

192

ADDIS ABABA UNIVERSITY
FACULTY OF INFORMATICS
DEPARTMENT OF INFORMATION SCIENCE

Application Of Data Mining Technology To Identify Determinant
Risk Factors Of HIV Infection And To Find Their Association Rules.

The Case of Center for Disease Control and Prevention (CDC)

By

Abraham Tesse Nedjo

June 2005

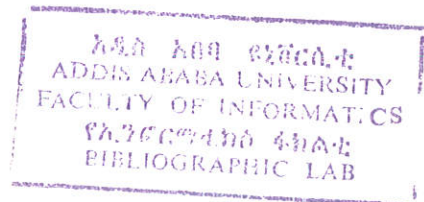
A Thesis Submitted To The Graduate Studies Of Addis Ababa University
In Partial Fulfilment Of The Requirements For The Degree Of Masters Of
Science In Information Science.

ADDIS ABABA UNIVERSITY
LIBRARIES
P.O. BOX 1176
ADDIS ABABA ETHIOPIA

ADDIS ABABA UNIVERSITY
SCHOOL OF GRADUATE STUDIES
Faculty of Informatics
Department of Information Science

APPLICATION OF DATA MINING TECHNOLOGY TO IDENTIFY DETERMINANT RISK
FACTORS OF HIV INFECTION AND TO FIND THEIR ASSOCIATION RULES: THE
CASE OF CENTER FOR DISEASE CONTROL AND PREVENTION (CDC)

BY
ABERHAM TESSO



Name and Signature of Members of the Examining Board

Dr. Gashaw Kebede, Chairman, Examining Board

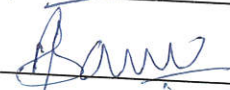
Prof. B.R. Krishna Rao, Advisor

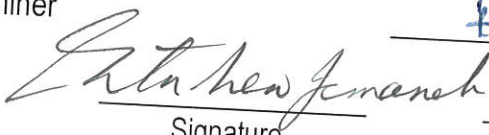
Dr. Kumudha Raimond, Examiner

Chairman, Faculty

Signature

Date


12/07/05

Dedication

Dedicated to my mother Jabane Obsi, who dedicated everything, she had to me and my brother Mathwos Tesso, who was unable to taste the joy of our family membership.

Acknowledgments

First and for most I would like to thank my beloved family specially my brothers Teklu, Tamiru, and Daniel, as well as my sisters Tsige, Hana, Dinknesh and Wegnesh for their prayer, support, and encouragement during my study. Their assistance, love and care were always with me.

I would like to thank my advisor Dr. Bandaru Rama Krishna Rao for his encouraging and constructive comments and for refining my work.

I am very grateful to the management and staff of CDC Ethiopia, especially W/o Seble, for their constant support in accessing the data and Ato Tamirat Teshome from Organization for Social Services for AIDS in providing relevant information to my research work.

My especial thanks goes to my sister Elfinesh Giragn for her support and encouragement, who helped me in typing my paper. She was always there whenever I needed someone. She was with me from the beginning to the end.

I am also grateful to my friends, staff members of Faculty of Informatics, and my classmates, for their encouragement, support, and the good friendship they shared. Special thanks to W/t Meseret Ayano, who offered good hospitality whenever her cooperation was needed.

Abbreviations

AIDS	Acquired immunodeficiency Syndrome
ANC	Antenatal Care
ART	Antiretroviral Therapy
ARV	Antiretroviral Drugs
BCC	Behavioural Change Communication
BSS	Behavioural Surveillance Survey
CDC	Center for Disease Control and Prevention
CL	Confidence Limit
CTR	Counseling, Testing and Referral
EHNRI	Ethiopian Health and Nutrition Research Institute
HIV	Human Immunodeficiency Virus
HSS	HIV Sentinel Surveillance
KDD	Knowledge Discovery in Databases
MOH	Ministry of Health
MTCT	Mother -To-Child Transmission
NSS	National Sentinel Surveillance
NSSR	National Sentinel Surveillance Report
NGO	Non-Governmental Organization
PLWHA	People Living With HIV/AIDS
PMTCT	Prevention of Mother -to-Child Transmission
SMCC	Sun Microsystems Computer Company.
STI	Sexually Transmitted Infection
UNAIDS	Joint United Nations Program on HIV/AIDS
VCT	Voluntary Counselling and Testing
WHO	World Health Organization

Table of Contents

DEDICATION.....	III
ACKNOWLEDGMENTS.....	IV
ABBREVIATIONS.....	V
LIST OF TABLES.....	VIII
LIST OF FIGURES.....	IX
LIST OF EQUATIONS.....	X
ABSTRACT.....	XI
CHAPTER ONE.....	I
1. INTRODUCTION.....	1
1.1. BACKGROUND.....	1
1.2. MOTIVATION OF THE STUDY.....	3
1.3. APPLICATION OF THE STUDY.....	6
1.4. OBJECTIVES OF THE STUDY.....	7
1.4.1. GENERAL OBJECTIVE.....	7
1.4.2. SPECIFIC OBJECTIVES.....	7
1.5. RESEARCH METHODOLOGY.....	8
1.5.1. IDENTIFYING THE DATA SOURCE.....	8
1.5.2. DETERMINING THE DATA QUANTITY.....	9
1.5.3. PRELIMINARY DATA ANALYSIS.....	9
1.5.4. DATA PREPARATION.....	10
1.5.5. TRAINING AND BUILDING MODELS.....	10
1.6. SCOPE AND LIMITATIONS.....	12
1.7. ORGANIZATION OF THE THESIS.....	12
CHAPTER TWO.....	12
2. HIV/AIDS AND CURRENT RESEARCH OUTPUTS.....	12
2.1 GENERAL OVERVIEW OF HIV/AIDS.....	12
2.2. GLOBAL HIV/AIDS CIRCUMSTANCES.....	12
2.3. HIV/AIDS IN ETHIOPIA.....	15
2.4 IMPACT OF HIV/AIDS IN ETHIOPIA.....	17
2.5. EFFORTS MADE TO ALLEVIATE HIV/AIDS IN ETHIOPIA.....	19
2.6. VOLUNTARY COUNSELLING AND TESTING.....	21
2.6.1. Elements of VCT HIV Counselling.....	21
2.6.2. Voluntary Testing.....	22
2.6.3. The Counselling Process.....	22
2.7. THE REASON OF CHOOSING VCT DATA FOR THIS RESEARCH.....	22
CHAPTER THREE.....	24
3. DATA MINING AND KNOWLEDGE DISCOVERY.....	24
3.1. INTRODUCTION.....	24

3.2. KNOWLEDGE DISCOVERY IN DATABASES	25
3.3. STEPS OF KNOWLEDGE DISCOVERY IN DATABASES	26
3.3.1. Problem Definition.....	27
3.3.2. Selecting a Target Data Set	27
3.3.4. Data Reduction and Projection	27
3.3.5. Choosing Mining Tasks and Algorithms.....	28
3.3.7. Interpreting and Evaluation of Mined Pattern.....	28
3.3.8. Consolidating Discovered Knowledge.....	29
3.4. MULTIDISCIPLINARY NATURE OF KDD	29
3.5. DATA MINING VERSUS OTHER STATISTICAL TOOLS	31
3.6. DATA MINING ACTIVITIES	32
3.7. APPLICATION OF DATA MINING	35
3.8. APPLICATION OF DATA MINING IN HEALTH SECTORS	36
CHAPTER FOUR.....	38
4. DATABASES AND ASSOCIATION MINING.....	38
4.1. MINING ASSOCIATION RULES IN LARGE DATABASES	38
4.2. INTERESTINGNESS MEASURES	39
4.3. ASSOCIATION MINING ALGORITHMS	42
4.3.1 Apriori Algorithm.....	42
4.3.2. Frequent Pattern Tree.....	43
CHAPTER FIVE.....	45
5. EXPERIMENTATION.....	45
5.1. THE GOAL OF THIS DATA MINING TASK	45
5.2. DATA MINING TOOL SELECTION	46
5.3. KNOWLEDGESTUDIO SOFTWARE	47
5.4. WEKA TOOL/SOFTWARE.....	48
5.5. IDENTIFYING THE SOURCE DATA	49
5.6. CREATING THE TARGET DATASETS	50
5.7. FEATURE SELECTION.....	52
5.8 DATA PREPROCESSING.....	55
5.8.1 Data transformation and reformation.....	55
5.8.2 Handling Variables with Missing Values.....	65
5.8.3 Data Organization for Model Building.....	66
5.9. MODEL BUILDING	67
5.10 GENERATING ASSOCIATION RULES.....	81
CHAPTER SIX	95
6. SUMMARY, CONCLUSION AND RECOMMENDATIONS	95
6.1. SUMMARY AND CONCLUSION	95
6.2. RECOMMENDATIONS.....	98
REFERENCES:	100

List of Tables

Table 2.1 Global Summary of HIV/AIDS Epidemic -----	13
Table 2.2 Estimated Adult and Child Deaths Due to HIV/AIDS During 2004 -----	14
Table 2.3 Estimated Number of Adults and Children Newly Infected with HIV During 2004 -----	14
Table 2.4. Summarized List of HIV/AIDS incidents in Ethiopia -----	15
Table 2.5 Estimated and projected adult HIV Prevalence (%) by sex and setting .selected years--	18
Table 5.1. Attributes Selected from the original data file -----	53
Table 5.2 Actual Values of "SEX" code and the original numeric code -----	56
Table 5.3. Actual Values of each "MARTSTAT" and the original numeric code -----	57
Table 5.4 The Original "EDUEXPE" code the decoded symbolic value -----	57
Table 5.5. The original code and Transformed values of "EMPLOYED" attribute-----	58
Table 5.6. The Original code and transformed values of "OCCUPAT" attribute-----	58
Table 5. 7. The original data set code and transformed value of "REFERBY" attribute -----	59
Table 5.8 The original data set code and transformed value of "REASHERE" attribute-----	60
Table 5.9 The original data set code and transformed value of "PRTESTED" attribute-----	61
Table 5. 10 The Original "EVERHAD" code and the new decoded values-----	61
Table 5.11 The Original "EXPOTIME" code and the new decoded values-----	61
Table 5.12 The Original "USECOND" code and the new decoded values-----	62
Table 5.13 The Original "LASTSEX" code and the new decoded values-----	62
Table 5.14. The Original "STIHIST" code and the new decoded values-----	63
Table 5.15 The Original "SEXWORK" code and the new decoded values-----	63
Table 5.16 The Original "EXPECT" code and the new decoded values-----	63
Table 5.17 The Original "NOTIFPRE" code and the new decoded values-----	64
Table 5.18 The Original "HIV" code and the new decoded values-----	64
Table 5.19 The Original "NOTIFPOS" code and the new decoded values-----	64
Table 5.20 Modal Values to Represent Missing Values for Nominal variables -----	66
Table 5. 21 List of variables excluded and the performance of those models -----	71
Table 5.22. Sample data prepared as single word value for all attributes-----	82
Table 5. 23 principal options for the APRIORI association rule learner.-----	85

List of Figures

Figure 2. 1 Estimated number of PLWH	13
Figure 5.1 Main Screen of Epi Info Software.....	51
Figure 5.2 Data insertion process to the KnowledgeSTUDIO.....	68
Figure 5.3 Decision Tree for attribute AGE	73
Figure 5. 4 Decision Tree for attribute USECOND	74
Figure .5. 5. Decision Tree for attributes USECOND and EXPECT	75
Figure 5. 6 Decision Tree for attribute MARSTAT	76
Figure 5. 7 Decision Tree for attribute STIHIST	79
Figure 5. 8. Decision Tree for attribute REASHERE	79
Figure 5. 9. Decision Tree for attribute SEX and EMPLOYED.....	80
Figure 5. 10 The WEKA Explorer screen	83
Figure 5. 11. Output from the APRIORI Association Rule Learner for the First Experiment.....	84
Figure 5.12. Output from the APRIORI Association Rule Learner for only positive records	87
Figure 5.13. Output from the APRIORI Association Rule Learner for only positive records without HIV field.	88
Figure 5.14 Output from the APRIORI Association Rule Learner for few selected fields.....	89
Figure 5. 15 Output from the APRIORI Association Rule Learner for another fields	90
Figure 5. 16 Output from the APRIORI Association Rule Learner for HIV negative instances.....	91
Figure 5 17 Output from the APRIORI Association Rule Learner for few fields of HIV negative instances	92

List of Equations

Equation 4. 1. Formula to Measure Confidence of Rule	40
Equation 4. 2. Formula to Measure Support of Rule.....	41

Abstract

Nowadays, people all over the world have been challenged with the confronting problem of HIV/AIDS. There is large scale of HIV-related data available at VCT centers and other clinics and hospitals. What is demanding is knowledge from the data that may help HIV/AIDS prevention activities.

This paper reports on the findings of a research that had the objective to apply data mining technology to find determinant risk factors of HIV infection and their association rules, that would broaden the insight about HIV/AIDS. The study focused on using VCT for the reason that the data availability in electronic format. A data set of 5267 visitors is used.

In this paper the process of knowledge discovery and data mining functions are described. Different activities are performed for the purpose of data preprocessing using Ms Excel. Among the total records 70% is used to train the model and 30% is used to validate classifier with unseen data. The KnowledgeSTUDIO (to identify the factors) and WEKA (to mine their association rules) softwares are used for experimenting the research.

In general, the research has resulted in determining the frontier risk factors, analyzing the influence of each with decision tree and assessing their affinity. The reported findings are promising, making the result usable to the health professional, government, policy makers, and the society at large. And the whole research process can be a good input for further research.

CHAPTER ONE

1. INTRODUCTION

1.1. BACKGROUND

The global wide spectrum of HIV/AIDS (Human Immunodeficiency Virus/Acquired Immunodeficiency Syndrome) problem has made all human beings to worry. Millions are suffering from HIV all over the world today. The number of people who pass away from HIV/AIDS Epidemic is very astonishing to hear about. The number of people living with HIV/AIDS, according to UNAIDS (2004), HIV/AIDS Epidemic report is closing to 40 million. The same source showed about the 5 million of newly infected people in 2004. HIV/AIDS has become a prevalent disease and presents a global problem and its impact, if not controlled, will be profound and is considered to be highly destructive. Especially it's the main causes of death among youth and middle age adults (highly productive and reproductive of the society) (WHO 2002).

Looking at the Ethiopian context, Ethiopia is one of the countries in sub-Saharan Africa most severely affected by HIV. In Ethiopia, AIDS is now recognized as the leading cause of adult morbidity and mortality in the country. In Ethiopia, AIDS has already increased the number of deaths in Ethiopia by 6 percent. The projection of the data, by Lisa Garbus 2003, showed that between 2000 and 2015, it will increase the number of deaths in Ethiopia by 27 percent. More than 42 percent of all hospital beds in the country are estimated to be occupied by AIDS patients. The risk factors associated with this problem may be many in number but some variables that make people to be infected with high probability are considered to be main risk factors. Most research papers reported so far by health professionals show the infection patterns mostly in terms of only the age and sex group of the individual. However, the association patterns of these attributes beyond these two variables have not been investigated.

According to Shegaw (2002), identifying health-related problems of the community is one of the most important steps in the planning of health care intervention. So, the availability of valid data on HIV/AIDS treatment, and on patterns of volunteers

counselling and testing units and their utilization, very much facilitates planning, monitoring, and evaluation of HIV treatments and counselling services.

Although various researches have been carried out elsewhere to investigate the applicability of data mining technology for various mining problems, to the knowledge of the researcher, no data mining technology in general has been investigated in finding major factors (risk factors) and their association that can increase the understanding for the HIV infection patterns for people in Ethiopia specially using the voluntarily counselled and tested individuals.

Center for Disease Control and Prevention (CDC) has a global AIDS program to help prevent HIV infections, improve care and support and build capacity to address the global HIV/AIDS pandemic. CDC provides financial and technical assistances through partnerships with community, governments, and national and international entities working in resource-constraint countries. Additional HIV prevention, education and research programs are also conducted in other CDC centers, institutes, and offices. CDC employs a comprehensive approach to preventing further spread of HIV and AIDS. These activities are critically important, as it is estimated that millions are currently living with HIV. And the number of people seeking the treatment, counselling, and test services is highly increased. The Health statistics center of CDC collects HIV/AIDS -related data in many of its data systems, including HIV voluntary counselling and testing data.

A comprehensive approach to prevent HIV includes the opportunity for the mass to discover whether they are infected together with appropriate counselling and support. Whatever the results of testing, people need to learn how to reduce their risk of getting or transmitting the infection. Counselling and Testing services are vital for prevention and treatment strategies, which is much better when done on voluntary bases. At the same time investigating the degree of association of the risk factors elevating HIV infection is equally important. That is what CDC and other health institutions looking for.

The capabilities to collect and store patients/clients data in large computer databases has increased significantly. But the relational database technology of today offers little functionality to process and explore data and establish a relationship or pattern among data element that are hidden or previously unknown

(Raghavan, et. al., 1998). Prather et. al. (2002) on his part states that, although health care databases have accumulated large quantities of information about patients and their medical conditions, there are only few tools to evaluate and analyze this clinical data after it has been captured and stored. The authors further stated that evaluation of stored data might lead to the discovery of trends and patterns hidden within the data that could significantly enhance our understanding of disease progression and management.

As Fayyad, Patesky- Phapcro, and Smyth (1996) describe, the traditional method of turning data into knowledge relies on manual analysis and interpretation. The writers further stated that, for example, in the health industry, it is common for specialists to periodically analyze current trends and changes in health data, say, on quarterly basis. The specialists then provide a report detailing the analysis to the sponsoring health organization; this report will be used for further decision making and planning for health management.

Specifically, as data volumes grow dramatically, data analysis based on manual methods is becoming completely impractical in many domains. Parther et.al., (2002) argued that to evaluate and analyze data stored in large databases, new techniques and methods are needed to search large quantities of data and to discover new patterns and relationships hidden in the data. It is due to the challenges of searching for knowledge in relational databases and our inability to interpret and digest these data as readily as they are accumulated, which has created a need for generation of tools and techniques for automated and intelligent database analysis. It is from this consequent that the discipline of *Knowledge Discovery in Databases (KDD)*, which deals with the study of such tools and techniques, has evolved into an important and active area of research (Raghavan, et. al. 1998).

1.2. MOTIVATION OF THE STUDY

In clear sense, it is an established fact that HIV case is an important issue as it could, directly or indirectly, affect all human being irrespective of job, life style, and their activities. HIV infection pattern understanding can be used by almost all sectors of the society at large. Mainly the government, medical sectors and steamed

researchers to protect life and to improve the efficiency of operations, and to plan a wide range of activities.

As it is mentioned in Anteneh (2004), a few decades have been elapsed since HIV/AIDS has drawn the attentions of all people in the world. HIV infection has been infamous for a number of reasons including its rapid spread across the globe endangering the lives of millions (especially in developing countries) and its firm resistance to the vigorous search for remedy and vaccine.

The underlying research problem that necessitated this study is the existence of high HIV/AIDS prevalence at a national level in general and in Addis Ababa city in particular. Specifically at the Addis Ababa city administration data on HIV-related cases is gathered daily by many health service centers and CDC have agreement to obtain the data. But the availability of this data analysed with the traditional statistical data analysis alone is not enough for health professionals, planners and policy makers to identify major determinants of HIV infection and transmission.

On the other hand, the problem is that all those previously done studies on health were conducted by using a very small proportion of the data bases. Yet, in those studies, data analysis was conducted by using simple statistical techniques, such as regression and verification techniques (Shegawu, 2002). Since the analysis made by using traditional methods focuses on problems with much more manageable number of variables and cases than may be encountered in real world data bases, they have limited capacity to discover new and unanticipated patterns and relationships that are hidden in conventional relational databases (Plate.et, al, 1997)

Particularly, in order to identify patterns in HIV-infection and associated risk factors, it is difficult to conclusively attribute the problem to a certain set of factors since the parameters (attributes or factors) involved are too many. So studying the affinity of the factors rather than simply listing them is of crucial importance. On this point Last and Kandel (2002) quoted in Shegaw(2002) mentioned that the tools used in the research on death causality have been so far limited to the statistical techniques, like summarization, regression, analysis of variance, etc. However, such methods of data analysis may concentrate on identifying the leading causes of infection and testing

the association of single factor with certain disease. No complex models involving more than one factor have been built.

Though the Center for Disease Control and Prevention in Addis Ababa office has collected and stored large volume of HIV- related data, including the data of counselled and tested volunteers, the accumulated data is not as such fully utilized to support HIV/AIDS prevention and control activities in the country. So, in such situations, new information technologies like data mining may be appropriate means in discovering hidden patterns in data.

Thus, in this study, the researcher has investigated how the attributes regarding the social, economical, behavioural, environmental, educational, and health related factors determine HIV infection and transmission in the selected area by applying the new computational methods of data mining technology. And this study also has examined the association of these frontier factors detailing with what degree of confidence and support they co-exist on an individual that has high probability of HIV infection.

The ultimate goal of this research is to assess how well a model can be built by using data mining techniques that could perform well in identifying the risk of HIV infections and transmission among the mass based only upon the personal, partner, environmental factors, etc, without using diagnostic tests or physical exam findings. And also the research had aimed at finding the degree to which the influential factors are associated. This type of model might have application out side of clinical settings to support health workers in taking preventive actions to reduce HIV related problems as well as to assist health service planners, policy makers, and decision makers as a decision support aid in planning and implementing health intervention programs aimed at in the fights against HIV/AIDS. Thus, the results of this data mining process can be used to distinguish individuals at high risk of infection from those with low risk of infection by looking at the probability and determining power of each attribute.

The research question of this study is that if the existing data in the VCT (Voluntary Counselling and Test) database is analyzed by using data mining techniques and tools, it can provide knowledge (beyond simply being a data) that can facilitate the

improvement of task against HIV/AIDS by identifying significant and meaningful patterns among different data elements which are hidden or previously unknown. Specifically, in relation to the problem domain of this study, if the right data included in the VCT system is selected, organized, cleaned, and analyzed using data mining tools and techniques it might show a result that could help to develop a model that predicts the risk of infection among individuals and to identify major determinants (factors that lead to probability) of HIV infection and transmission in Ethiopia together with their degree of co-occurrence.

