their family?**	**Nominal**	**Addressed to the child's parents or guardians**
40	**Has the child been living with present household since birth?**	**Nominal**	**Addressed to the child's parents or guardians**
41	**What type of education of training institution does the child attend in this academic year?**	**Nominal**	**Addressed to the child's parents or guardians**
42	**The grade level the child is currently attending**	**Nominal**	**Addressed to the child's parents or guardians**
43	**Did the child attend school or training institution during last week?**	**Nominal**	**Addressed to the child's parents or guardians**
44	**Did the child attend formal education before this academic**	**Nominal**	**Addressed to the child's parents or guardians**

	year?		
45	**Main reason for not attending formal education?**	**Nominal**	**Addressed to the child's parents or guardians**
46	**Engagement in housekeeping activities without payment**	**Nominal**	**Addressed to the child's parents or guardians**
47	**Average number of working hours per day**	**Numeric**	**Addressed to the child's parents or guardians**
48	**Main housekeeping activities last week-Housekeeping**	**Nominal**	**Addressed to the child's parents or guardians**
49	**Cleaning of the household dwelling**	**Nominal**	**Addressed to the child's parents or guardians**
50	**Preparing meals**	**Nominal**	**Addressed to the child's parents or guardians**
51	**Serving meals**	**Nominal**	**Addressed to the child's parents or guardians**
52	**Mending, washing and pressing clothes**	**Nominal**	**Addressed to the child's parents or guardians**
53	**Shopping**	**Nominal**	**Addressed to the child's parents or guardians**
54	**Gathering firewood and dong cake**	**Nominal**	**Addressed to the child's parents or guardians**
55	**Caring for infants**	**Nominal**	**Addressed to the child's parents or guardians**

56	**Message**	**Nominal**	**Addressed to the child's parents or guardians**
57	**Fetching water**	**Nominal**	**Addressed to the child's parents or guardians**
58	**Other**	**Nominal**	**Addressed to the child's parents or guardians**
59	**Didn't work**	**Nominal**	**Addressed to the child's parents or guardians**
60	**Productive activity of child last week (guardians)**	**Nominal**	**Addressed to the child's parents or guardians**
61	**Did the child do non productive activities or attend school last week?**	**Nominal**	**Addressed to the child's parents or guardians**
62	**Injury or illness at work place**	**Nominal**	**Addressed to the child's parents or guardians**
63	**Did the child do productive work last week(guardians)**	**Nominal**	**Addressed to the child's parents or guardians**