Voluntary Counselling and Test (VCT) is the most important services provided by the medical profession and, as already mentioned, HIV/AIDS test and counselling is an area where techniques of data mining technology can be successfully applied to come up with various results in the area based on the pattern of the available data. For instance, data mining techniques can be used to investigate the attributes of a visitor that makes a risky volunteer (with a positive HIV test) and safe visitor (with a negative HIV test) clearly showing the probability of being so and measuring the degree of co appearance for the most significant attributes.

Knowing one's HIV status is considered now days to be the best way towards the fight against HIV/AIDS. And it is believed that the good way to check one's status of HIV is through voluntary counselling and test service. This is the main reason why, presently, number of health centers and institutions are providing the service by encouraging visitors (or volunteers) by lowering the counselling and test fee down to a free of charge.

Here, as to the objective of this research, the main concern of interest should not only be the service provision but also to understand and investigate the volunteers' data and treat them accordingly. So, this study is an attempt that can encourage the experts in the area to exploit the wide range of applications of data mining technology and techniques in various facets of the discipline.

1.3. APPLICATION OF THE STUDY

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

In the first place, the Center for Disease Control and Prevention as well as other similar organization may use the intended model to know deeper about their data and contribution of each attributes of the customer that determine the probability of infection. Beyond simply knowing the factors, it enables also to measure degree of association of the attributes resulting in infection. This helps to device different strategies for the efforts towards the struggle against HIV/AIDS, like for instance, planning education programs to enhance awareness, in case lack of awareness, result in high probability of infection. In doing so they can improve the services they render to save and protect life of the majority.

Secondly researchers in the domain area can use the result of this study to collaborate with the basic researches they conduct on HIV/AIDS and related studies and operations. Like for instance, if the study reveals that gender differences show differences in infection pattern, medical researchers may get some ground to step for further investigation.

The government, minister of health, and policy makers can also use the out put of this research as a decision support means when setting policies on the issues of HIV/AIDS prevention and elimination.

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

1. 4. OBJECTIVES OF THE STUDY

1. 4.1. GENERAL OBJECTIVE

The objective of this study is to explore the potential applicability of Data Mining Techniques to identify the major factors that result in HIV Infection and Transmission and to investigate their association; that can support counselling and test service providers, policy makers, planners, etc, to prevent the prevalence of HIV/AIDS.

1. 4.2. SPECIFIC OBJECTIVES

In order to achieve the specified general objective, the researcher has the following targets as its specific objectives:

- Literature Review
- Identify determinant risk factors

- Assess the influence of the risk factors
- Find association rules for the factors
- Identify group of the society highly affected
- Identify group of the society most benefited from VCT
- Report the result and make recommendations

1. 5. RESEARCH METHODOLOGY

The methodology applied for this research was adopted from the techniques used by Stat soft and Two Crows (1999). It uses the different data mining steps that need to be carried out to develop a decision support system, assuming that the problem is well understood.

The different methods employed for the research and how they were applied for the study is provided below.

1. 5. 1. IDENTIFYING THE DATA SOURCE

In this research, the data that were relevant to the problem under question was identified based on the available literatures and by obtaining assistance from the experts in the area. The VCT data was found to be an appropriate source and was collected from CDC center of USA in Addis Ababa. Center for Disease Control and Prevention (CDC) collects and records HIV/AIDS-related treatment data including the VCT data from service centers like Zewuditu Memorial Hospital in Addis Ababa and Organization for Social Services for AIDS. The researcher used this VCT data available at this center. This is because; the VCT data that are required for this research are already available in electronic format at this station. And secondly, VCT is an HIV treatment (service) that encompasses the general population and is best type of data to represent the mass with no restriction.

1. 5. 2. DETERMINING THE DATA QUANTITY

It has been important to make a reasonable estimate of how much data would be needed for undertaking the study. In this regard, based on available literatures and experts' opinion, all the existing data were collected and selection was made before applying data mining process.

1. 5. 3. PRELIMINARY DATA ANALYSIS

Data mining can be regarded as extensions of standard statistical techniques, and such techniques can give us an idea of the performance, that the model is likely to achieve. In this respect, assuming that the factors that influence HIV infection, such as marital status, sex, age, occupation, reason for the test may be good factors and hence the researcher has made use of visual observation as well as statistical correlation analysis to identify determinant attributes in addition to the experts' suggestions.

1. 5. 4. DATA PREPARATION

Since it is necessary to convert the row data to a more suitable format, the researcher has done the data preparation task as follows: In the first step, the data was checked for validity in size and quality. Appropriate measures were taken to correct inconsistent and missing values. Since data mining generally provides improved performance with the normalized data, all the selected client data sets were transformed into values that are convenient for the selected data mining tools. Especially, on preparing the data set for the association mining using WEKA software (with Apriori algorithm), tough effort was made to change the nominal values of all the attributes to single label that Java environment of WEKA understands.

1.5. 5. TRAINING AND BUILDING MODELS

The next step after collecting, cleaning, and transforming the data was training and model building. Model building, which is the core work in the data mining problem, was accomplished using available data mining software. The selection criteria for the software were made on the bases of availability and convenience of the software. Among a number of available data mining software, KnowledgeSTUDIO and WEKA were used in this project at different level and purpose of the research.

The experiment was conducted by dividing the data into training and validation sets. The training set was used to actually train the model and the validation set was used to check the generalization ability of the model with a data set that is not part of the training set. Of the total data, 70% was used to train the model, and 30% was used for validation.

1. 6. SCOPE AND LIMITATIONS

The scope of this research was limited to the voluntary counselling and test clients of Center for Disease Control and Prevention, where the required customer data was available. Even though the findings of this study can fairly be adopted as relevant to explore and implement the potential applicability of data mining tools in other centres of similar service, related to HIV/AIDS, this research is limited to the case of Center for Disease Control and Prevention.

Furthermore, the study was limited to the voluntary counselling and test data in investigating the contribution of each attribute resulting in HIV infection and finding their association rule. This pattern can be used both to understand the existing data and to predict how new instances will behave.

The time that was given to undertake this research work was very short and counted a considerable limitation in developing the underlying model.

1. 7. ORGANIZATION OF THE THESIS

The whole paper or thesis is organized (chunked) into six chapters. In the first chapter, introduction, mainly the problem for which the research was conducted is presented. Moreover, the general and specific objectives, and the methodology employed to carry out the research is discussed.

The second chapter contains reviews of the works done in the problem area of the research. The current state of HIV/AIDS and related research problems were identified. Moreover, possible solutions to these problems are indicated.

In the third chapter, literatures on the technology, which is at the center of this study, namely, Data Mining, Knowledge Discovery in Databases, steps of KDD, will be reviewed. Knowledge Discovery in Databases activities and mining association rules were also appraised.

The fourth chapter is mainly concerned with the particular data mining activity adopted in solving the problem under consideration. This is about association mining in a bit detail level of abstraction.

In the fifth chapter, which is the experimentation part of this research, the different steps to be followed in collecting and preparing data were described. Furthermore, the model building process using appropriate techniques and algorithms, including the various parameters were discussed in detail. Finally, summary and interpretation of the experiment's results are given. And in the sixth chapter, the concluding remarks and recommendations are discussed.

CHAPTER TWO

2. HIV/AIDS AND CURRENT RESEARCH OUTPUTS

2.1 General Overview of HIV/AIDS

HIV/AIDS is the most far-reaching and damaging epidemic the world had ever seen (WHO 2002). WHO, further stated that HIV/AIDS, within a single generation, has grown into an individual and social tragedy with huge implications for human security, for social and political stability and for economic development. Originally viewed as just another disease, HIV/AIDS has long since moved beyond the boundaries of the health system. It is now generally acknowledged that addressing the pandemic required concerted efforts across all sectors involving a wide array of actors. Nevertheless, the health sector must remain central to all efforts to halt transmission of HIV and to mitigate its impact.

2.2. Global HIV/AIDS Circumstances

As all may perceive it, the effect of HIV/AIDS spans all over the world, and it is very measurable when viewed by the number of people infected and the number of AIDS and HIV/AIDS-related deaths. For this review on the global effect of HIV/AIDS, UNAIDS, December 2004 reported global summary of the HIV/AIDS epidemic indicating number of people living with HIV/AIDS, number of newly infected people in 2004, and AIDS death in 2004 reviewed as shown in the Table 2.1 on the next page.

This table indicates that the total number of people living with the Human Immunodeficiency Virus (HIV) rose in 2004 to reach its highest level ever, compared with years before: an estimated 39.4 million [35.9 million–44.3 million] people are living with the virus. This figure includes the 4.9 million [4.3 million–6.4 million] people who acquired HIV in 2004. The global AIDS epidemic killed 3.1 million [2.8 million–3.5 million] people in the past year. The following graph, as adopted from the same report, displays graphically the estimated number of people living with HIV.

Number of people living with HIV/AIDS		
	Total	39.4 million (35.9–44.3 million)
	Adults	37.2 million (33.8–41.7 million)
	Women	17.6 million (16.3–19.5 million)
	Children under 15 years	2.2 million (2.0–2.6 million)
People newly infected with HIV in 2004		
	Total	4.9 million (4.3–6.4 million)
	Adults	4.3 million (3.7–5.7 million)
	Children under 15 years	640 000 (570 000–750 000)
AIDS deaths in 2004		
	Total	3.1 million (2.8–3.5 million)
	Adults	2.6 million (2.3–2.9 million)
	Children under 15 years	510, 000 (460 000–600 000)

Table 2.1 Global Summary of HIV/AIDS Epidemic

[Source: *AIDS Epidemic Update, UNAIDS, December 2004*]

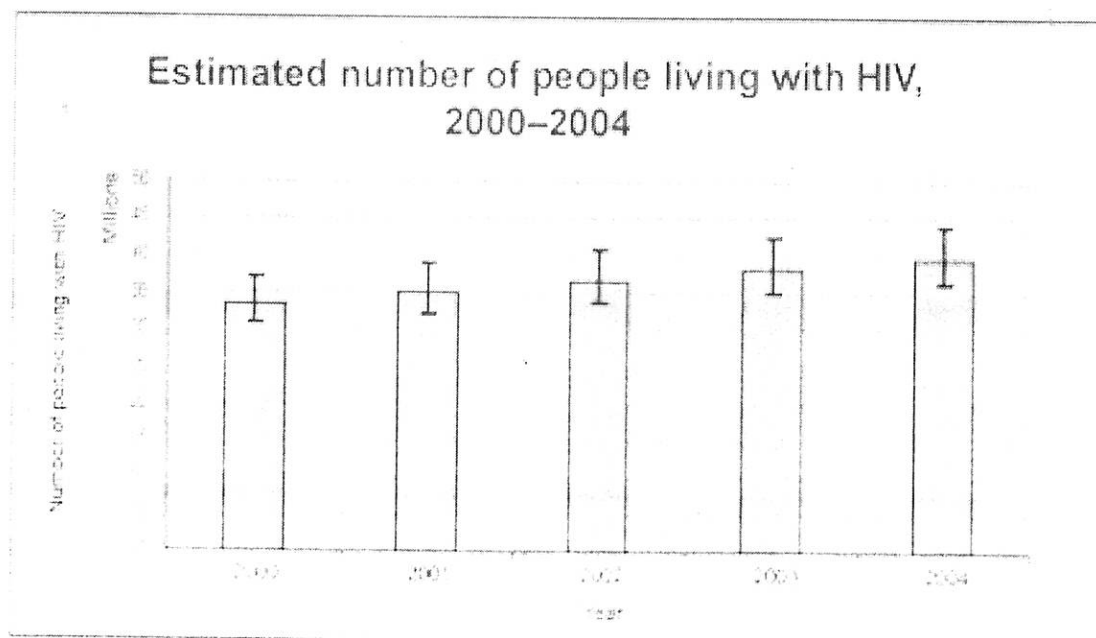


Figure 2. 1 Estimated number of PLWH

Taken from *AIDS Epidemic Update, UNAIDS, December 2004*]

The report yet indicate that Sub-Saharan Africa remains by far the worst-affected region, with 25.4 million [23.4 million–28.4 million] people living with HIV at the end of 2004, compared to 24.4 million [22.5 million–27.3 million] in 2002. Just under two thirds (64%) of all people living with HIV are in sub-Saharan Africa. The following table, quoted from the same source showed estimated adult and child deaths due to HIV/AIDS during 2004 for various continent of the world.

Sub-Saharan Africa	2,300,000
Asia	540,000
Latin America	95,000
Eastern Europe and Central Asia	60,000
North America, Western and Central Europe	23,000
North Africa and Middle East	28,000
Caribbean	36,000
Oceania	700

Table 2.2 Estimated Adult and Child Deaths Due to HIV/AIDS During 2004

[Source: *AIDS Epidemic Update, UNAIDS, December 2004*]

What is still important to give attention is that the number of people estimated to be newly infected keeps rising. From the same source (*UNAIDS, December 2004*) it is indicated that millions are newly infected during the year 2004, as shown in the following table.

Sub-Saharan Africa	25,400,000
Asia	3,100,000
Latin America	240,000
Eastern Europe and Central Asia	210,000
North America, Western and Central Europe	64,000
North Africa and Middle East	92,000
Caribbean	53,000
Oceania	5,000

Table 2.3 Estimated Number of Adults and Children Newly Infected with HIV During 2004

[Source: *AIDS Epidemic Update, UNAIDS, December 2004*]

Having all the review of HIV/AIDS in the world, it is not a right time to simply worry about HIV but every body should do something about it. It is the responsibility of all human being to act against HIV/AIDS. For same reason, the researcher of this experimental study is striving from his part towards the fight against HIV/AIDS by implementing the possible and available information technology tools to the problem domain.

2.3. HIV/AIDS in Ethiopia

Before starting writing about the current HIV/AIDS, it may be of importance to review its history of emergence. The first HIV infections in Ethiopia were identified in 1984. In 1986, the first AIDS cases were reported. In September 1987, the government established an HIV/AIDS department within the Ministry of Health, and in 1988, an HIV surveillance system was established. In 1989, the Health Bureau of the Addis Ababa City Administration began HIV sentinel surveillance. (Lisa Garbus 2003). The list of these incidents in Ethiopia is summarized as in the Table 2.4. Lisa Garbus added that, in the year 2003, there were 37 HIV sentinel surveillance sites (31 urban, 6 rural), though only 34 report data to the MOH. But what needs to be noticed here is that as the overwhelming majority of HSS sites are in urban areas, an enormous segment of the rural population remains uncovered by the current HIV sentinel surveillance system. Moreover, the six rural sites represent only two regions. This despite that 85 percent of the population lives in rural areas.

Year	Incidents
1984	The first HIV infection identified
1985	National HIV/AIDS task force established
1986	The first AIDS cases reported
1987	HIV/AIDS Department established
1988	HIV Surveillance system established
1989	Addis Ababa Health Bureau began Sentinel Surveillance
1991	National HIV/AIDS treatment policy was drafted
2000	The National AIDS Prevention & Control was established
2001	VCT service established as HIV Intervention strategy

Table 2.4. Summarized List of HIV/AIDS incidents in Ethiopia

According to Lisa Garbus 2003, HIV/AIDS in Ethiopia increased rapidly during the 1990s. The report of this writer showed that by 1989, HIV prevalence among the general adult population was estimated at 2.7 percent, increasing to 7.1 percent in 1997 and to 7.3 percent in 2000. In 2001, this figure was 6.6 percent. However, the MOH (Minister of Health) does not believe that this fall indicates that the HIV epidemic in Ethiopia is declining; rather, it is primarily a result of the reclassification of one sentinel site. The MOH estimated that 2.2 million Ethiopians were living with HIV/AIDS in 2001, of whom 2 million were adults.

During the early stages of the HIV/AIDS epidemic, there was a major effort to conduct serious surveys in Addis Ababa and other major urban centers among core transmitter groups. However, post-1990, there are very few data to indicate the level or progression of the epidemic among sex workers and truck drivers, as well as traders/merchants and the military.

When we look at the nature of the HIV case data as seen from the paper of Lisa Garbus 2003, it indicate that heterosexual and MTCT (Mother-To-Child Transmission) transmission account for almost all HIV infections in the country. The few data available have not found an association between harmful traditional practices and acquisition of HIV. HIV transmission via unsafe injections appears to be very low.

AIDS case reporting began soon after the establishment of the HIV/AIDS department within the MOH in 1987. AIDS cases are grossly underreported. Among women, AIDS cases peak between ages 20 and 29; for men, between ages 25 to 34. (MOH 2004). Lisa Garbus 2003, also mentioned that from the ANC (Antenatal Care) data, the group with the highest HIV prevalence in the country is women ages 15 to 24. Data from blood donors, visa applicants, and police and army recruits indicate that HIV prevalence among men peaks between ages 25 and 29. This is likely related to age mixing, wherein young women have older male sex partners, primarily for economic reasons.

According to ANC data, (on MOH 2004), the group with the highest HIV prevalence in the country is women ages 15 to 24. Data from blood donors, visa applicants, and police and army recruits indicate that HIV prevalence among men peaks between

ages 25 and 29. This is likely related to age mixing, wherein young women have older male sex partners, primarily for economic reasons.

When we look at the political, economy and socio-behavioral context of the country, the relationship between HIV prevalence and socioeconomic factors is highly complex. Increasingly, risk of HIV infection is recognized as related to, inter alia, one's socioeconomic status as well as the socioeconomic profile of the community in which one is situated. Additionally, Ethiopia's political history, civil war, conflicts with Eritrea, and current food crisis also affect HIV/AIDS dynamics (Lisa Garbus 2003)

There is also a general belief that Ethiopia's mobile populations including the rapidly increasing number of rural residents seeking employment in urban areas; military personnel; those displaced by war, drought, and/or environmental degradation; male transport workers; sex workers; émigrés; traders; orphans and vulnerable children; humanitarian and relief workers; and prisoners are core transmitter group.

Lisa Garbus (2003) has indicated that gender disparities in enrolment ratios and educational attainment levels are high which may obviously be one reason to imbalance the chance of infection between male and female group. The country's high maternal mortality ratio is also an indication not only of poor reproductive health, but of women's low status and poor access to basic health services. Many Ethiopian women have little power in sexual negotiation with their husbands. Almost 14 percent of currently married women in Ethiopia are in a polygynous union. About 80 percent of have been circumcised. Poverty and unemployment are leading to a dramatic increase in the trafficking of Ethiopian women. Other issues that render Ethiopian women vulnerable to HIV include rape, abduction, and early marriage (Lisa Garbus 2003).

2.4 Impact of HIV/AIDS in Ethiopia

The impact of HIV/AIDS, just like in any other country, is quite perceivable even in a more severe state. Ethiopia is one of the countries in sub-Saharan Africa most severely affected by HIV. In Ethiopia, AIDS is now recognized as the leading cause of adult morbidity and mortality in the country. According to Lisa Garbus 2003, emphasized Ethiopia's population will be up to 16 percent smaller than it would have been in a "no-AIDS" scenario. The writer also mentioned that AIDS will reduce life

expectancy by 9 to 13 percent through 2050. AIDS has already increased the number of deaths in Ethiopia by 6 percent. Between 2000 and 2015, it will increase the number of deaths in Ethiopia by 27 percent. By the end of 2002, 1.7 million Ethiopians had died because of AIDS. By 2014, there will be a cumulative total of 5.3 million AIDS deaths (Lisa Garbus 2003).

The work of Lisa Garbus showed that, more than 42 percent of all hospital beds in the country are estimated to be occupied by AIDS patients. Ethiopia has the sixth-highest number of TB cases in the world. About 42 percent of adult (15-49) TB cases were HIV-positive during 2000. The adult HIV prevalence in Ethiopia as estimated and projected from various data of NSS results showed constantly incremental behavior. The following table implies the national HIV prevalent by sex and setting for selected years (1982-2008).

	1982	1985	1990	1995	2000	2001	2002	2003	2008
National	0.0	0.2	1.6	3.2	3.9	4.1	4.2	4.4	5.0
Male	0.0	0.2	1.5	2.8	3.4	3.4	3.7	3.8	4.4
Female	0.0	0.2	1.7	3.6	4.4	4.8	4.8	5.0	5.7
Urban	0.0	0.7	7.0	13.4	13.0	12.8	12.7	12.6	12.6
Rural	0.0	0.1	0.3	0.8	1.9	2.1	2.4	2.6	3.4

Table 2.5 Estimated and projected adult HIV Prevalence (%) by sex and setting .selected years
[Source MOH 2004]

According to MOH 2004, possible impact analysis on different variables shows that HIV/AIDS has a visible effect and detrimental impact on Ethiopian society and economy. The report also points out that, findings from studies conducted at schools, workplaces, and among orphans also indicate similar effects. Some of the important impact analysis on the 2004 result and extrapolations to future years, made by MOH 2004, are shortly reviewed in the following few paragraphs.

HIV/AIDS counted for an estimated 38% or 54,000 of all TB case incidences in 2003. And this proportion is expected to continue to rise in the coming years, contributing to a total projected TB caseload of 180,000 in 2008. The impact of AIDS on the total population size of the country is another devastating effect. The report on MOH 20003, figure out that the population lost to AIDS by the year 2003 and was 900,000 and is projected to reach 1.8 million by 2008 if present trend continues. The report

also mentioned that adult (15-49 years) deaths due to AIDS are expected to rise tremendously in the coming years and directly account for about a third of all young adult deaths in the country.

According to the survey report of MOH 2004, there was a 5% increase in death among teachers, some of which might be attributed to AIDS. Moreover, absenteeism of at least one week out of a semester was reported among a third of the teachers due to sickness of the teacher or members of his/her family. Educational costs are increasing due to the replacement of non-productive teachers. Overall school drop out rates is reported increasingly possibly because of sickness and death of parents. The same study showed that orphan students who lost their parents presumably due to AIDS and associated disease were more likely to drop out of school and repeat classes than their non-orphan counterparts.

The industrial sector is also greatly affected by the presence of the HIV/AIDS epidemic among the work force. A study conducted by the Employers Federation of Ethiopia in 2002 (quoted in MOH 2004), showed that the impact is most severe in the wholesale and retail trades followed by the manufacturing, agriculture and public service sectors. The major effects are found to be due to the following: (MOH 2004)

- HIV-related illnesses in workers resulting in reduced productivity
- Many skilled workers quit their jobs because of sickness
- Shortages of skilled manpower due to HIV/AIDS
- Increased mortality in the work force due to HIV/AIDS
- Increased absenteeism
- Rising medical costs due to HIV/AIDS

2.5. Efforts Made To Alleviate HIV/AIDS in Ethiopia

In 1985 (before the first AIDS case had been officially diagnosed), the government of Ethiopia established a national HIV/AIDS task force. It issued the first AIDS control strategy at the end of 1985. In 1987, the government established an HIV/AIDS department within the Ministry of Health. The MOH developed a short-term plan in 1987 and medium-term plans in 1987 and 1992. As in many countries, these interventions were inadequate in scale; largely ineffective in implementation; lacked sufficient stakeholder involvement in planning and implementation, especially at the

community level: were poorly or not at all coordinated and integrated across sectors and among service providers; and received relatively low priority within government, society in general, and in the international community, with a resultant low level of allocated financial and human resources.

In 1989, the MOH drafted a four-point policy statement on AIDS prevention. The first draft of a national policy was created in 1991, though not approved until 1998. The National AIDS Prevention and Control Council was established in 2000 and is charged with implementing the Strategic Framework for the National Response to HIV/AIDS in Ethiopia for 2000-2004. The council, chaired by the president of Ethiopia and comprising members from government, NGOs, religious bodies, and civil society, has declared HIV/AIDS a national emergency.

The participatory efforts of various bodies of the society against HIV/AIDS are seen gradually. Lisa Garbus discussed that the Ministry of Education has yet to provide clear guidelines for a comprehensive HIV education curriculum. Despite that the agricultural and livestock sectors account for over 85 percent of the labor force, the Ministry of Agriculture is not actively involved in HIV prevention.

Although Ethiopia began the HIV/AIDS policy process in 1989 - far earlier than most African countries - it took nine years to complete. Moreover, the process involved almost no participation by actors outside government. There was little high-level political involvement in HIV/AIDS during the 1990s. By 1999, however, President Negasso Gidada had become an active spokesman in the fight against HIV/AIDS. (Lisa Garbus 2003). According to this writer, Ethiopia's response is constrained by extremely limited human, financial, technical, material, and management capacities. The country's health care infrastructure is massively under-equipped to address HIV/AIDS, particularly outside Addis Ababa. The country's size and poor transport infrastructure are also key factors. Bureaucratic constraints are impeding the response. Although much more funding is required, building additional human capacity in the health sector, improving coordination, and using existing funds efficiently and effectively are paramount.

The data on HIV and HIV-related cases in the country has getting huge in size. But still when we look at the nature of the reports above, the reports are aggregated by age and sex while the case may have a number of attributes to be studied together.

So the intent of this experimental research is to change/improve the view about the data rather than only seeing it with few dimensions like age and sex. So in this research, the core and frontier risk factors for HIV infection had been identified using data mining tools and their associations were also investigated and good association rules have been generated and discussed.

2. 6. Voluntary Counselling and Testing

Voluntary HIV counselling and testing (VCT) is the process by which an individual undergoes counselling enabling him or her to make an informed choice about being tested for HIV. This decision must be entirely the choice of the individual and he or she must be assured that the process will be confidential.

HIV voluntary counselling and testing (VCT) has been shown to have a role in both HIV prevention and, for people with HIV infection, as an entry point to care. HIV provides people with an opportunity to learn and accept their HIV serostatus in a confidential environment with counselling and referral for ongoing emotional support and medical care. People who have been tested seropositive can benefit from earlier appropriate medical care and interventions to treat and / or prevent HIV - associated illnesses. Pregnant women who are aware of their seropositive status can prevent transmission to their infants. Knowledge of HIV serostatus can also help people to make decisions to protect themselves and their sexual partners from infection. According to UNAIDS (2000) technical update, studies have indicated that VCT may be a relatively cost-effective intervention in preventing HIV transmission.

2.6.1. Elements of VCT HIV Counselling

HIV counselling has been defined as "a confidential dialogue between a person and a care provider aimed at enabling the person to cope with stress and make personal decisions related to HIV/AIDS. The counselling process includes an evaluation of personal risk of HIV transmission and facilitation of preventive behaviour." (WHO 1994). The objectives of HIV counselling are the prevention of HIV transmission and the emotional support of those who wish to consider HIV testing, both to help them make a decision about whether or not to be tested, and to provide support and facilitated decision-making following testing. With the consent of the client, counselling can be extended to spouses and/ or other sexual partners and other supportive family members or trusted friends where appropriate.

HIV counselling can be carried out anywhere that provides an environment that ensures confidentiality and allows for private discussion of sexual matters and personal worries. Counselling must be flexible and focused on the individual client's specific needs and situation.

2.6.2. Voluntary Testing

HIV testing may have far-reaching implications and consequences for the person being tested. Although there are important benefits to knowing one's HIV status, HIV is, in many communities, a stigmatizing condition, and this can lead to negative outcomes for some people following testing. Stigma may actively prevent people accessing care, gaining support, and preventing onward transmission. That is why UNAIDS stipulates testing should be voluntary, and VCT should take place in collaboration with stigma-reducing activities.

2.6.3. The Counselling Process

The VCT process consists of pretest, post-test and follow-up counselling. HIV counselling can be adapted to the needs of the client/s and can be for individuals, couples, families and children and should be adapted to the needs and capacities of the settings in which it is to be delivered. The content and approach may vary considerably for men and women and with various groups, such as counselling for young people, married people, unmarried individuals, or sex workers (UNAIDS 2000). So, as mentioned above, counselling as part of VCT ideally involves at least two sessions (pretest counselling and post-test counselling). More sessions can be offered before or after the test, or during the time the client is waiting for test results.

2. 7. The Reason of Choosing VCT Data for this Research

As already been described in the previous section, VCT is a relatively cost-effective intervention in preventing HIV transmission is considered as best strategy in HIV/AIDS interventions. VCT service encompasses or addresses general population groups. "When properly carried out, voluntary counselling and testing can help break the vicious circle of fear, stigma and denial"(UNAIDS, 2000).

Other STD-related treatments or services more directly focused on pregnant women, young people and TB clinic clientele, but VCT service models include the general population. Though, VCT looks a relatively new programme intervention with the few services available being concentrated, the service remains to be of high importance since the general population of the society is benefited. So VCT is not limited to any population group and the data from such service will confidently represent the mass unlike the case of ANC data which is taken from only few portion of the society. Secondly, the availability of the data in electronic format has made the VCT data to be appropriate source to conduct this experimental research.

CHAPTER THREE

3. Data Mining and Knowledge Discovery

3.1. Introduction

With the positive effect of capacity and technology advancement to collect, store and maintain data is raised with drastically high rate, the ability to analyse and understand a huge commutation of data was not so easy until recently. The resolution for such problem of understanding huge data has been possible as effective of the advent of data mining technology. The data that are generated and stored continuously grow into large database scaling to giga, and even tera, bytes as mentioned in Deogun et.al (2001). Similarly, Witten & Frank, (2000) emphasises the matter that, the amount of information in the world doubles every 20 months at a rate of 100%. The downside of this phenomenal growth is inability to analyse & understand the data.

The price affordability and wide spread of devices especially, the ever decrease of computer price, enabled organisations to collect and store large collection of data. As Al-Attar (1999) states most organisations are currently labelled as 'data rich'. But the data are usually used only to provide endless facts and figures. Such facts and figures do not represent knowledge and can be considered, in the view of knowledge seekers as a virgin Oar that needs to be exploited further in stead of simply creating information overload. Al-After (1999), further discuss that most organisations are still knowledge poor. This is because it is the pattern in the data that represent knowledge.

Organisations keep on maintaining records of data generated even longer after their life time has expired because they believe that there is valuable information implicitly coded within it (Fayyad, 1996). When data collection mechanism was not advanced as it is so today, it was relatively easy to analyse data manually. But as the data collections continues to be an unmanageable in size as well as complexity, the demand for more sophisticated techniques & tools for data analysis has emerged. One of these techniques is data mining (SMCC 1997)

The concept data mining had been defined /explained/ by different scholars with different phrases, but almost conveying same understanding. The intent of all the meaning is that data mining is a technique and tool that enable data analyst to easily reach at depth in large data and explore the important knowledge hidden in it.

Data mining is defined as the process of extracting valid, previously unknown, comprehensible and actionable information from large databases and using it to make critical business decisions (Connolly et. al 1999). The authors stress that the main target of data mining is in searching important information that is hidden and unexpected. This is because most organizations have large accumulated data while what is important to them is information.

Data mining can be considered as one of the most important frontiers in databases systems and one of the most promising interdisciplinary developments in information industry (Han and Kamber, 2001). The writers further discuss that interesting knowledge, regularities or high level information can be extracted from databases and viewed from different perspectives. After discovering previously unknown features of data, organizations apply the knowledge to their business processes in general.

The explosive growth in data collection in business and scientific fields has literally forced the need to analyze and mine useful knowledge from it. Data mining refers to the entire process of extracting useful and novel patterns/models from large datasets. It is a process through which interesting and previously unknown patterns and correlations can be extracted automatically from a large database of information. The implicit information within databases, and mainly the interesting association relationships among sets of objects, that lead to association rules, may disclose useful patterns for decision support, financial forecast, marketing policies, medical diagnosis and many other applications. Association rule mining involves detecting items, which tend to occur together in transactions, and the association rules that relate them.

3.2. Knowledge Discovery in Databases

In conducting a data mining activity, series of steps are to be followed to complete the entire process. The steps of which will be discussed in one of the subsequent

sections. The phrase KDD which has manifested itself at the first KDD workshop in 1989, emphasize that knowledge is the end product of database findings. Some misunderstanding avails among researchers and intellectuals in the field of data mining about the two phrases (Data mining & Knowledge Discovery in Databases). With many people data mining is treated as a synonym for KDD, while others view it as simply an essential step in the process of KDD.

Fayyad et.al.(1996) describe the difference as; KDD is the overall process of discovering useful knowledge from data, and data mining is a particular step in this process. The authors further explain that data mining is the application of specific algorithms for extracting patterns from data. Despite of the view of the quoted authors, other authors like Han & Kamber (2001) believe that the term data mining has become more popular as a synonym for KDD and can be used interchangeably. And through out this research paper this view of understanding and using the two phrases interchangeably is applicable.

It is also of important to mention some of the means by which the KDD addressed. Fayyad et.al.(1996), listed a number of phrases with which the notion of finding useful patterns in raw data including KDD, data mining, knowledge extraction, information discovery, information harvesting, data archaeology, and data pattern processing.

3.3. Steps of Knowledge Discovery in Databases

As aforementioned earlier the KDD process is not a single step activity. Rather KDD is an iterative, multi-level sequence of steps, which are, data cleaning, data integration, data selection, data transformation, data mining, pattern evaluation, and knowledge presentation (Han & Kamber 2001). But according to other writers (TCC 1999), the process of building and implementing a data mining solution even starts at formulating the business problem at hand.

The entire Knowledge Discovery in Databases process as discussed by Fayyad et.al(1996) is indicated as follows:

3.3.1. Problem Definition

Developing an understanding of the application domain, the relevant prior knowledge, and the goals of the end-user are considered to be the first step towards a successful data mining activity (Fayyad et.al, 1996). In another words, the starting block in a data mining activity is formulating the problem, which is very crucial step of the overall set of activity. So any mining process should start with a clear definition of the business problem involved and the objective function.

3.3.2. Selecting a Target Data Set

The second step involves selecting of data, which is also the main ingredient in data mining activity, because the success of the result depends on the data used. Meretwork (2004) emphasis the importance of selection step in the KDD process because databases are heterogeneous, containing a wide variety of data, not all of which may be appropriate for the analysis at hand. So focusing on a subset of variables or data samples, on which discovery is to be performed is crucial. Lavin and Zehavi (1999) quoted in Meretwork (2004) pointed that one can use subjective judgment to extract the relevant target set, but in other cases segmentation analysis have to be used.

3.3.3. Data Cleaning and Preprocessing

Data cleaning preprocessing is another important step of KDD that is about basic operations such as the removal of noise or outliers, deciding on strategies for handling missing data fields. Fayyad et.al (1996) indicated that this step is also about accounting for time sequence information and known changes as well as joining data from various sources to create hemogeneous source by resolving representations and encoding differences. Various literatures mention that data collection, cleaning and preprocessing is the part that takes high percent of the time and effort of the entire data mining process.

3.3.4. Data Reduction and Projection

Since data mining generally provides improved performance with the normalized data, all the data sets need to be transformed into values that are convenient for the selected tools. This fourth step, data reduction and transformation, is the step where

data is transformed into forms appropriate for mining by performing summary or aggregation operation. This also includes finding useful features to represent the data depending on the goal of task. In this stage dimensionality reduction and transformation methods are used to reduce the effective number of variables under consideration or to find invariant representation for the data.

3.3.5. Choosing Mining Tasks and Algorithms

The mining task starts by deciding whether the goal of the KDD process is classification, regression, clustering, association mining, etc. After the goal is defined, selecting appropriate mining algorithm(s) and method(s) to be used for searching patterns in the data follows. This process includes deciding which models and parameters might be appropriate and matching a particular data mining method with the overall criteria of the KDD process.

3.3.6. Data Mining and Model Building

The sixth step, data mining, is an essential process where intelligent methods are applied in order to extract new and interesting data patterns. This data mining step is an essential step because it uncovers hidden pattern for evaluation. Kamber and Han (2001) pointed out that this step may involve users' interaction or knowledge base. These interesting patterns are presented to the user and may be stored as a new knowledge in a knowledge base.

3.3.7. Interpreting and Evaluation of Mined Pattern

In the seventh step, which is pattern evaluation and knowledge representation, the model is tested for accuracy and validation on an independent dataset that has not been used to create the model. This validation stage is a step where the truly interesting patterns, based on some interesting measures, are identified and assessment of the model for sensitivity as well as pilot testing of the model for usability is also performed. The step may also involve visualization of the extracted pattern and may need possible return to any of the previous steps for further iteration.

3.3.8. Consolidating Discovered Knowledge

The final step, deployment of the model, is incorporating this knowledge into the performance system, or simply documenting it and reporting it to interested parties. This deployment also involves checking for and resolving potential conflicts with previously believed or extracted knowledge.

After deploying the model and successful usage of the discovered knowledge, the only further idea to think off is monitoring the model. This is because of the fact that there may be always changes in human needs and conditions at the domain may vary as function of time. This means, a model that is perfect to the need of an organization at one time may no longer be regarded as a perfect one after long time. So, frequently and constantly revalidation of model on new data is required to check if the model lasts appropriate.

However, it does not mean that the steps of KDD process are carried out sequentially. Data mining is an iterative and multi step process: 'The basic step (extracting the data from a database, prepare and otherwise cleaning it, run the knowledge discovery algorithms on it, analyze the results, and write it as repeated again and again. And each step can conceivably returned to any previous ones' (Askale, 2001).

The current experimental research work also followed accepted stages of a data mining process to identify the core risk factors that are related to HIV infection and mine their affinity analysis.

3.4. Multidisciplinary Nature of KDD

Many researchers in the field of machine learning, pattern recognition, databases, statistics, artificial intelligence, knowledge acquisition for expert system and data visualization have been shown keen interest in KDD.

Whitten and Frank (2000) mentioned about relationship of statistics and machine learning in a more detail expressions. Regarding their differences, statistics has been more concerned with testing hypothesis, where as machine learning is has been concerned with formulating the process of generalization as a search through possible hypothesis. And regarding their similarity, very similar schemes have been

developed in parallel, in both fields. Statisticians published a book on classification and regression trees. While about closer period of time, a machine learning researches were developing a system for inferring classification trees from examples. Another similarity between the two is use of nearest-neighbour method for classification. These are standard statistical techniques that have been extensively adopted by machine learning researches.

Fayyad et.al (1996), explained that the knowledge discovery in databases systems typically draw upon methods, algorithms and techniques from these diversified fields with a unifying goal of extracting knowledge from data in the context of large databases. The authors further discussed the links, relations and commonality between KDD and other fields as enumerated below.

- In the fields of machine learning and pattern recognitions, their overlap with KDD lies in the study of theories and algorithms for systems which extract patterns and models from data (mainly data mining methods). KDD focuses on the extension of these theories and algorithms to the problem of finding special patterns in large sets of real world data.
- Knowledge Discovery in databases also has much common land with statistics, particularly exploring data analysis (EDA). Knowledge Discovery in Databases systems often embeds particular statistical procedures for modelling data and handling noise within an overall knowledge discovery framework.
- Machine discovery which targets the discovery of empirical laws from observation and experimentation (Shrager and Langley 1990) as quoted in Fayyad (1996), and causal modelling for the inference of causal models from data are also related research areas.
- Another related area is data warehousing which refers to the recently popular Management Information Systems trends for collecting and cleaning transactional data and making them available for on-line retrieval. A popular approach for analysis of data warehouses has been OLAP (on-line analytical processing) whose tools focus on providing multidimensional data analysis. Hence both knowledge discovery and OLAP are viewed as

related facets of a new generation of intelligent information extraction and management tools.

The conceptual and practical convergence of these fields has contributed and applied together for specific application. These all together made the field of data mining to be considered as a multidisciplinary practical and experimental research area that has capability of solving business problems. At this point, it is still necessary to notice that data mining has not replaced any of these techniques but strengthens them in one way or another.

3.5. Data Mining Versus Other Statistical Tools

Data analysts have been using different tools for data analysis even for centuries to understand the natural world. Moxon(1996) stated that there have long existed many analytical tools that support verification based approach. Irrespective of the existence and capabilities of the many analytical tools, data mining tools exhibited differences in many aspects. Askale (2001) pointed that in the analytical tools, it is the user that puts a hypothesis about specific data relationships and then the tools would help in verifying or disproving the hypothesis. In other phrases, statistical analysis concerns itself with analysing primary data that has been collected to check specific research hypothesis, while data mining can also concern itself with secondary data collected for other reasons.

The other three aspects that distinguish statistical data analysis from data mining, as listed by Giudici 2003, are:

- First, data mining analyses huge amount of data while statistical analysis limits itself to small data set
- Second, many databases do not lead to the classic forms of statistical data organizations, but data mining approach can still supports the analysis.
- Third, data mining results must be of some consequence. This means that constant attention and emphasis need be given to business results achieved with analysis model.

Business problems can be solved using various statistical methods and many algorithms. And the choice of a method depends on the underlying problem or the type of available data. Statistical tools are typically used to address the business

problem of generating an overview of the data in databases. This is done by using techniques that summarize information about the data into statistical measures that can be interpreted without requiring every record in the databases to be understood in detail. An example can be the application of statistical functions like finding the maximum and minimum, the mean or the variance.

Askale (2001) indicated that query and reporting tools integrate the data and report on any pattern (query) that is requested by the user. This type of discovery is possible only if the user prompts it. The writer further described this as, unless the users suspect a pattern they will never find it. This shows that the traditional tools have limitations in that results depend on the ability of the analyst to formulate appropriate questions. Moreover, when the number of the variables being analysed are too many it would be difficult to identify the important variables (TCC, 1999). Data mining is different because rather than verifying hypothetical patterns, it uses data itself to uncover such patterns.

According to Trybula (1997), machine learning and statistical techniques can be used to automatically extract information from data, but data mining which is also called knowledge discovery in databases, is an efficient discovery of valuable, non-obvious information from large collection of data. Data mining centres on the automated discovery of new facts, hidden patterns and relationships in data.

Though, data mining technique is discussed to a better means of data understanding utilization, it does not mean that data mining has replaced other techniques such as OLAP, query, reporting, etc. Rather data mining complements and integrates with other decision support system capabilities such as query and reporting, on-line analytical processing, data visualization and traditional statistical analysis.

3.6. Data Mining Activities

When we see it at high level mining, the two primary goals of data mining, in practice, are the prediction and description in general. Predictive modelling consists of several types of model such as classification, regression and AI-based models and are referred to as supervised learning, since calculated or estimated values are compared with known results. Whereas descriptive techniques are referred to as unsupervised learning which interrogate the database to identify patterns and

relationships in the data. Association mining, clustering analysis, pattern recognition models, visualization methods, among the many are categorized under descriptive mining.

According to Fayyad et.al.(1996), prediction involves using some variables or fields in the database to predict unknown or future values of other variables of interest. And description focuses on finding human-interpretable patterns describing the data. The authors further state that the relative importance of prediction and description for particular data mining applications may vary considerably; however, in context of KDD, description tends to be more important than prediction.

Data mining could be implemented to carry out different tasks. Some of the major activities, among the many, that could be carried out using data mining are: classification, estimation, prediction, affinity grouping or association rules, clustering, description and visualization [Berry & Linoff 2000]. The goals of predictions and description are achieved by using these stated tasks or activities of data mining.

Classification which many researchers in the area of data mining, perceive as the most common mining task, is a process of finding a set of models that describe and distinguish data classes or concepts, which enables to predict the class of objects whose class label is known. So in developing a classification model, a set of preclassified examples are used to develop a model that can classify new data into one of predefined set of classes. Different authors like Han & Kamber 2001, indicated that classification is a two step process. In the first step is model building using existing data and in the second step the model is used for classification of new objects after the accuracy of the model is checked with test data.

The second data mining activity is estimation, which deals with continuous valued outcomes. Unlike classification which deals with discrete outcomes (like "Yes" or "No"), estimation deals with continuous valued outcomes. Estimation approach measures a worthiness score of some unknown continuous variable rather than categorizing the outcome into predefined label.

The other common data mining task is affinity grouping; also called finding association rules, which is used to determine which things often go together. Association approaches are most common in market basket analysis. Affinities are

expressed in terms of support and confidence rules such as percent of all objects in which some condition holds true when some other factor is true. For this experimental research, affinity grouping of data mining was carried out since the model to be built is finding association of the attributes (or risk factors) that mostly go together for individuals whose HIV test result reviles positive. The association rule mining was not conducted for all the attributes of the clients' data. Attribute selection was carried out using other mining activities (Decision tree classification in KnowledgeSTUDIO), and the affinity grouping were examined for the risk factors contribute more for HIV infection.

Prediction is also another mining activity that can be thought of as classification or estimation (Berry & Linoff 2000). Prediction tasks feel different because the records are classified according to some predefined future behaviour or estimated future value.

Any of the techniques used for classification and estimation can be obtained for use in prediction by using training examples where the value of the variable to be predicted is already known, along with historical data for those examples. The historical data is used to build a model that explain the current observed behaviour and this model is applied to current inputs and the result is a prediction of future behaviour.

The other data mining activity is clustering which deals with the task of segmenting a diversified population of objects into a number of more homogeneous subgroups or clusters. Clustering is similar to classification in that the task is about organization of data into classes, but clustering is different from classification since there is no predefined class in clustering. Most of the time clustering is the first step in data mining analysis for it would be used to identify related records and then other data mining tasks could be performed to explore further relationships.

There are other data mining tasks that increase our understanding of the people, products, or processes that produced the data in the first place. Some times the purpose of data mining may be simply to describe what is going on in a completed database. Such mining activities which simply uncover a good description of behaviour is description and visualization. Therefore, data visualization is also one powerful descriptive data mining activity.

3.7. Application of Data mining

Much of the data mining technology, including mining activities, mining process, its relevance with other tools, are seen in the previous sections of this part. Now it is time to take a step to see the promising applications of data mining.

Data mining has become very popular because of its successful application in diversified areas. Nowadays, many companies are benefiting from data mining. The popularity and substantial contribution it can make spans from service giving organizations to production enterprises.

Many organizations are using data mining to help improve their entire business processes including acquiring new customer, in increasing revenue from existing customers, and retaining good customers. By profiling customers who have bought a particular product a company can focus its attention on similar customers who have not bought that product. By profiling customers who have left, a company can act to retain customers who are at risk for leaving (Two Crown, 1999). The authors also mention that telecommunications and credit card companies are the two leaders in applying data mining to detect fraudulent use of their services. Insurance companies and stock exchanges have also shown interest in applying data mining to reduce fraud.

According to Kamber and Ham 2001, data mining is also applicable for biological and DNA data analysis. Biomedical researches include the development of new pharmaceuticals and advances in cancer therapies to the identification and study of human gene by discovering large-scale sequencing patterns. The authors added that recent research in DNA analysis has led to the discovery of new medicine and approaches for disease diagnosis, prevention, and treatment.

Data mining now days is also used in pharmaceutical forms to mine large data bases of chemical compounds and of genetic material to discover substances that might be candidates for development as agents for the treatments of disease, as well as analysing the recent sales of the company (Berry & Linoff 2000).

Data mining techniques have been found fruitful in medical applications where it can be used to analyse and even to predict the effectiveness of surgical procedures,

medical tests or medications. On this point Shegaw (2002), On developing predictive model for child mortality, has applied data mining technology with a good achievement research report.

3. 8. Application of Data Mining In Health Sectors

Ballenger et.al.(1999) write that data mining is a 'horizontal application meaning it is not specific to any industry.' Ballenger added that the ingredients that are important for data mining are availability of data and the willingness to explore the possibility of hidden knowledge that resides in the data. However, there are industries that have already made significant progress in the application of data mining. One of these domains is the health sector.

According to Larvac (1998), in the health sector, the widespread use of medical information systems and explosive growth of medical databases require traditional manual data analysis to be completed with methods for efficient computer assisted analysis. Such an extensive amount of data gathered in medical databases requires specialized tools and methods that can be used to discover new information and knowledge which is useful in decision making and problem solving.

With the pressure from various factors and the need to enhance quality of services, the health industry has started to mine its data to cut costs, improve quality and save lives. In this regard, Bresnahan (1997) stated that one way in which data mining is helping health service providers cut costs and improve service by showing which treatments statistically have been most effective. For instance, once the hospital counsellors, and hence the administrators, recognize that HIV/AIDS test volunteers from similar background (as collected from their personal data), are falling in same classes like positive HIV test results, the hospital can devise better strategies towards efforts to provide effective counselling and treatment including planning to educate the mass of that group.

The application of data mining is nowadays found to be of very much important and knowledge discovery can also close the loop between clinical data capture and evidence-based decision support by facilitating the conversion of data into evidence for future decision (Downs and Wallace, 2001). Data mining techniques have also been applied for survival analysis and the prediction of prognosis in cancer treatment

(Theeuwes, Kappen, and Neijt, 2001). Plate et.al (1997) also described that although epidemiological data is traditionally analyzed with very simple statistical techniques; flexible models such as neural networks have the potential to discover unanticipated features in data.

Shegaw(2002) pointed that the application and utilization of data mining technology in the health sector is steadily growing fast in the developed world. But its applicability remains to be unappreciated in the developing countries in general, and in the Ethiopian health context specifically. Given experiences elsewhere in terms of the benefits acquired in applying data mining technology in the health sector, it is only proper to explore the relevance and potential advantage of such a powerful and promising technology in the Ethiopian context. Thus, in this study, since there is high and astonishing rate of HIV/AIDS related problem, the researcher is motivated to assess the potential applicability of data mining technology to develop a model that can support clinics, hospitals, policy makers, and planners etc, in Ethiopia by using a VCT database for two reasons: the first reason is the availability of the data in electronic format, and the second reason is the fact that the VCT service encompasses the general population. The researcher has chosen the Center for Disease Control and Prevention, (CDC) to conduct the study, because CDC has collection of relevant data.

CHAPTER FOUR

4. DATABASES AND ASSOCIATION MINING

4.1. Mining Association Rules in Large Databases

In section 3.6 of this chapter, where data mining functionalities were discussed, it has been pointed out that association rules mining is one of the types of patterns that can be mined. Data mining functionalities are used to specify the kind of patterns to be found in data mining task. In general, as indicated in Han & Kamber (2001), data mining tasks can be classified into two main categories as descriptive mining and predictive mining. Affinity grouping or association rules mining is one of the descriptive mining tasks that characterize the general properties of the data in the database.

Association rules mining finds interesting association or correlation relationships among a large set of data items (Han & Kamber, 2001). With massive amounts of data continuously being collected and stored, many organizations are becoming interested in mining association rules from their databases. The discovery of interesting association relationships among large amounts of transactional records can help in decision making processes, such as catalog designing, cross marketing, loss leading analysis and support in medical diagnosis as well.

So according to Han & Kamber (2001), association mining which is also called association analysis or affinity grouping is the discovery of association rules showing attribute values conditions that occur frequently together in a given set of data. Although association analysis is widely used for market basket or transaction analysis, it has been also used in various areas.

More formally, association rules are of the form $X \Rightarrow Y$, that is:

" $A_1 I_{\lambda_1} \dots I_{\lambda_m} \Rightarrow B_1 I_{\lambda_1} \dots I_{\lambda_n}$ " Where

A_i (for $i \in \{1, \lambda, m\}$) and B_j (for $j \in \{1, \lambda, n\}$)

are attribute value pairs. The association rule $X \Rightarrow Y$ is interpreted as database tuples that satisfy the conditions in X will also likely to satisfy the conditions in Y. So

association rule mining searches for interesting relationships among objects in a given data set.

To illustrate this concept with examples, given a database, a data mining system may find association rules like:

age(*person*, "20---29) and **marriage**(*person*, "single") ==> **test**(*person*, "infected") with [support = 2%, confidence = 60%]

where *person* is a variable representing a customer and age, marriage, & test are attributes that describe the person. "20 --29", "single" and "infected" are the respective attribute values. The rule in this example indicates that, of all the customers under study, 2% (support) are 20 to 29 years of age with marital status of single and its HIV test result reads positive(infected) at the database of the given service center. There is 60% probability (confidence or certainty) that the customers in this age and who is unmarried is likely to be infected with the virus.

According to Whitten and Frank (2000), Association rules like classification rules attempts to predict the class given a set of conditions on the LHS (Left Hand Side). The conditions are typically attribute value pairs, also referred to as item-sets. However, the main difference is that the prediction associated with the RHS (Right Hand Side) of the rule is not confined to a single class attribute, instead it can be associated with one or more attribute combinations.

4.2. Interestingness measures

Interestingness measures are one of the data mining primitives that need to be considered in the data mining process. These functions are used to separate uninteresting patterns from knowledge discovered. They may be used to guide the mining process or, after discovery, to evaluate the discovered patterns. Different kinds of knowledge may have different interestingness measures.

According to Han & Kamber (2001), although specification of the relevant data, the kind of knowledge to be mined, and background knowledge may substantially reduce the number of patterns generated: a data mining process may still generate a large number of patterns. Typically, only a small fraction of these patterns will actually be of interest to the given user. Thus, users need to further confine the number of

uninteresting patterns returned by the process. This can be achieved by specifying interestingness measures that estimate the simplicity, certainty, utility, and novelty, of patterns.

Simplicity, as a measure of patterns interestingness, is a factor contributing to the overall simplicity for human comprehension about the discovered pattern. Objective measures of pattern simplicity, as discussed in Han & Kamber (2001), can be viewed as functions of the pattern structure, defined in terms of the pattern size in bits, or the number of attributes or operations appearing in the pattern. This implies that the more complex the structure of the rule, the more difficult it is to interpret, and hence the less interesting it is likely to be. Rule length[▼] is also another interestingness measure.

Rule support and confidence are two measures of rule interestingness that can validate the usability of the rules generated. They respectively reflect the usefulness and certainty of discovered rules. For instance, a support of 2% for an association rule means that 2% of all the transactions under analysis show that some variable (X) and another one (Y) usually go together. A confidence of 60% means that 60% of the customers who exhibited some feature on one variable (X) also have similar feature on another variable(Y). Typically, association rules are considered to be interesting if they satisfy both a minimum support threshold and a minimum confidence threshold. Such thresholds can usually be set by users or domain experts.

Each discovered pattern should have a measure of certainty associated with it that assess the validity or trustworthiness of the pattern. A certainty measure for association rules of the form " $A \Rightarrow B$ ", where A and B are sets of items in a database (D), is confidence. Given a set of task-relevant data tuples the confidence of " $A \Rightarrow B$ " is defined as

▼ For rules expressed in conjunctive normal form (as a set of conjunctive predicates), rule length is typically defined as the number of conjuncts in the rule.

$$\text{confidence } (A \Rightarrow B) = \frac{(\# \text{ tuples containing both } A \text{ and } B)}{(\# \text{ tuples containing } A)}$$

Equation 4. 1. Formula to Measure Confidence of Rule

A confidence value of 100%, or 1, indicates that the rule is always correct on the data analysed and such rules are called *exact*.

The potential usefulness of a pattern is a factor defining its interestingness. It can be estimated by a utility function, such as support. The support of an association pattern refers to the percentage of task-relevant data tuples for which the pattern is true. For association rule of the form " $A \Rightarrow B$ " where A and B are sets of items, it is defined as:

$$\text{Support}(A \Rightarrow B) = \frac{(\# \text{ tuples containing both } A \text{ and } B)}{(\text{total } \# \text{ tuples})}$$

Equation 4. 2. Formula to Measure Support of Rule

In other illustrative terms, the rule $A \Rightarrow B$ that holds in the transaction set D with support s , where s is the percentage of transactions in D that contain $A \cup B$ (i.e., both A and B). This is taken to be the probability, $P(A \cup B)$. This means, support $(A \Rightarrow B) = P(A \cup B)$. Similarly, the rule $A \Rightarrow B$ has confidence c in the transaction set D if c is the percentage of transaction in D containing A that also contain B . This taken to be the conditional probability, $P(B|A)$. That is, confidence $(A \Rightarrow B) = P(B|A)$.

Association rules that satisfy both a user specified minimum confidence threshold and user defined minimum support threshold are referred to as strong association rules and are considered interesting. Rules with low support likely represent noise, or rare or exceptional cases.

In general, each measure is associated with a threshold that can be controlled by the user. Rules that do not meet the threshold are considered uninteresting, and hence are not presented to the user as knowledge.

Association rule mining is a two-step process as Han & Kamber (2001) explained it as follows. This is the most common approach to find association rules by breaking up the problem into two parts.

1. The first step is finding all frequent itemsets which, by definition, must occur at least as frequent as a pre-defined minimum support count.
2. The second step is generating strong association rules from the frequent itemsets which, by definition, must satisfy minimum support and minimum confidence.

According to Sivanandam et.al.2004, the definition of a frequent pattern relies on the following considerations. A set of items is referred to as an itemset (pattern). An itemset that contains k items is a k -itemset. The set $\{X, Y\}$ is a 2- itemset. The occurrence frequency of an itemset is the number of transactions that contain the itemset. This is also known as the frequency or the support count of an itemset. An itemset satisfies *minimum support* if the occurrence frequency of the itemset is greater than or equal to the *minimal support threshold value* defined by the user.

Discovering the frequent itemsets requires a lot of computation power, memory and I/O, which can only be provided by parallel computers. As data is increasing in terms of the dimensions (number of items) and size (number of transactions), one of the main attributes needed in an association rule-mining algorithm is scalability[¶]. Sequential algorithms cannot provide scalability, in terms of the data dimension, size, or runtime performance, for large databases. Therefore, high-performance parallel and distributed computing can be employed for enhancing the performance, scalability (Sivanandam et.al.2004).

A frequent itemset is called *closed* if it does not have any superset with the same support. A frequent itemset is said to be *maximal* if it has no supersets that are frequent. The collection of maximal frequent itemsets is a subset of the collection of closed frequent itemsets, which is a subset of the collection of all frequent itemsets. Maximal frequent itemsets are necessary for generating association rules.

[¶] the term scalability refers the ability to handle massive data stores

4. 3. Association Mining Algorithms

4.3.1 Apriori Algorithm

Apriori is an influential algorithm for mining frequent itemsets for Boolean association rules. The name of the algorithm is based on the fact that the algorithm uses prior knowledge of frequent itemsets properties. Han & Kamber (2001) pointed that Apriori employs an iterative approach known as level-wise search, where k -itemsets are used to explore $(k+1)$ -itemsets. First, the set of frequent 1-itemsets is found. This set is denoted L_1 . L_1 is used to find L_2 , the set of frequent two itemsets, which is used to find L_3 , and so on, until no more frequent k -itemsets can be found. The finding of each L_k requires one full scan of the database (Han & Kamber 2001). The pseudo code for the apriori algorithm is shown on the AppendixD.

According to Sivanandam et.al.(2004), Apriori is based on *bottom-up search approach*. In this approach, the search starts from 1 itemsets and extends one level in each pass until all maximal frequent itemsets are found. This approach performs well if the length of the maximal itemset is short. If the maximal itemset is longer, top down search is suitable. The authors further discussed that for a transaction with a medium sized maximal frequent set, a combination of both these approaches performs well. This algorithm integrates both the bottom up search as well as the top-down search. This algorithm is suitable for itemsets of any size. It uses the top down approach to find the frequent subsets of itemsets. The bottom up approach is used to find the supersets of the frequent itemsets.

Other than Apriori algorithm, there are also a number of association mining algorithms which are more or less rooted their bases on the basic idea of the apriori algorithm. Some of these which are adopted as a solution for limitations of apriori algorithm are as listed below.

4.3.2. Frequent Pattern Tree

Frequent Pattern growth is another way of mining frequent itemsets without candidate generation. The Apriori generates a huge number of candidates which, obviously, increase the time and space complexity of the process. The FP growth adopts the a divide –and-conquer strategy as follows: It compresses the database representing frequent items into a frequent pattern tree, but retain the items set association information, and then divide such a compressed database into a set of conditional databases (a special kind of projected database), each associated with one frequent item, and mine each such database separately.

The FP-growth method transforms the problem of finding long frequent patterns to looking for shorter ones recursively and then concatenating the suffix. Appendix B shows the algorithm. When the database is large, it is some times unrealistic to construct a main memory-based FP-tree. So an interesting alternative is to first partition the database into a set of projected databases, and then construct an FP-tree and mine it in each projected database. Such a process can be recursively applied to any projected database if its FP-tree still cannot fit in main memory.

According to Han & Kamber (2001), a study on the performance of the FP-growth showed that it is efficient and scalable for mining both long and short frequent patterns, and is about an order of magnitude faster than the apriori algorithm.

In the case of this research, the algorithm used for association mining was just the APRIORI algorithm and other algorithms like Count Distribution algorithm, Data Distribution algorithm, and Candidate Distribution algorithm are not employed. These advanced algorithms are left as recommendation for any interested in the area.

CHAPTER FIVE

5. EXPERIMENTATION

This chapter details the different data mining steps that were carried out in this experimental research work. This section makes up the central and most important part of the entire study. As to this research, the initial task is to identify the leading risk factors. Accordingly, the main objective of this experimentation part was to determine the main risk factors that determine probability of HIV infection to be high in the first place and find out their association rule or degree of co-occurrence.

The major data mining steps undertaken/involved in this research include: identifying the goal of the data mining task, identifying source of data, data cleaning and preprocessing, as well as developing classifier and association rules generation.

5.1. The Goal of This Data Mining Task

The underlying research problem that necessitated this study is the existence of high HIV/AIDS prevalence. Since, data on HIV-related cases is gathered daily by many health centers, it has become practically impossible to use only the traditional statistical data analysis on huge volume of data. Health professionals, planners and policy makers need to identify major determinants of HIV infection as well as their association in order to plan and implement case-specific intervention strategies to reduce HIV/AIDS prevalence and related disasters in country at large.

As stated in the background of this research, previously done studies on health were conducted by using a very small proportion of the data bases. Besides, in those studies, data analysis was conducted by using simple statistical techniques, such as regression and verification techniques. Since the analysis made by using traditional methods focuses on problems with much more manageable number of variables and cases than may be encountered in real world data bases, they have limited capacity to discover new and unanticipated patterns and relationships that are hidden in the databases. Particularly, in order to identify patterns in HIV-infection and associated risk factors, it

was difficult to conclusively attribute the problem to a certain set of factors since the parameters (attributes or factors) involved are too many. *So studying the affinity of the factors rather than simply listing them is of crucial importance.*

Before passing to the discussion on the remaining mining steps applied in this research an attempt was first done to describe the technology or tools used together with the steps involved in conducting the experiment. In this research work, the researcher has conducted two activities. To identify or determine the frontier attributes, just by the data itself, the researcher has used *KnowledgeSTUDIO* software and to mine the associations of these risk factors *WEKA* was used. Thus, before going to the details of the remaining steps which were carried out in this research, the researcher would like to give an overview of *KnowledgeSTUDIO* and *WEKA* software.

5. 2. Data mining Tool selection

Tool selection criteria may vary based on different considerations. Among the factors considered in the selection process for an appropriate data mining tool, the important ones were:

- ✚ The data mining tasks that the tool is intended for (classification and association)
- ✚ The algorithms supported (Apriori and Decision trees)
- ✚ Architecture and Operating systems: the computer architecture on which the software runs (stand alone and a MS Windows operating system)
- ✚ Data sources: Possible formats for the data that is to be analyzed (MS Excel)
- ✚ Data Size: the maximum number of records the software can comfortably handle (more than 15,000 records)
- ✚ Visualization capabilities

The identification of an appropriate data mining tool was a time taking process and consumed a considerable amount of the research time. Reviews of data mining tools, by Elder et. al. (1998) and Goebel et. al. (1999), were used as basis for further investigation on the Internet. The researcher finally approached the venders of two software tools that more or less satisfy the above requirements. These preferably

selected software are Knowledge Studio version 4.1.1 of Angoss Software Corporation (Angoss), and WEKA version 3-4.

5.3. KnowledgeSTUDIO Software

KnowledgeSTUDIO is a new generation of software that employs today's most advanced data mining techniques packaged in a familiar Microsoft Office-style working environment. KnowledgeSTUDIO functions as a workbench for creating and managing data mining projects. With it, it is easy to gather data, perform analyses, create and test models, graph objects, generate a variety of reports, and perhaps most importantly, score datasets so that you can make predictions based on your models.

KnowledgeSTUDIO provides a number of ways to visually explore and express patterns in data. It can help with the whole data mining process – from preparing data through producing final graphs and reports. Some of the facilities it provides for data mining processes are:

- ✦ **Decision Trees:** Explore relationships visually by creating splits that are displayed in a tree-like fashion.
- ✦ **Graphs:** Display the values of variables and decision tree splits with a number of chart types.
- ✦ **Predictive Models:** Create, test, and use software objects that can render predictions on new data.
- ✦ **Data Management:** View data in raw or mapped form. Control and adjust variable types and other meta-data. Handle missing values and more.
- ✦ **Cluster Analyses:** Create, test, and use software objects that can categorize or group records from new data.
- ✦ **Rules and Code:** Generate decision tree rules in a variety of formats. Generate code from predictive models.
- ✦ **Reports:** View and print reports for dataset layouts, decision tree splits, and predictive model and cluster analysis data.

5. 4. WEKA Tool/Software

The system is written in Java, an object-oriented programming language that is widely available for all major computer platforms. 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. There are several different levels at which WEKA can be used. First of all, it provides implementations of state-of-the-art learning algorithms that can be applied to a dataset from the command line. It also includes a variety of tools for transforming datasets, like the algorithms for discretization. One can pre-process a dataset, feed it into a learning scheme, and analyze the resulting classifier and its performance—all without writing any program code at all.

As an example to get started, better to explain how to transform a spreadsheet into a dataset with the right format for this process, and how to build a decision tree from it. Spreadsheet data can be transformed into a dataset as comma delimited data. And the file need be in a ARFF format. WEKA includes implementations of algorithms for learning association rules and for clustering data for which no class value is specified. One of the algorithm included for association mining is Apriori algorithm. The pseudo-code for this algorithm is shown on AppendixD.

Apriori algorithm iteratively reduces the minimum support until it finds the required number of rules with the given minimum confidence or until it reaches lower bound support. The algorithm provides various options to adjust and control the rule generation; some these valid options are listed as follows.

- N required number of rules. The required number of rules (default: 10).
- C minimum confidence of a rule. The minimum confidence of a rule (default: 0.9).
- D delta for minimum support. The delta by which the minimum support is decreased in each iteration (default: 0.05).

- U upper bound for minimum support. The upper bound for minimum support (default = 1.0). Don't explicitly look for rules with more than this level of support.
- M lower bound for minimum support. The lower bound for the minimum support (default = 0.1).
- S significance level. If used, rules are tested for significance at the given level. Slower (default = no significance testing).
- R If set then columns that contain all missing values are removed from the data.

These valid options were effectively used in this research to get rules at different confidence and support. Based on the level of support and confidence, the rules soundness was assessed by domain experts. Discussions for the result were made only on those rules which sound good as judged by experienced experts.

5. 5. Identifying the Source data

It has been stated under section 1.2 as well as 1.4 of the first chapter that the primary goal of this research work was to assess how well an association rules for the risk factors can be generated by using data mining techniques. This work would also help in predicting the risk of infection among clients based solely upon VCT client datasets consisting of demographic, partner, professional, educational, public health history factors, without using diagnostic tests or physical exam findings. Thus the result of this data mining process can be used to distinguish persons at high risk of infection from those with low risk of infection, and the association of the risk factors is investigated using the selected association mining tool (Apriori).

Even though clear definition of the problem is the first step in any data mining task, with high importance, the primary requirement is a availability of data in any format. And in most cases the ideal source of such data is the corporate data warehouse[▼]. Especially, in supervised learning system, we use pre-classified historical data (past data) to build a

[▼] Data warehouse refers to the collection of data from many different sources and its storage in a common format with consistent definition for keys and fields.

model of the future. Thus, in relation to this research work, in order to distinguish people who are likely to have high risk of infection from those with low risk of infection, the classifier should be trained using historical data collected from all VCT clients (consisting of both infected and uninfected people) or positive and negative test results.

In this research, in order to identify source of historical data collected about HIV/AIDS related cases by CDC, discussions were held with various experts at the domain area. As a result of this discussion, the researcher identified VCT to be main source of data which is available in electronic format that can be used for this research project. This choice has two reasons as a base. The first reason is for the fact that VCT services address the general population and hence data from such area has reliability to represent the entire population of the society. Secondly, the availability of the data in electronic format has made the VCT data to be appropriate source to conduct this experimental research.

5. 6. Creating the Target Datasets

Once the source of the data has been identified, the next step taken was creating target datasets from the obtained data. The VCT client data set was originally stored in Epi Info Version 6 (EPI6) with text format (which is a statistical package usually used for the analysis of medical data; (a brief of IPI6 will be detailed in the following indented paragraphs). Data entry into the specified software is performed at the VCT service sites and incorporated into the CDC data warehouse. Figure 5.1. shows the main screen of Epi Info software.

Epi Info, a series of computer programs produced by CDC and the World Health Organization provides public-domain software for word processing, database and statistics work in public health. Version 6 of Epi Info was released in May, 1994.

Epi Info is a complete system for word processing, data entry, database management, and public health statistics for IBM-compatible microcomputers (DOS). Facilities are provided for importing and exporting a variety of file types. Although Epi Info has its own file format, dBASE files also can be analyzed directly.

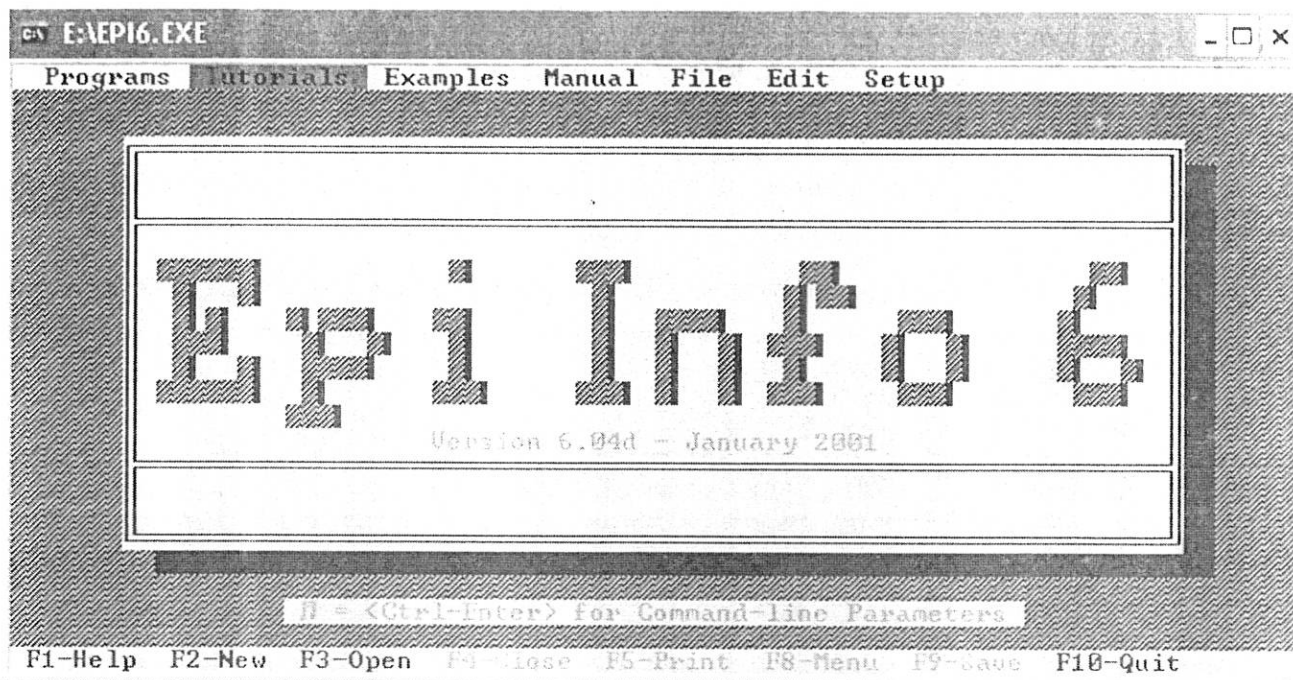


Figure 5.1 Main Screen of Epi Info Software

The data set was organized in rows and columns where each column represents an attribute (Variable) and each row stands for a single record of an individual. In this data set a total of 82 attributes (columns) and 18646 records (rows) were identified. This data is recorded in two years and nine months time from which the researcher decided to work the experiment using only two years data (09/12/2002 - 09/12/2004) having 15801 records.

The researcher collected a copy of this original data set on CD and all the necessary operations of data selection and data preparation were carried out on this data set. Following the successful selection of the required data set from the original two years data (09/12/2002 - 09/12/2004), the next important issue considered by the researcher was importing the selected data set, which was in Ep16 document format in to MS-Excel format.

Although the original data set was organized in rows and columns using EP16 statistical software, it was not possible to import this data as is into the data mining software that were used in this research project. Since the KnowledgeSTUDIO software has a good

facility to import data from Lotus, Excel, dBASE, Metastock, ASCII and binary files, the researcher decided to export the original data set in to MS-Excel format, which was then imported to KnowledgeSTUDIO software. As a result of this attempt, the selected data set was successfully imported in to Excel format. This is also for good reason of ease preprocessing operation with Ms-Excel. The exported data to a worksheet is shown as in AppendixA.

In order to determine the major risk factors that make an individual to lie in the HIV infected class, a data set consisting of both classes of clients (infected, and uninfected) was created and prepared for analysis. Since this research work was a preliminary work intended to assess the applicability of data mining technology based on VCT data sets, the experiment was conducted by samples taken from the original two years data. Thus for the purpose of this research work, a sample data set consisting of records from each class of people was selected and used to identify leading factors. With same sample data, association rules for the selected attributes were studied. Sample records and attributes of the original data is partially displayed on AppendixA.

5. 7. Feature Selection

As it is described previously, the original data set from which the target data set for this research work has been selected consists of a total of 82 attributes. However, based on experts' opinion, of these attributes, only some of them were considered as relevant (influential) for this specific research work.

Liu and Motoda (1998), as quoted in Shegaw (2002) discussed on the importance of selecting relevant features (attributes) in any data mining task as "the abundance of potential features constitute a serious obstacle to the efficiency of most learning algorithms. Popular methods such as K-nearest neighbor, Apriori, C4.5, and back propagation are slowed down by the presence of many features, especially if most of these features are redundant and irrelevant to the learning task." Furthermore, some algorithms may be confused by irrelevant or noisy attributes and lead to a poor mining result. Therefore, eliminating some attributes, which are assumed to be irrelevant to this

association mining, can increase the accuracy, save the computational time, and simplify results obtained.

It was by considering the importance of selecting relevant features (attributes) in increasing the efficiency of the algorithm that the researcher opted to apply dimensionality reduction on the data set created for analysis by selecting the minimum set of features (attributes) that associated with the research problem.

As a result of dimensionality reduction process, a total of 20 candidate attributes including client ID (as suggested by domain experts) were selected as relevant features to train the required decision tree classifier in order to identify the major factors for farther selection that was used for association mining. Selection of these first relevant features was conducted in consultation with health specialists and researchers who have good knowledge and experience on the data set of VCT database.

The following table gives a brief description of the candidate attributes selected as relevant features but further dimensionality reduction was carried out for the final mining task, which is association mining, based on the result obtained from the KnowledgeSTUDIO classifier.

Table 5.1. Attributes Selected from the original data file

No.	Attribute Name	Description	Values
1	CLIENT	A unique reference number given to each client	A serial number
2	AGE	The age of the clients	Continuous numerical value
3	SEX	The gender of the client	Two symbolic values
4	MARSTAT	Marital status of the client	Seven unique categorical values
5	EDUEXPE	The level of educational of the client	Eight unique categorical values
6	EMPLOYED	Employment condition of the client	Two unique categorical values
7	OCCUPAT	Occupation of the client	Twelve unique categorical values

8	REFERBY	A referral case visit	Twelve unique categorical values
9	REASHERE	The most important reason of the client on the visit day	Twenty-two unique categorical values
10	PRTESTED	The previously tested client case	Six unique categorical values
11	EVERHAD	The previous sex experience of the client with penetration	Two unique categorical values
12	EXPOTIME	Suspected exposure time	Continuous numerical values
13	USECOND	The client condom use in the last 3 months	Four unique categorical values
14	LASTSEX	The condom use of the client during the last time had sex	Four unique categorical values
15	STIHIST	The history of STI of the client	Four unique categorical values
16	SEXWORK	Is the client a commercial sex worker	Five unique categorical values
17	EXPECT	A test result that the client expects	Four unique categorical values
18	NOTIFPRE	The pre-test partner notification plan of the client	Five unique categorical values
19	HIV	The HIV test result of the client	Three unique categorical values
20	NOTIFPOS	The post-test partner notification plan of the client	Five unique categorical values

The above selected fields were chosen with various purposes in mind. Some of those features were selected to be used as an independent (input or factors) and as dependent. The dependent (output) variable i.e. status of the client was used to classify the visitors as infected or uninfected. Based on the outcome of the dependent variable, the contribution (influence) of predicator (input) variables for HIV infection were evaluated.

There are fields that were available in the original two years VCT data set but that were not considered as irrelevant for the task of this research project. These unused attributes include: "COUNTRY", "REGION", "VDATE", "COUNCODE", "PARTCODE", "COUPCODE", etc.

Since the data source selected for this research project consisted of two years and nine months VCT data (13/03/1002- 09/12/2004), records collected before March 2002 were not considered in this research project just for the sake of conducting the study on the two years data which is still more than 85 percent of the entire records and reasonable amount for a research. As it is already discussed previously, VCT records of clients who visited (were served) between December 2002 and December 2004 were used to train the model, which not only can predict the likelihood that a given client would be infected or not, but also for further selection of the most few attributes before association mining was employed.

5.8 Data Preprocessing

As it is described in the data mining methodology of Berry and Linoff (1997), data cleaning and preparation is the second step of any data mining task. Though, the process of data cleaning and preprocessing is highly dependent on the technique (method) to be employed, it can not be a step to be skipped in any data mining research.

So, data preparation was conducted by taking into consideration the requirement of the software used. Therefore, in preparing the data set into a format that is convenient for

knowledgeSTUDIO and WEKA, the following data cleaning and preprocessing steps were undertaken on the data set created for analysis.

5.8.1 Data transformation and reformation

California Scientific Software (1998) describe that, in general, how one represents data is often crucial to the success or failure of data mining project as well as the efficiency of the process. There are some common areas of confusion that should be avoided during the data preparation stage of data mining task. Particularly, when some data appears as numbers (as the case of data in this research in which almost all the codes are numeric), they may deceive us. So, according to California Scientific Software, if numbers represent unique concepts and not values within a continuous range of some quantity or rating, they should be converted into symbols.

In the original data set created for this data mining task, the values of some attributes were represented using numerical codes. Such numerical codes that are used to represent unique concepts had to be converted into symbols or columns to avoid any confusion during the mining process as follows:

- (i) **Decoding the "AGE" Code** The attribute "AGE" in the original data set consists the real age of the individuals. The values of this field were numeric and needed to be discretized into various ranges. Before segmenting the age range the researcher consulted domain experts to know on what basis they used to generate age related reports. So according to the experts' information, it was identified that they make data analysis as the following segments: "10 -To-14", "15 -To-19", "20 -To-24", "25 -To-29", "30 -To-34", "35 -To-39", "40 -To-44", "45 -To-49", and "50+". The transformation was made possible using MS-Excel functions.
- (ii) **Decoding the "SEX" Code:** In the original data set, the attribute "SEX" which stands for gender was represented using two different numeric codes(1 & 2). So, for the purpose of this research work, these numeric codes were converted into their respective symbolic values (the gender of the client). Table 5.2 shows the SEX code used in the original data file and the transformed name of the SEX.

Table 5.2 Actual Values of "SEX" code and the original numeric code

Original "SEX" Code	Transformed Value Used
1	Male
2	Female

(iii) **Decoding the "MARSTAT" code:** in the original data set the variable "MARSTAT" stands for "Marital Status" and its values were represented using seven different numeric codes one for each value. For this attribute, the original numeric codes were converted into the respective symbolic values (the marriage condition of the individuals). The following table shows the numeric code used in the original data set and the actual symbolic value used for this research work.

Table 5.3. Actual Values of each "MARTSTAT" and the original numeric code

Original "MARSTAT" Code	Transformed Value Used
1	Married
2	Never Married
3	Separated
4	Divorced
5	Widowed
6	Living together
99	Others

(iv) **Decoding the "EDUEXPE" code:** in the original data set, the variable "EDUEXPE" stands for "Educational Experience of the client"(i.e. the level of education of the visitor) and its values were represented using eight different numeric codes where each code represents one unique symbolic value. So, for this attribute, the numeric codes used in the original data set were converted into their respective symbolic values. The following table shows the original codes used to represent the values of "EDUEXPE" attribute, and the transformed value for each code.

Original "EDUEXPE" Code	Transformed Value Used
0	Illiterate
1	Able to read
2	Primary Level (1-6)
3	Secondary Level (7-12) Old System
4	Secondary Level (7-10) New System
5	A level (11-12)
6	Vocational (11-12)
7	Tertiary Level ≥ 12
99	Others

Table 5.4 The Original "EDUEXPE" code the decoded symbolic value

(v) **Decoding the "EMPLOYED" code:** in the original data set file this variable stands for the employment condition of the client and its values were represented by using two different numeric codes where each code represents one unique categorical value. These numeric codes were therefore converts into their actual values at this stage of

the research. The following table shows the original numeric codes used to represent the values of EMPLOYED attribute, and the decoded actual value for each code.

Table 5.5. The original code and Transformed values of "EMPLOYED" attribute

Original "EMPLOYED" Code	Transformed Value Used
0	No (Unemployed)
1	Yes (Employed)

(vi) **Decoding the "OCCUPAT" code:** this attribute is used to describe the occupation of the client and in the original data set file its values were represented by using twelve different numeric codes where each code represents one unique categorical value. The following table shows the original code and transformed values of each code.

Original "OCCUPAT" Code	Transformed Value Used
1	Legislators, Sr. Officials, Managers
2	Professionals
3	Technicians, Ass. Professionals
4	Clerks
5	Service, Shop, market, Sales
6	Skilled Agr. And Fishery workers
7	Crafts and Trades
8	Plant/Machine Op., Assembly
9	Elementary Occupation/unskilled
10	Students
11	Housewife
12	Armed forces
99	Others

Table 5.6. The Original code and transformed values of "OCCUPAT" attribute

(vii) **Decoding the "REFERBY" code:** in the original data set file this variable stands for the referral case or condition of the client (if any) and its values were represented by using twelve different numeric codes where each code represents one unique categorical value. These numeric codes were therefore converts into their actual values at this stage of this research work. The following table shows the original

numeric codes used to represent the values of "REFERBY" attribute, and the decoded actual value for each code.

Original "REFERBY" Code	Transformed Value Used
1	Self- not referred
2	Public Health Institutions
3	Private Health Institutions
4	Military Health Institutions
5	Friend or Relative
7	NGO
8	School
9	Religious Institutions
10	Client
11	APLWH/A
12	PMTCT
99	Others

Table 5. 7. The original data set code and transformed value of "REFERBY" attribute

(viii) **Decoding the "REASHERE" code:** in the original data set file this variable stands for the "Most Important Reason Here" which describes the reason why the client has visited the VCT service center and its values were represented by using twenty-two different numeric codes where each code represents one unique categorical value. These numeric codes were therefore converts into their actual. Table 5.8 shows the original numeric codes used to represent the values of "REASHERE" attribute, and the decoded actual value for each code.

Original "RASHERE" Code	Transformed Value Used
1	Client risky/had risk
2	Partner risky/ Had risk
3	not trust partner
4	Ill/ symptoms
5	Premarital
6	Marital reunion
7	Family planning
8	Visa Applicant
9	Referred
10	2 nd test (window)
11	Confirm positive result
12	Get results preventive test
13	Need counselling
14	Test before pregnant
15	Pregnant, must know
16	Plan for future
17	Death/illness of partner
18	Occupational exposure
19	Other blood/fluid exp.
20	Sexual assault
21	Preliminary ART
99	Others

Table 5.8 The original data set code and transformed value of "REASHERE" attribute

(ix) **Decoding the "PRTESTED" code:** In the original data set file this variable stands for the "Previously Tested" which describes the previous test condition of the client and its values were represented by using seven different numeric codes where each code represents one unique categorical value. These numeric codes were therefore converted into their actual value. The following table shows the original numeric codes used to represent the values of "PRTESTED" attribute, and the decoded actual value for each code.

Table 5.9 The original data set code and transformed value of "PRTESTED" attribute

Original " PRTESTED" Code	Transformed Value Used
0	No
1	Yes, HIV +
2	Yes, HIV-
3	Yes, inconclusive
4	Result not given
5	Didn't take results
99	Others

(x) **Decoding the "EVERHAD" code:** in the original data set this variable is used to store information about sexual experience of individual ("EVER HAD SEX"). In the original file, the values of this variable were represented using two numeric codes. So, for the purpose of this research work, the numeric codes in the target data set created for analysis were converted into their respective actual values as shown in Table. 5.10

Table 5. 10 The Original "EVERHAD" code and the new decoded values

Original " EVERHAD" Code	Transformed Value Used
0	No
1	Yes

(xi) **Decoding "EXPOTIME" code:** this variable stands for "Suspected Exposure Time" and used to maintain information about the suspected exposure of infection risk. In the original data file, the values for this attribute were represented using five numeric codes. Therefore, for this variable as well, these numeric codes were converted into their respective actual values. The following table shows the original codes and the respective values for those codes.

Original " EXPOTIME" Code	Transformed Value Used
1	Less than one month
2	Between 1 month and 3 months
3	Between 4 month and 6 months
5	Over six months
98	N/A

Table 5. 11 The Original "EXPOTIME" code and the new decoded values

(xii) **Decoding "USECOND" code:** this variable stands for "Condom use for the last three months" and in the original data file, the values for this attribute were represented using four numeric codes. Therefore, for this variable these numeric codes were converted into their respective actual values. The following table shows the original codes and the respective values for those codes.

Table 5.12 The Original "USECOND" code and the new decoded values

Original " USECOND" Code	Transformed Value Used
0	Never
1	Always
2	Sometimes
98	N/A

(xiii) **Decoding the "LASTSEX" code:** this variable stands for "Used condom last time had sex" and used to gather information about if the client used condom during his/her last time sex engagement. For this variable four numeric codes were used in the original data set. The actual values required for this variable were decoded as in the following table:

Table 5.13 The Original "LASTSEX" code and the new decoded values

Original " LASTSEX" Code	Transformed Value Used
0	No
1	Yes
97	Doesn't remember
98	N/A

(xiv) **Decoding the "STIHIST" code:** this attribute stands for "History of STI" and used to draw information about the sexually transmitted infection history of the client. In the original data set, the values of this variable were represented using four numeric codes. So, for the purpose of this research work, the numeric codes in the target data set were converted into their respective actual values.

Table 5.14. The Original "STIHIST" code and the new decoded values

Original " STIHIST" Code	Transformed Value Used
0	No
1	Yes
97	Don't know
98	N/A

(xv) **Decoding "SEXWORK" code:** SEXWORK stands for "Is client commercial sex worker" and is used to store information about the client if he/she is committed to sexual activities for business purpose. In the original data set file, the values of this variable were represented using five numeric codes. The original code and their respective Transformed values are shown in the table below:

Table 5.15 The Original "SEXWORK" code and the new decoded values

Original " SEXWORK" Code	Transformed Value Used
0	No
1	Yes
2	counselor thinks yes
97	Don't know
98	N/A

(xvi) **Decoding the "EXPECT" code:** the "EXPECT" attribute stands for the "Result Client Expects" and used to collect the opinion of the client if he/she thinks the result of test will be positive or negative. The values for this variable were represented using four numeric codes. The following table depicts the original and transformed values of the attribute.

Table 5.16 The Original "EXPECT" code and the new decoded values

Original " EXPECT" Code	Transformed Value Used
0	Negative
1	Positive
97	Don't know
98	N/A

(xvii) **Decoding the "NOTIFPRE" code:** this variable is used to maintain information about the willingness of the client to notify his/her partner prior to the HIV test. "NOTIFPRE" stands for pre-test partner notification plan and in the original data set file the values of this variable were represented using five numeric codes. The following table shows the original code the transformed values of the variable.

Table 5.17 The Original "NOTIFPRE" code and the new decoded values

Original " NOTIFPRE" Code	Transformed Value Used
0	Refused to notify
1	Agree to notify
2	Plan to notify
3	Unsure
98	N/A

(xviii) **Decoding the "HIV" code:** the HIV attribute is used to record information of the test result of the client. In the original data set the values of this attribute were given in three numeric codes and their decoded values were shown on the following table.

Table 5.18 The Original "HIV" code and the new decoded values

Original " HIV" Code	Transformed Value Used
0	HIV-
1	HIV+
98	N/A

(xix) **Decoding the "NOTIFPOS" code:** the "NOTIFPOS" attribute stands for "post-test partner notification plan" which is used to store information of the willingness of the client to notify his friend about his/her HIV test result. In the original data set the values of this variable were filled with five numeric codes. The next table is about the original code and the transformed values of this variable.

Table 5.19 The Original "NOTIFPOS" code and the new decoded values

Original " NOTIFPOS" Code	Transformed Value Used
0	Refused to notify
1	Agree to notify
2	Plan to notify
3	Unsure

These decoded or transformed values for all attributes were used for this experimental research and their transformation was done by using Excel built-in functions. The formula employed to reform all the attribute values is shown on *AppendixB*.

5.8.2 Handling Variables with Missing Values.

Missing values of some variables, in any data mining task, affects the mining output and should be handled systematically. In order to handle the problem of variables with missing values in the data set created of a given data mining task, there are different approaches recommended for use. For this specific research work, the approach suggested by Two Crows corporation (1998) was followed to represent missing values identified in the target data set.

For continuous variables or variables with numerical values, such as the AGE variable in this project, Two Crows Corporation (1998) suggest that missing should be replaced with the mean value for that field. Based on this suggestion, for the numeric variable (AGE) missing values were replaced by the mean value of this field. The average value for this variable was 26.

In similar way, to handle the problem of missing values for categorical variables, the same suggestion is to group such fields into nominal and ordinal variables, and to take the median for ordinal variables and the modal for the nominal variables. The ordinal variables are categorical variables whose values have meaningful order and such a variable was not involved in this research, where as nominal variables refer to the categorical variables whose values are unordered.

According to the data set created for this research work, categorical variables whose missing values were handled based on the above suggestion include the following variables: "SEX", "MARSTAT", "EDUEXPE", "EMPLOYED", "OCCUPAT", "REFERBY", "REASHERE", "PRTESTED", "EVERHAD", "EXPOTIME", "USECOND", "LASTSEX", "STIHIST", "SEXWORK", "EXPECT", "NOTIFPRE", "HIV", "NOTIFPOS". All the above fields are nominal variables and for any missing in those fields, the modal (most

frequent) value was filled. The modal values for the above these nominal categorical variables listed in the table below.

Table 5.20 Modal Values to Represent Missing Values for Nominal variables

Field name	Modal value
SEX	Male
MARSTAT	Never Married
EDUEXPE	Secondary Level
EMPLOYED	No
OCCUPAT	Others
REFERBY	Self-not referred
REASHERE	Had Risk or Suspect or Symptoms
PRTESTED	No
EVERHAD	Yes
EXPOTIME	Over Six months
USECOND	Never
LASTSEX	No
STIHIST	No
SEXWORK	No
EXPECT	Negative
NOTIFPRE	Agree to Notify
HIV	Negative
NOTIFPOS	Agree to Notify

5.8.3 Data Organization for Model Building

The previous sections were a preprocess activities and still there are more step towards the model building. Data preparation for the KnowledgeSTUDIO was not a big deal except the reformation as shown in the previous section. Because this software can process data in plain English even if there are punctuation marks included. Specially for interpreting in the Decision tree algorithm, such data format was not considered problematic.

But in the case of data preparation for the WEKA, another considerable amount of time was exhausted. The data which were in Excel format was not directly used unlike that of for KnowledgeSTUDIO. The values for each attribute were made to single label (without having spaces in between if any). When the values are made of many words, they were written in a jointed way to be read as single label as Comma Delimited or comma separated values file/data. In order for better understandability to humans, the first letters of the joined words were capitalized. For instance, the value "Fifteen to Nineteen" of AGE was converted to "*FifteenToNineteen*". Then the data was saved in (*data.csv*) file format.

Then the file was opened with Notepad program for further preparation. Here the relation and attributes were declared for the data in such away that the WEKA software understands. The sample dataset for declaration of attributes is shown on *AppendixC*. After preparing the data this way it was saved in (*data.arff*) file format that Java environment of WEKA reads.

5. 9. Model Building

At this stage everything was ready to create and train the model. The first action to accomplish this was to import the organized data set into the workspace of the KnowledgeSTUDIO. The data insertion screen of the software looks Figure.5.2.

Basically, the work of this research was a two-step process. In the first step a classifier was modelled to identify the frontier risk factors, performed by decision tree classifier in KnowledgeSTUDIO. In the second step the association rules were generated for the most determinant risk factors. So the present research was more concentrated in generating rules to explain affinity group of the factors or to determine which factor most often go together in the VCT records with positive test result.

Thus, after the data set in Excel format was imported into KnowledgeSTUDIO, classification model was trained by inserting decision tree with KnowledgeSEEKER algorithm. The classification model was conducted by using all the 19 attributes that were selected except the "CLIENT" field because the values of the client attribute are unique ID for each record has no power to discriminate. The variables that were used to build the first model were: "SEX" , "MARSTAT", "EDUEXPE", "EMPLOYED", "OCCUPAT", "REFERBY", "REASHERE", "PRTESTED", "EVERHAD", "EXPOTIME", "USECOND", "LASTSEX", "STIHIST", "SEXWORK", "EXPECT", "NOTIFPRE", "NOTIFPOS" and the "HIV" field was the output variable.

For the first trial, all the default parameters provided by the KnowledgeSTUDIO (i.e., algorithm, Information Gain Measure, and Auto select settings) were accepted and used as they were. The default parameters of auto select setting used for this initial has the following vales.

Maximum attr. info	10.00
Minimum attr. info	0.1
Maximum attr. count	5
Algorithm	KnowledgeSEEKER
Info. Measure	Entropy Variance - Non P-value

In this software there are two types of algorithm that are available for construction of decision tree. These are KnowledgeSEEKER and HeatSEEKER. In this particular research KnowledgeSEEKER was used. The reason is that it is a flexible algorithm that is especially good for exploration purposes and manual tree building and it can also handle a large amount of attributes with either a continuous or discrete dependent variable.

With this setting, the model determined classes of the attribute HIV. The entire record was classified into HIV+ and HIV- values as shown on the next figure. There is a 17.06 percent chance that HIV will be positive and 82.94 percent chance that HIV will be Negative.

HIV+	635	17.06%
Negative	3,087	82.94%
Total	3,722	100.00%

The model were iteratively built and checked by varying the parameters from the default value to see whether the probability of the class changes from the obtained percentage. But after many trials it was found that the probability remained the same as the first result. Therefore, the researcher decided to continue using the default parameters and setting as provided by the software.

After modelling algorithm and modelling measure have been selected and before the model was actually built, a mechanism was generated to test the models quality and validity. In this step what was performed is the separation of the data set into training and test set so as to build the model. This is to build the model on the training set and estimate its quality on the testing set. Partitioning the dataset into learning and test sets was done using the insert partition feature of the software.

Before passing to the next model building, the classifier was checked for validity with validation dataset. Validation showed better performance when the tree is forced to grow automatically. The auto grow stop size was set up to 20 and number of split to up to 5. This decision for parameter setting was done by first trying on different combinations and this one was found to give better percentage of validation. The following table indicates the Validation Summary Report or confusion matrix for the validation set.

Confusion Matrix - HIV			
	Predicted		
Actual		HIV+	HIV-
	HIV+	92	178
	HIV-	13	1297

Statistics	
Total records	1580
Correctly predicted	1389
Percentage	87.91%
Valid records	1580

When the classifier is used on validation data which was not part of the training data, the test showed that 87.91 percent of the total records were classified correctly. Among the 1580 records the confusion matrix show that Only 13 records were predicted positive and observed negative while 178 were predicted negative and observed positive.

Next to this, upon an effort to identify the most determinant factors experimentally, different classifiers were made with various combinations of attributes by omitting one variable at a time. This was done in consultation with experts in the area who had better knowledge and experience as to which variables are more determinant (among the already selected ones) and essential to identify infection probability. To build this classifier too, the default parameters were used.

With this option used by the researcher every model was checked for validation to see the influence of the omitted variable. The exclusion of some of the variables resulted in a less performance classifier, which indicates the importance of the excluded variable. For some variables, their exclusion resulted in a better performance as seen with validation set, which means that these variables are not important. A summary of the precision rate observed for the different variables with their respective number of records correctly predicted are shown on the following table.

Table 5. 21 List of variables excluded and the performance of those models

Excluded Variable	Percentage	# Records correctly predicted
AGE	87.53*	1383
SEX	87.91	1389
MARSTAT	87.03*	1375
EDUEXPE	87.91	1389
EMPLOYED	87.91	1389
OCCUPAT	87.91	1389
REFERBY	87.91	1389
REASHERE	87.59*	1384
PRTESTED	88.23	1394
EVERHAD	87.97	1390
EXPOTIME	88.04	1391
USEDOND	87.59*	1384

LASTSEX	87.97	1390
STIHIST	87.59*	1384
SEXWORK	87.91	1389
EXPECT	87.22*	1378
NOTIFPRE	87.91	1389
NOTIFPOS	87.66*	1385

For the results presented in the above table, it was observed that the exclusion of the variables AGE", "MARSTAT", "REASHERE", "USEDOND", "STIHIST", "ECPECT", and "NOTIFPOS" flagged with star (*) resulted in a poor performing model, relative to the original model, which indicates the importance of these variables. The researcher used 0.25% deviation as threshold to justify the importance of these variables. Where as better performing models were obtained when the variables: "PRTESTED", "EVERHAD", "EXPOTIME", "LASTSEX", "NOTIFPRE" were excluded. And the exclusion of the variables "SEX", "EDUEXPE", "EMPLOYED", "OCCUPAT", "REFERBY", and "SEXWORK", showed no change at all and this still indicates that the variables are not as such important ones.

Therefore, for the further experimentation, the researcher decided to concentrate more on those accepted as important variables. In contrary to the opinion of the domain experts, the variable SEX and EMPLOYED remained to be non-influential. There was also another problem observed from the previous experiment. The attribute NOTIFPOS was judged as unimportant by the experts but in the above experiment the exclusion of this variable revealed poor accuracy as if it were an important variable. The scientific selection of the other attributes was appreciated with views of domain experts.

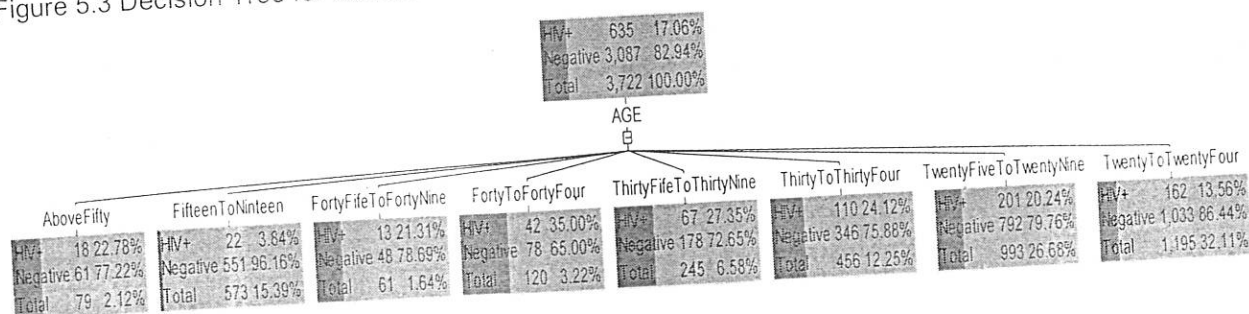
To get rid of such strange phenomena, the researcher has made subsequent changes based on StatSoft (2003) suggestion that it may often be most useful to combine the automatic methods for building trees with "educated guesses" and domain specific expertise. To be wise the researcher has built various models with these attributes (AGE", "MARSTAT", "REASHERE", "USEDOND", "STIHIST", "ECPECT") together with the suspicious fields ["NOTIFPOS", "SEX", and "EMPLOYED") in different combination

and checked for validity. This approach compromised both experts' suggestion and the experimentally obtained evidence.

Subsequent changes were made and different parameters utilized, to analyze whether SEX, EMPLOYED and NOTIFPOS are good determinant to be considered as influential risk factors or not. But upon checking for validity it was experimentally observed that the precision of all classifiers remained unchanged. That means, regardless of exclusion or inclusion of them didn't show difference than using only the previously selected variables. So with this experimental evidence and experts' suggestion, the researcher proceeded towards the remaining part of the research identifying attributes "AGE", "SEX", "MARSTAT", "EMPLOYED" "REASHERE", "USEDOND", "STIHIST", "EXPECT" as determinant factors for HIV infection.

The individual discussion on the contribution of each of these factors is presented as follows one after another.

Figure 5.3 Decision Tree for attribute AGE



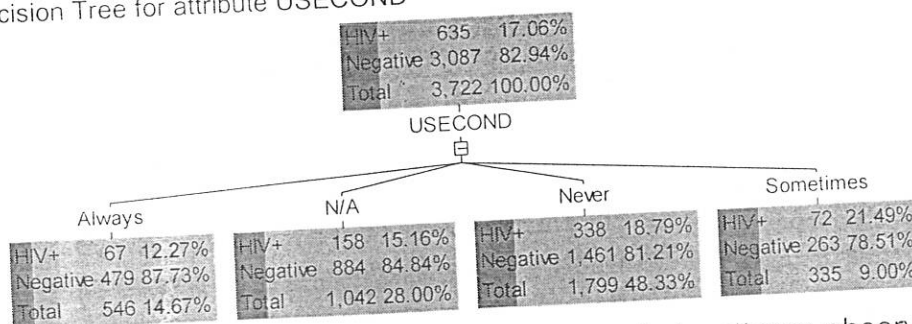
Taking the extreme case, the figure indicates that higher proportion of the sample (32.11%) is in the age of 20 to 24 among which 13.56% was found HIV positive. The maximum ratio of infection per class (35%) was seen in the range of 40 to 44 years age. and minimum infection ratio (3.84%) was observed in the range of 15 to 19 years.

If AGE is Above Fifty then there is a 22.7 percent chance that HIV will be positive and a 77.22 percent chance that HIV will be Negative. And if AGE is 45 to 49 then there is a 21.31 percent chance that HIV will be HIV+ and a 78.69 percent chance that HIV will be Negative.

When we look at the 20's and 30's, If AGE is 30-34 there is a 24.12 percent chance that HIV will be positive and a 75.88 percent chance that HIV will be Negative. If AGE is 35-39 there is a 27.35 percent chance that HIV will be positive and a 72.65 percent chance that HIV will be Negative. Similarly, if AGE is 25-29 then there is a 20.24 percent chance that HIV will be positive and a 79.76 percent chance that HIV will be Negative.

The above findings generally indicate that the higher probability of infection spans from age 25 and above. For all ages above 25 there is more than 20% chance to be HIV positive, regardless of other factors to be discussed later.

Figure 5. 4 Decision Tree for attribute USECOND

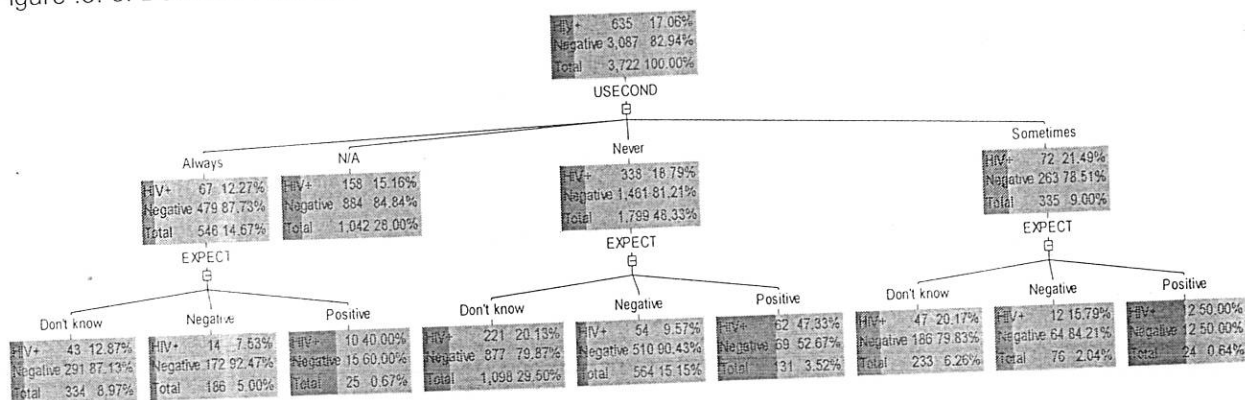


From this figure there are two interesting and unexpected patterns observed. The first one is about the records in the N/A(Not Applicable) class. The clients with this value reflected that there is no need to use condom, regardless of their reason. Yet they have 15.16% chances to be infected. *May be they are the one who do not exercise sex, or who think they trust their partner, or so. But for unknown reason the probability of infection is not negligible. So the researcher recommends another investigation on this issue.*

The other important finding here is comparison of the people who never use condom and those who use sometimes. Unlucky they are; the probability of infection for the people who use condom sometimes is higher than those who never use it. That means, If USECOND is equal to "Sometimes" then there is a 21.49% chance that HIV will be positive, where as If USECOND is equal to "Never" there is a 18.79% chance that HIV will be positive. *The difference is 3% chance which is very significant with this large scale data. This disclosed that the irregular use is even the worst of all the cases.*

To increase the insight view of these patterns, the researcher considered the fields USECOND and EXPECT together and came up with the following results:

Figure .5. 5. Decision Tree for attributes USECOND and EXPECT



From the above figure, it is shown that if USECOND is "Always" and EXPECT is "Negative" there is a 7.53% chance that HIV will be positive and this is very small when compared to those who expect positive (40%) even though they say that they use condom always. This an indication that people who properly use condom have better confidence to expect negative test result and also have better chance to be negative than those who expect positive.

When looking at the classes who sometimes use and who never uses condom, whatever they may expect (positive or negative), the higher probability is still with the one who sometimes use condom. The following rules generated from the tree describe their probability in clear sense.

- Rule # 1:** If USECOND is equal to Never and EXPECT is equal to Don't know then there is a 20.1275 percent chance that HIV will be HIV+ and a 79.8725 percent chance that HIV will be Negative.
- Rule # 2:** If USECOND is equal to Never and EXPECT is equal to Negative then there is a 9.57447 percent chance that HIV will be HIV+ and a 90.4255 percent chance that HIV will be Negative.
- Rule # 3:** If USECOND is equal to Never and EXPECT is equal to Positive then there is a 47.3282 percent chance that HIV will be HIV+ and a 52.6718 percent chance that HIV will be Negative.
- Rule # 4:** If USECOND is equal to Sometimes and EXPECT is equal to Don't know then there is a 20.1717 percent chance that HIV will be HIV+ and a 79.8283 percent chance that HIV will be Negative.

Rule # 5: If USECOND is equal to Sometimes and EXPECT is equal to Negative then there is a 15.7895 percent chance that HIV will be HIV+ and a 84.2105 percent chance that HIV will be Negative.

Rule # 6: If USECOND is equal to Sometimes and EXPECT is equal to Positive then there is a 50 percent chance that HIV will be HIV+ and a 50 percent chance that HIV will be Negative.

The other important point which was not been possible to display on the tree (larger tree size unable to fit paper size) was the following facts:

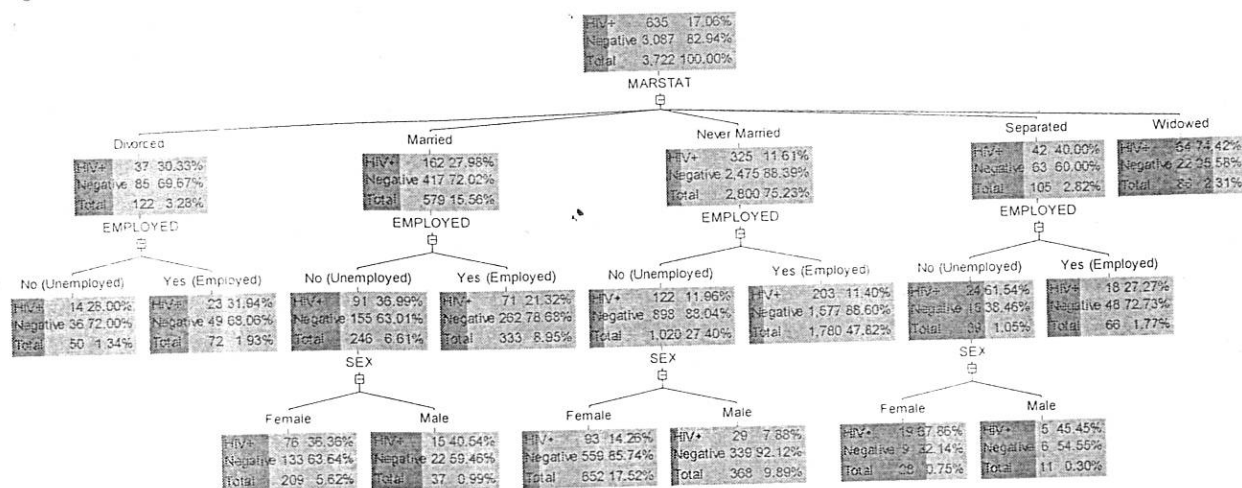
Rule # 7 If USECOND is equal to N/A and SEX is equal to Female then there is a 17.1681 percent chance that HIV will be HIV+ and a 82.8319 percent chance that HIV will be Negative.

Rule # 8: If USECOND is equal to N/A and SEX is equal to Male then there is a 12.7883 percent chance that HIV will be HIV+ and a 87.2117 percent chance that HIV will be Negative.

From this it can be generalized that most female compared to male feel/believe that the concern of condom is not applicable. 17.17% of females who say using condom is not applicable are infected. But in the case of male only 12.79% of those who say using condom is not applicable are infected.

The next graph displays the decision tree for the attribute MARSTAT. And associated set rules are discussed after.

Figure 5. 6 Decision Tree for attribute MARSTAT



From this decision tree, lots of previously unrecognized things related to chance of HIV infection were observed. These are made clear by looking at the rules obtained from the tree. This tree indicates that above 75% of the data is collected from unmarried people. The implication of this figure is that unmarried individuals are concerned about HIV case

(are more fearful, or have better awareness) than others and get benefits of VCT. But in contrary, the ration of people infected with HIV as seen from the result is very small in unmarried group. The probability is only 11.61% which is much smaller than the general probability (17.06%).

According to this result, married people are found to be less concerned about the matter (VCT), though the infection probability has shown very significant figure. The result shows 27.98% of married clients are infected which almost triples the percent for the unmarried one. This gap invites two other investigations: the first one is to know why married people are so affect with HIV and the second one is to know why this group could not be benefited from VCT services.

Widowed, separated, and divorced classes have exhibited the higher probability in general. That is 74.42%, 40% and 30.33% of them respectively are HIV positive. And hence the researcher analyzed the case of marital status together with other factors like employment and gender in order to discover new understanding. Regarding this, rules from same tree indicates the next important behaviour of the data.

- Rule # 1: If MARSTAT is equal to Married and EMPLOYED is equal to No (Unemployed) then there is a 36.99 percent chance that HIV will be HIV+ and a 63.01 percent chance that HIV will be Negative.*
- Rule # 2: If MARSTAT is equal to Married and EMPLOYED is equal to Yes (Employed) then there is a 21.3213 percent chance that HIV will be HIV+ and a 78.6787 percent chance that HIV will be Negative.*
- Rule # 3: If MARSTAT is equal to Married and EMPLOYED is equal to No (Unemployed) and SEX is equal to Female then there is a 36.36% chance that HIV will be HIV+ and a 63.64% chance that HIV will be Negative.*
- Rule # 4: If MARSTAT is equal to Married and EMPLOYED is equal to No (Unemployed) and SEX is equal to Male then there is a 40.54 % chance that HIV will be HIV+ and a 59.46% chance that HIV will be Negative.*
- Rule # 5: If MARSTAT is equal to Never Married and EMPLOYED is equal to No (Unemployed) then there is a 11.9608 percent chance that HIV will be HIV+ and a 88.0392 percent chance that HIV will be Negative.*

- Rule # 6: If MARSTAT is equal to Never Married and EMPLOYED is equal to Yes (Employed) then there is a 11.4045 percent chance that HIV will be HIV+ and a 88.5955 percent chance that HIV will be Negative.*
- Rule # 7: If MARSTAT is equal to Never Married and EMPLOYED is equal to No (Unemployed) and SEX is equal to Female then there is a 14.26% chance that HIV will be HIV+ and a 85.74% chance that HIV will be Negative.*
- Rule # 8: If MARSTAT is equal to Never Married and EMPLOYED is equal to No (Unemployed) and SEX is equal to Male then there is a 7.88% chance that HIV will be HIV+ and a 92.12% chance that HIV will be Negative.*
- Rule # 9: If MARSTAT is equal to Separated and EMPLOYED is equal to No (Unemployed) then there is a 61.54% chance that HIV will be HIV+ and a 38.46% chance that HIV will be Negative.*
- Rule # 10: If MARSTAT is equal to Separated and EMPLOYED is equal to Yes (Employed) then there is a 27.2727 percent chance that HIV will be HIV+ and a 72.7273 percent chance that HIV will be Negative.*
- Rule # 11: If MARSTAT is equal to Separated and EMPLOYED is equal to No (Unemployed) and SEX is equal to Female then there is a 67.86% chance that HIV will be HIV+ and a 32.14% chance that HIV will be Negative.*
- Rule # 12: If MARSTAT is equal to Separated and EMPLOYED is equal to No (Unemployed) and SEX is equal to Male then there is a 45.45% chance that HIV will be HIV+ and a 54.55% chance that HIV will be Negative.*

It is presented in rules 1 and 2, that employment variations between the married communities revealed marked differences in HIV infection. As per the discussions made with health experts and researchers, *the researcher proved that these variations in HIV infection among the employed and unemployed seems appropriate since unemployed people usually have excess free time to be exposed to the problem. This case proves that busy individuals are by far (about 15.67% probability difference) safer than others.*

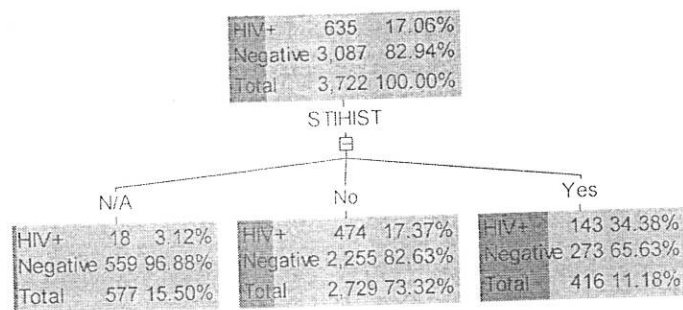
Rules 3 and 4 reinforces the above case relating to the gender difference. Taking the unemployed records of married people in to consideration, these rules disclosed that the male are more vulnerable than the female provided that they are married. The possible reason for this observation is left open for further investigation.

Another important understanding drawn from the tree is that appeared on rules 7 and 8. Rule number seven revealed that 14.26% of the unmarried and unemployed female

clients are HIV positive. This number is greater than the total probability (11.96% shown seen in rule # 5). But on the rule number 8 it is shown that only 7.88% of unmarried and unemployed males are HIV positive. So the conclusion of these two rules is that, in the case of this group (unmarried and unemployed) female are found to be more vulnerable than male. This proved the earlier understanding that women commit themselves to unsafe sexual practice for material benefit.

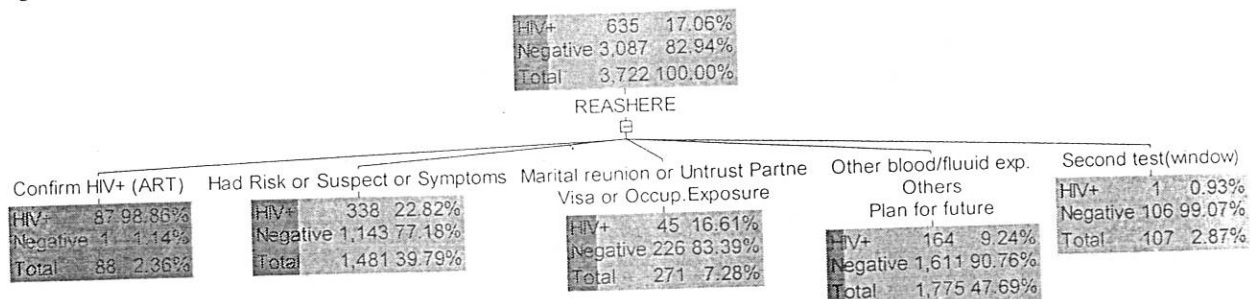
Next is discussion on relation of sexually transmitted infection and HIV infection. Figure 5.7 shows the decision tree for the attribute STIHIST.

Figure 5. 7 Decision Tree for attribute STIHIST



From the above tree, it was observed, that HIV/AIDS is associated with sexually transmitted infection. Especially, looking at the "No" and "Yes" node of the tree, 34.38% of the people with STI positive are under category of HIV positive. But on the "No" STIHIST case the probability of HIV infection is 17.37% which is by half less than the probability for the "Yes" group. The chance of the "No" STIHIST is slightly equal to the general probability of HIV positive (17.06%) observed for the entire data set. So from this understanding it is proved that having sexually transmitted infection increases the chance of HIV infections as well.

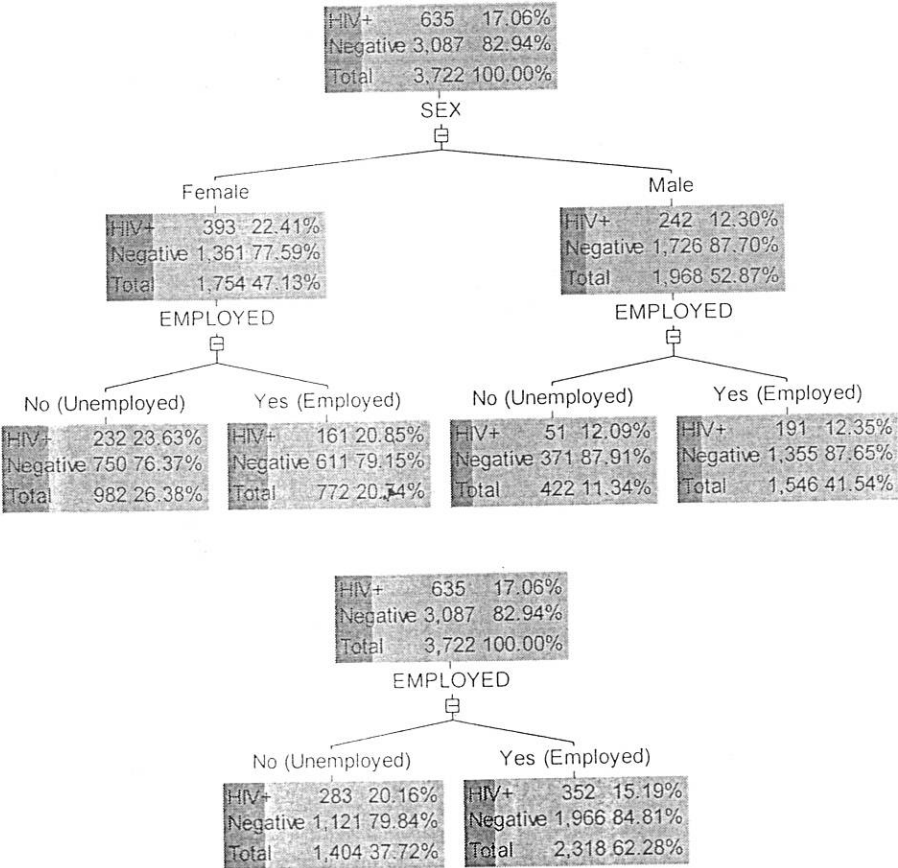
Figure 5. 8. Decision Tree for attribute REASHERE



From this decision tree, it is shown that only one person out of 88 clients who want to confirm HIV+ was seen negative. So 98.86% of peoples who previously known their seropositive status have confirmed exactly. *This was recognized as an indication for the perfection of the HIV test service.*

On the other hand, among the clients who visit for second test or windowing, only one person was found HIV positive. The 99.07% of the people with this reason are observed HIV negative. *This shows another insight that people who had used VCT service are much benefited and have strong stack to take care of themselves from ways that leads to HIV infection. This was made possible as a result of counseling service given before and is an encouraging assessment or feedback that counseling precisely helps society.*

Figure 5. 9. Decision Tree for attribute SEX and EMPLOYED



From this decision tree for SEX attribute, it was observed that higher percent of the clients are male (52.87%) while 47.13% is female. But higher proportion of infection was

seen on the female node with probability of 22.41% while in the case of male there is only 12.3% chance to be HIV positive. The probability seems balanced for both the unemployed and employed male group (12.09% and 12.35% respectively). But in the case of unemployed females, the probability of infection is increased to 23.63%. *This variation was observed as another testimony that there is material benefit behind multi partnership for women. Therefore, unemployment is one critical problem to be considered on the way towards HIV/AIDS intervention.*

5.10 Generating Association rules

In the previous section of this chapter the primary task done was identifying the main risk factors in a systematic way. That step was also supported with approval of domain experts. Now, in this section the association of these attributes was mined and their association rules were generated and analyzed. This is a section in which the second part of objective of this particular research has met its goal.

For this part of the project, the researcher used WEKA data mining tool for the only reason that it has better features for association mining with Apriori algorithm. But as it has been mentioned in the data preparation section of this chapter, preparing the data was a little bit tiresome task and accounted considerable amount of the research time. The data which were in Excel format was not directly used unlike that of for KnowledgeSTUDIO. The values for each attribute were made to single label (without having spaces in between if any). When the values are made of many words, they were written in a jointed way to be read as single label as Comma Delimited or comma separated values file/data. For the sake of better human understandability, the first letters of the joined words were capitalized. For instance, the value "Fifteen to Nineteen" of AGE was converted to "FifteenToNineteen". Then the data was saved in (data.csv) file format. Table 5. 22 sample data prepared as single word value.

Then the file was opened with Notepad program for further preparation. Here the relation and attributes were declared for the data in such away that the WEKA software

understands. The sample dataset for declaration of attributes is shown on AppendixC. After preparing the data this way, it was saved in (*data.arff*) file format that Java environment of WEKA reads.

AGE	SEX	MARSTAT	EMPLOYED	REASHERE	USECOND	STIHIST	EXPECT	HIV
TwentyToTwentyFour	Female	NeverMarried	Yes	HadRiskOrSuspectOrSymptoms	Never	No	DontKnow	Negative
TwentyToTwentyFour	Male	NeverMarried	Yes	PlanForFuture	Never	Yes	Negative	Negative
TwentyFiveToTwentyNine	Female	NeverMarried	No	HadRiskOrSuspectOrSymptoms	Always	No	DontKnow	Positive
TwentyToTwentyFour	Female	Divorced	No	VisaOrOccupExposure	Never	No	Negative	Negative
FifteenToNineteen	Female	NeverMarried	No	PlanForFuture	NotAppl	NotAppl	DontKnow	Negative
TwentyToTwentyFour	Male	NeverMarried	Yes	HadRiskOrSuspectOrSymptoms	NotAppl	No	Negative	Negative
FifteenToNineteen	Male	NeverMarried	Yes	PlanForFuture	NotAppl	NotAppl	Negative	Negative
TwentyToTwentyFour	Female	NeverMarried	Yes	PlanForFuture	NotAppl	No	DontKnow	Negative
TwentyFiveToTwentyNine	Male	NeverMarried	Yes	PlanForFuture	Always	No	Negative	Positive
ThirtyToThirtyFour	Female	Married	No	HadRiskOrSuspectOrSymptoms	Never	No	DontKnow	Negative
TwentyFiveToTwentyNine	Male	NeverMarried	No	HadRiskOrSuspectOrSymptoms	NotAppl	No	Positive	Positive
FifteenToNineteen	Female	NeverMarried	Yes	OtherBloodOrFluidExp	NotAppl	NotAppl	DontKnow	Negative
TwentyToTwentyFour	Male	NeverMarried	No	HadRiskOrSuspectOrSymptoms	Never	No	Negative	Negative
ThirtyToThirtyFour	Male	NeverMarried	Yes	PlanForFuture	NotAppl	No	DontKnow	Negative
ThirtyFiveToThirtyNine	Male	Married	Yes	HadRiskOrSuspectOrSymptoms	Never	No	DontKnow	Positive
TwentyToTwentyFour	Female	Married	No	MaritalReunionOrUntrusstPartner	Never	No	DontKnow	Negative
ThirtyToThirtyFour	Female	Married	No	PlanForFuture	NotAppl	No	Negative	Negative
TwentyFiveToTwentyNine	Female	Widowed	Yes	HadRiskOrSuspectOrSymptoms	Never	No	Positive	Positive

Table 5.22. Sample data prepared as single word value for all attributes

WEKA Explorer has good interface to open files in ARFF format and load to the software. Hence the file *data.arff* was loaded successfully. Figure 5.10 below shows the data loading screen of WEKA explorer.

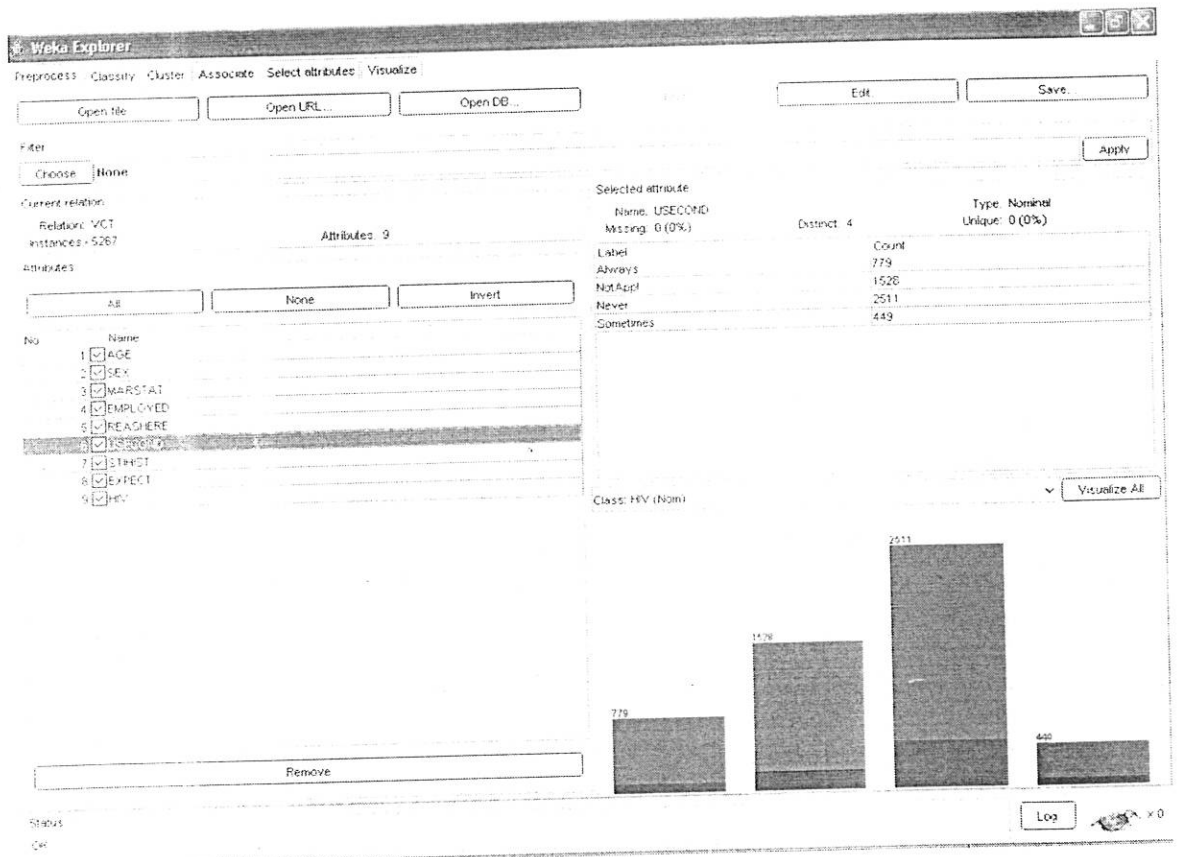


Figure 5. 10 The WEKA Explorer screen

WEKA includes an implementation of the APRIORI algorithm for generating association rules. The APRIORI algorithm can only deal with nominal attributes that the data for this research has already been prepared for.

The association mining output of Apriori is shown in Figure 5.11. The last part gives the association rules that are found. The number preceding the $\Rightarrow$ symbol indicates the rule's support, that is, the number of items covered by its premise. Following the rule is the number of those items for which the rule's consequent holds as well. In parentheses is the confidence of the rule- in other words, the second figure divided by the first. In this

research, with first trial (with default parameters and with all the attributes) the confidence of the best rule (rule with highest confidence) was 0.94.

Apriori

Minimum support: 0.2

Minimum metric <confidence>: 0.9

Number of cycles performed: 16

Generated sets of large itemsets:

Size of set of large itemsets L(1): 15

Size of set of large itemsets L(2): 55

Size of set of large itemsets L(3): 50

Size of set of large itemsets L(4): 16

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

Best rules found:

1. MARSTAT=NeverMarried EXPECT=Negative 1473 ==> HIV=Negative 1388 conf:(0.94)
2. MARSTAT=NeverMarried USECOND=Never HIV=Negative 1464 ==> STIHIST=No 1371 conf:(0.94)
3. SEX=Female USECOND=Never 1286 ==> STIHIST=No 1191 conf:(0.93)
4. MARSTAT=NeverMarried REASHERE=PlanForFuture 1522 ==> HIV=Negative 1406 conf:(0.92)
5. SEX=Male MARSTAT=NeverMarried STIHIST=No 1626 ==> HIV=Negative 1496 conf:(0.92)
6. EXPECT=Negative 1843 ==> HIV=Negative 1695 conf:(0.92)
7. SEX=Male MARSTAT=NeverMarried EMPLOYED=Yes STIHIST=No 1291 ==> HIV=Negative 1183 conf:(0.92)
8. MARSTAT=NeverMarried USECOND=Never 1685 ==> STIHIST=No 1542 conf:(0.92)
9. USECOND=Never HIV=Negative 2018 ==> STIHIST=No 1845 conf:(0.91)
10. USECOND=Never EXPECT=DontKnow HIV=Negative 1199 ==> STIHIST=No 1090 conf:(0.91)

Figure 5. 11. Output from the APRIORI Association Rule Learner for the First Experiment

APRIORI orders rules according to their confidence and uses support as a tie-breaker. Preceding the rules are the numbers of item sets found for each support size considered. By default, APRIORI tries to generate ten rules. It begins with maximum support of 100% of the data items and decreases this in steps of 5% (default Delta) until there are at least ten rules with the required minimum confidence, or until the support has reached a lower bound of 10%. The minimum confidence is set to 0.9 by default. As it is seen from the beginning of Figure 5.11, the minimum support decreased to 0.2, or 20%, before the required number of rules could be generated which involved a total of 16 iterations. These parameters were changed by setting the corresponding options. The principal options are summarized in Table 5. 23.

Table 5. 23 principal options for the APRIORI association rule learner.

Options	Function
-N < required number of rules>	Specify required number of rules
-C ,<minimum confidence of a rule>	Specify minimum confidence of a rule
-D <delta for minimum support >	Specify delta for decrease of minimum support
-M < lower bound for minimum support	Specify lower bound for minimum support

Now the researcher proceeded by changing these settings in different combination with varying parameters mainly: minimum support, minimum confidence, delta for minimum support and number of rules to be generated. The rules were extracted and found to be mostly sound by domain experts. The discussions of the results were made on those, that domain experts were found to be interested in.

In the first experiment, all the records of dataset and all the previously selected attributes or risk factors including the target attribute [AGE, SEX, MARSTAT, EMPLOYED, REASHERE, USECOND, STIHIST, EXPECT & HIV] were treated with default parameters. That is, with minimum support 0.1, minimum confidence 0.9, delta 0.05, and 10 number of rules.

The best rules were obtained after 16 cycles (Figure 5.11). Among the rules found, it was seen (Rule #1) that if some one is unmarried and he/she expects negative result there is 94% confidence with 27.97% support that the person's result is HIV negative. The second rule indicated that if, a person is not married, never use condom and his/her HIV test is negative there is 94% confidence with 27.8% support that history of sexually transmitted infection (STIHIST) of the client is NO. Rule #4 indicates that for all unmarried people whose reason of test is plan for future there is 92% confidence with 28.9% support to be HIV negative.

Similarly, from rule #7, unmarried male (or clients) who are employed have 92% confidence with 24.51% support that their HIV test result will be negative. This was found as indication that employment for the men is one factor to reduce the chance of infection.

Though there were more rules found with this trail the rest did not sound good to the view of the experts and discussion was resumed with other trials.

Next experimentation was conducted with increasing parameter for number of rules to 20. On the second experiment, changing the number of rules to 20, the observed results are as follows. The number of cycles performed to attain this was 17 epochs. Here the additional information that the result showed was:

If some one is not married, his/her STIHIST is not applicable, and USECOND is also not applicable; there is 98% confidence that HIV test is negative. The support of this rule is 16.55%. That means lower number of records relatively exhibits this feature.

All the above results were observed from experiment done with all the records (records with both positive and negative test results) and selected fields including HIV attribute. The experiment continued for each class separately. For records with positive results first and negative results next.

For the HIV positive class, since it is clear that occurrence of any item will have 100% confidence to result in HIV Positive class, the researcher used minimum support to control the interestingness of the rules. Accordingly, for the entire attributes including the target field (HIV), minimum support was set to 0.3 and only rules that can meet this number of records were considered. Figure 5.12 depicts the output of this experiment.

With these parameters, the rules were found after 11 epochs. The result showed that records whose STIHIST is NO has also shown positive HIV with 75.53% support. That means we have 100% confidence that 679 records of the 899 HIV+ total are also NO in STIHIST (Rule #1)

Apriori

=====

Minimum support: 0.45

Minimum metric <confidence>: 0.9

Number of cycles performed: 11

Generated sets of large itemsets:

Size of set of large itemsets L(1): 9

Size of set of large itemsets L(2): 10

Size of set of large itemsets L(3): 2

Best rules found:

1. STIHIST=No 679 ==> HIV=Positive 679 conf:(1)
2. EXPECT=DontKnow 600 ==> HIV=Positive 600 conf:(1)
3. SEX=Female 568 ==> HIV=Positive 568 conf:(1)
4. EMPLOYED=Yes 494 ==> HIV=Positive 494 conf:(1)
5. USECOND=Never 493 ==> HIV=Positive 493 conf:(1)
6. SEX=Female STIHIST=No 476 ==> HIV=Positive 476 conf:(1)
7. REASHERE=HadRiskOrSuspectOrSymptoms 473 ==> HIV=Positive 473 conf:(1)
8. MARSTAT=NeverMarried 469 ==> HIV=Positive 469 conf:(1)
9. STIHIST=No EXPECT=DontKnow 464 ==> HIV=Positive 464 conf:(1)
10. EMPLOYED=No 405 ==> HIV=Positive 405 conf:(1)

Figure 5.12. Output from the APRIORI Association Rule Learner for only positive records

The result tells that with 66.74% support whose expectation is "Don't Know" get HIV+ test result. Among 899 HIV positives, 600 don't expect that they are HIV infected or not (Rule #2). *Therefore, the experiment has shown that the expectations of the visitors may not have direct implication that the result will be HIV+.* Similarly, records of individuals who never use condom have shown more than fifty percent support in the HIV+ class. The support of the rule is 54.95% (494 of total 899 HIV+) Rule#5.

Just like having reason of future plan in the case of HIV negative, reasons like "had risk", "suspect partner" and "seen symptoms" are main purpose why people who have positive HIV test visit the service. *There is 100% confidence and 52.61% support that "Had Risk or Suspect or Symptoms" is reason for HIV positive class. That is, for all 473 records out of 899 HIV+ records their reason to visit VCT center is risk related.*

Now the investigation is narrowed to the risk factors alone without including the HIV attribute. Here experiments were conducted for the records with HIV+ value in the target field. As seen in the previous experiments, the total number of records with positive values is 899. The attributes involved are: AGE, SEX, MARSTAT, EMPLOYED, REASHERE, USECOND, STIHIST, and EXPECT.

Apriori

Minimum support: 0.1

Minimum metric <confidence>: 0.9

Number of cycles performed: 18

Generated sets of large itemsets:

Size of set of large itemsets L(1): 21

Size of set of large itemsets L(2): 84

Size of set of large itemsets L(3): 113

Size of set of large itemsets L(4): 54

Size of set of large itemsets L(5): 5

Best rules found:

1. AGE=TwentyToTwentyFour EMPLOYED=No STIHIST=No 111 ==> SEX=Female 105 conf:(0.95)
2. AGE=TwentyToTwentyFour EMPLOYED=No 138 ==> SEX=Female 127 conf:(0.92)
3. AGE=TwentyToTwentyFour USECOND=Never STIHIST=No 109 ==> SEX=Female 99 conf:(0.91)
4. SEX=Female MARSTAT=NeverMarried USECOND=Never EXPECT=DontKnow 101 ==> STIHIST=No 91 conf:(0.9)
5. AGE=TwentyFiveToTwentyNine EMPLOYED=No STIHIST=No 111 ==> SEX=Female 100 conf:(0.9)
6. SEX=Female EMPLOYED=Yes REASHERE=HadRiskOrSuspectOrSymptoms 120 ==> STIHIST=No 108 conf:(0.9)

Figure 5.13. Output from the APRIORI Association Rule Learner for only positive records without HIV field.

As it is observed from the above rules, the algorithm has used some attributes that show strong co-occurrence to construct rules. To determine the importance of the above rules and the attributes used to construct those rules, the association of the attributes with the predicates were evaluated based upon comments given by domain experts and reports of previous research works.

Among the visitors of positive test results within the age range of 20-24 who are not employed and STI history is NO are also females. Rule #1 shows that there is association among these factors with 95% confidence on 111 records out of 899 total (12.35%) support. Together with Rule #3 it is also shown that "never use condom" is associated with these factors by 91% confidence. *Generalizing the results observed (especially Rules #1, #2, #5), it is presented that unemployment for women in the age of 20's is highly associated with HIV infection.*

From the above rules, it was also observed that marital status (rule#4), condom use (rule #3&4), and expectation (rule#4) have strong association with sexually transmitted infection history. These rules show the co-occurrence of behaviour of clients to be unmarried, never use condom and have no STIHIST. *Yet they are infected and they don't expect their result to be either positive or negative. This does not disprove that HIV infection is related to positive STIHIST but this may be because most of the people*

respond "NO" even when they do not know their sexually transmitted infection history (as commented by domain experts).

Keeping on narrowing the combination of these risk factors, the researcher conducted further experiments with fewer attributes. The next mining was done omitting the attribute "EXPECT". The obtained result was almost similar to the one seen before and no new knowledge was recognized from it. So, additional trials were performed further with another combination. Here only AGE, SEX, MARSTAT, EMPLOYED, REASHERE, USECOND fields were used to mine for another data pattern. With the default parameters, only rule which is one of the previously observed was generated. Hence another trial was made with 80% minimum confidence. The association came up with best rules after 18 cycles as seen in the Figure 5.14.

Apriori

=====

Minimum support: 0.1

Minimum metric <confidence>: 0.8

Number of cycles performed: 18

Generated sets of large itemsets:

Size of set of large itemsets L(1): 16

Size of set of large itemsets L(2): 46

Size of set of large itemsets L(3): 32

Size of set of large itemsets L(4): 1

Best rules found:

1. AGE=TwentyToTwentyFour EMPLOYED=No 138 ==> SEX=Female 127 conf:(0.92)
2. AGE=TwentyToTwentyFour USECOND=Never 131 ==> SEX=Female 115 conf:(0.88)
3. AGE=TwentyFiveToTwentyNine EMPLOYED=No 132 ==> SEX=Female 115 conf:(0.87)
4. EMPLOYED=No USECOND=Never 224 ==> SEX=Female 192 conf:(0.86)
5. MARSTAT=Married EMPLOYED=No 129 ==> SEX=Female 110 conf:(0.85)
6. SEX=Male MARSTAT=NeverMarried REASHERE=HadRiskOrSuspectOrSymptoms 114 ==> EMPLOYED=Yes 96 conf:(0.84)
7. AGE=TwentyToTwentyFour 233 ==> SEX=Female 196 conf:(0.84)
8. EMPLOYED=No 405 ==> SEX=Female 338 conf:(0.83)
9. EMPLOYED=No REASHERE=HadRiskOrSuspectOrSymptoms 198 ==> SEX=Female 165 conf:(0.83)
10. SEX=Male REASHERE=HadRiskOrSuspectOrSymptoms 188 ==> EMPLOYED=Yes 155 conf:(0.82)

Figure 5.14 Output from the APRIORI Association Rule Learner for few selected fields

Only selected rules commented by domain experts are picked up and discussed here. Interesting findings identified from this rules were, the association of marital status and gender, employment, reason of the test, and HIV infection. As it was observed in the rules #6, #9, and #10, it is emphasized that unemployment was not associated to having risk HIV infection for the male group. For instance, it is depicted in rule #6 that if some

one is *male, unmarried and his reason is having risk, suspect or symptoms there is 84% confidence that he is also employed*. But in the case of female, especially as indicated in rule #9, if a record shows *unemployed and reason for test is having risk or suspect, or symptoms, with 83% confidence that the client is female*.

Another trial was performed with combination of attributes AGE, MARSTAT, REASHERE, and USECOND. Figure 5.15 below shows the output of the association rule learner for this set of fields.

```

Apriori
=====
Minimum support: 0.05
Minimum metric <confidence>: 0.6
Number of cycles performed: 19
Generated sets of large itemsets:
Size of set of large itemsets L(1): 17
Size of set of large itemsets L(2): 34
Size of set of large itemsets L(3): 12
Best rules found:

1. MARSTAT=Married REASHERE=PlanForFuture 65 ==> USECOND=Never 49  conf:(0.75)
2. MARSTAT=Married 235 ==> USECOND=Never 174  conf:(0.74)
3. MARSTAT=Married REASHERE=HadRiskOrSuspectOrSymptoms 104 ==> USECOND=Never 75  conf:(0.72)
4. MARSTAT=Widowed 82 ==> REASHERE=HadRiskOrSuspectOrSymptoms 59  conf:(0.72)
5. AGE=TwentyFiveToTwentyNine MARSTAT=Married 72 ==> USECOND=Never 50  conf:(0.69)
6. AGE=TwentyToTwentyFour REASHERE=HadRiskOrSuspectOrSymptoms 121 ==> MARSTAT=NeverMarried 84  conf:(0.69)
7. USECOND=Sometimes 86 ==> MARSTAT=NeverMarried 59  conf:(0.69)
8. AGE=TwentyToTwentyFour 233 ==> MARSTAT=NeverMarried 159  conf:(0.68)
9. USECOND=Always 107 ==> MARSTAT=NeverMarried 67  conf:(0.63)
10. AGE=TwentyToTwentyFour USECOND=Never 131 ==> MARSTAT=NeverMarried 81  conf:(0.62)

```

Figure 5. 15 Output from the APRIORI Association Rule Learner for another fields

From the above rules, it was observed that married people whose reason of test is for future plan, or having risk or suspect or symptoms were found that they never use condom (rule #1 & #3). This association has 75% confidence and this indicates that they have good trust for their spouses but still their test result show HIV+. At the same time, same experiment showed that among the infected people, who say they always use condom are with 63% confidence unmarried (rule #9). *These two points show that the concern of using condom seems to be more for unmarried person though the probability of infection is almost closer for every body (married or unmarried).*

Experts at the domain area of this research problem have shown their interest to see also the relationship between these selected attributes with group of clients whose test result is HIV negative. Accordingly, the researcher committed himself to investigate association of these selected attributes on HIV-negative records. The number of records with negative value counted 4368 out of 5267. As a result of the experiment the following patterns were observed.

```

Apriori
=====
Minimum support: 0.45
Minimum metric <confidence>: 0.9
Number of cycles performed: 11
Generated sets of large itemsets:
Size of set of large itemsets L(1): 7
Size of set of large itemsets L(2): 11
Size of set of large itemsets L(3): 5
Best rules found:

1. MARSTAT=NeverMarried 3526 ==> HIV=Negative 3526   conf:(1)
2. STIHIST=No 3174 ==> HIV=Negative 3174   conf:(1)
3. EMPLOYED=Yes 2758 ==> HIV=Negative 2758   conf:(1)
4. EXPECT=DontKnow 2499 ==> HIV=Negative 2499   conf:(1)
5. MARSTAT=NeverMarried STIHIST=No 2476 ==> HIV=Negative 2476   conf:(1)
6. SEX=Male 2421 ==> HIV=Negative 2421   conf:(1)
7. MARSTAT=NeverMarried EMPLOYED=Yes 2217 ==> HIV=Negative 2217   conf:(1)
8. EMPLOYED=Yes STIHIST=No 2103 ==> HIV=Negative 2103   conf:(1)
9. SEX=Male MARSTAT=NeverMarried 2041 ==> HIV=Negative 2041   conf:(1)
10. USECOND=Never 2018 ==> HIV=Negative 2018   conf:(1)

```

Figure 5. 16 Output from the APRIORI Association Rule Learner for HIV negative instances

With default parameters the result was found to be of no much new information. Only rules with very few itemsets that do not give any insight to the data were observed. But this experiment showed that among the clients whose test results are negative, with 80.72% support they were found to be unmarried person. From the experiment we have 100% confidence to say this fact holds true with the mentioned support. This indicates that majority of VCT clients (especially with HIV-negative) are unmarried community.

The researcher took another option in order to get association for many factors. The option taken was increasing the number of rules to 100 with 90% minimum confidence and 20% minimum support. As a result of this a long list of rules (99 rules) was

generated among which only some meaningful, sounding rules were taken and discussed.

One of the important findings noticed here is new insight about risk feeling of the clients and HIV test result. Previously it was known that the clients whose reason for test is plan for future are associated with HIV-negative class. This truth has also been verified with experiment too, as pointed out earlier. But here at this stage of the research, another interesting finding (though may not sound good) was observed. The experiment disclosed that people whose reason for test is having risk, suspect or symptoms is also associated to HIV negative result with a promising evidence. *With 36.33% support of the records and 100% confidence, people having such risky feelings were found to be seronegative.* This was additional previously hidden information that domain experts were impressed to here about. They confessed that it helps counsellors to tell the clients with full evidence that a person who has risky perception of oneself has better chance to be uninfected.

Upon efforts made to study the association of the risk factors excluding the target class (HIV field), it was seen possible to come up with further hidden knowledge. Figure 5.17 displays output of one of the selected attributes checked.

```
Apriori
=====
Minimum support: 0.1
Minimum metric <confidence>: 0.9
Number of cycles performed: 18
Generated sets of large itemsets:
Size of set of large itemsets L(1): 12
Size of set of large itemsets L(2): 18
Size of set of large itemsets L(3): 7
Best rules found:

1. AGE=FifteenToNineteen USECOND=NotAppl 456 ==> MARSTAT=NeverMarried 448  conf:(0.98)
2. AGE=FifteenToNineteen 816 ==> MARSTAT=NeverMarried 771  conf:(0.94)
3. REASHERE=PlanForFuture USECOND=NotAppl 573 ==> MARSTAT=NeverMarried 541  conf:(0.94)
4. AGE=TwentyToTwentyFour REASHERE=PlanForFuture 578 ==> MARSTAT=NeverMarried 534  conf:(0.92)
5. USECOND=NotAppl 1315 ==> MARSTAT=NeverMarried 1187  conf:(0.9)
```

Figure 5 17 Output from the APRIORI Association Rule Learner for few fields of HIV negative instances

Among the many, one of the rules that sounded good was the relations exhibited among age, reason, and marital status of the customers (rule #4). It was observed that people

whose age is between twenty and twenty four, and who come for a reason of having risk or suspect, or symptoms are also not married. This was observed with 92% confidence that these factors are associated with 13.24% support. *Therefore, association of these factors revealed that between the specified ages, if a person is not married there is high chance that he/she had risk, or suspect or symptoms while the test result may still be negative.*

The overall process of this research made by employing both decision tree and association techniques demonstrated that data mining is a method that should be considered in HIV/AIDS prevention and control activities. From the results obtained, it can be generalized that the HIV risk factors can be experimentally recognized and their co-occurrence be investigated by using VCT dataset organized by service providers.

CHAPTER SIX

6. SUMMARY, CONCLUSION AND RECOMMENDATIONS

6.1. Summary and Conclusion

Since the past few decades, the world is moving through tremendous technological transformations which hugely affected the way people live and work. The power of information is duly recognized and information and communication technology flourished in a bid to manage this highly valued asset. Management Information Systems, Health Information Systems, Geographic Information Systems, Data Mining Technology, Artificial Intelligence, etc, are the core ones to be mentioned.

Computer Science has played a major role in this wave of changes. Some of its fields, Data Mining and Artificial Intelligence, have helped some previously unthinkable things materialize. Knowledge Discovery from Database (KDD) is one from the all set. KDD has emerged in recent years in order to bridge the gap of analysing large volume of data and extracting useful information and knowledge for decision making.

The application of data mining technology has increasingly become very popular and proved to be relevant for many sectors such as retail trade, health, telecommunications, and banking. Especially in health sector, data mining technology has been applied for patient survival analysis, diagnosis, for outcomes measurement, to improve patient treatment and decision making, etc.

This paper reports on a research that was set to explore the possible application of data mining technology in HIV/AIDS intervention in Ethiopian context, by using VCT data of CDC to identify determinant risk factors and assess their associations. The reported results are significant to CDC and other similar organizations for one major reason; the study involved a real life problem. That means the data used and the requirements set correspond to real problem.

The methodology employed consisted of three basic steps; data collection, data preparation, model building and testing. But, just like what any data mining project may

encounter, the steps were not strictly followed linearly. There were times where the research was forced to go back and forth between different steps in order to accomplish the project. Hence more than 65% of the total time allocated for the research was consumed by data collection and preparation.

The problem definition of the research was based on the accumulation and existence of huge amount of HIV related data and high demand of understanding and analyzing the data at the selected area. And also the HIV case demanded insight understanding of related factors. Following problem definition, defining data mining function was also one phase in the process. The data mining function of the study was defined to be mining association rules. But determinant risk factors were identified using decision tree classification. The other tasks performed during the process were feature selection and transforming the data. A total data set of 5267 VCT records were used.

The experimentation part was two phase activity. In the first part, decision tree classifier was built by using KnowledgeSTUDIO software in order to identify frontier risk factors that have high discrimination power of classifying into HIV positive and HIV negative. Several classifiers were built and only those with high performance as checked by unseen data were selected. Accordingly, attributes "AGE", "SEX", "MARSTAT", "EMPLOYED", "REASHERE", "USEDOND", "STIHIST", "EXPECT" were identified as determinant factors to discriminate VCT clients as positive and negative test result. The selection was done experimentally and supported by domain experts.

During the second phase of the experiment, WEKA Tool was used to mine and generate association rules for the identified risk factors. This has also involved additional preprocessing activities to reform the data that is suitable to WEKA software. A number of rules were generated and some that introduced new and meaningful information were discussed. Various association options/parameters were used to control the interestingness of the rules.

In general, encouraging results were obtained from activities and the findings were appreciated for their importance. To point out some of these:

One of the important finding in this research recognized as new knowledge was that people who sometimes use condom have 3% worse chance than even those who never use. Among the clients who say using condom is not applicable, in average 15% of them were found HIV infected. From which 17.17% was female only 12.70% was male. This new information has a public domain importance.

Variations in HIV infection among the unemployed especially in female was observed pronounced. This identification was also believed to assist government and policy makers in activities of HIV/AIDS intervention. The research had also proved the earlier understanding that women commit themselves to unsafe sexual practice for material benefit.

Proved that, 98.86% of peoples who previously known their seropositive status have confirmed exactly. This was recognized as an indication for the perfection of the HIV test service.

As encouraging feedback, 99.07% of the people who had used VCT service previously are much benefited and have strong stack to take care of themselves from ways that leads to HIV infection and realized consistent use of the service is worthwhile.

The experiment disclosed that people whose reason for test is having risk, suspect or symptoms is also associated to HIV negative result with a promising evidence. With 36.33% support of the records and 100% confidence, people having such risky feelings were found to be seronegative.

Observed that, above 75% of the data is collected from unmarried people. The implication of this figure is that unmarried individuals are concerned about HIV case than others and get benefits of VCT. But in contrary, the ration of people infected with HIV as seen from the result is very small in unmarried group. The probability is only 11.61% which is much smaller than the general probability (17.06%).

In concluding the report, the result obtained from the study can be considered as satisfactory and the research work has proved the potential applicability of data mining technology to identify determinant risk factors and assess their association. The result

has wide advantage in broadening the general understanding about HIV infection for the society, health professionals, government and policy makers at large.

6.2. Recommendations

This experimental research work has verified the potential applicability of data mining technology to investigate HIV risk factors and their association based on the available data of VCT clients. Thus, based on the findings of this research work, the researcher would like to make the following recommendations particularly in relation to the possible application of data mining technology in supporting HIV/AIDS prevention and control activities aimed at reducing HIV related problems.

In the process of this research work, it was learnt that more research and development efforts needs to be conducted to enable the full exploitation of data mining technology in the health sector of this country. In particular, the following areas were identified as deserving further research work:

1. In this research work, an attempt has been made to assess the applicability of data mining technology to find association rules for risk factors by using some set of variables that were considered important by experts. For a number of other variables, however, it remains to investigate further the effect of those variables to study with better accuracy and performance than the findings of this research work.
2. It is indeed very important to assess the applicability of data mining techniques in studying risk factors behaviour by using client datasets gathered from different health centres than only VCT dataset. This in turn would help to compare the results obtained with those VCT datasets to the results obtained using non-VCT datasets like blood donors ANC datasets.
3. Since this study has used a small percentage of the two years VCT dataset to pursue the investigation, it is appropriate to work on comprehensive study by large datasets taken from the main database of the CDC.

4. Although encouraging results were obtained in this research work, the researcher had experienced serious shortage of time to make some more trials to build and test more accurate and better performing classifiers. The researcher has noted that the accuracy rate tends to increase to a desirable level with more and more trials of different combinations of the variables. It is therefore, recommended that more time should be allocated to build and evaluates models with a desirable level of accuracy.
5. More evaluation should be conducted particularly to assess whether rules qualify as a hypothesis for conventional research activities.
6. The possibility of incorporating the findings of this study in another (and operational) application should be explored.

- Prather, Jonathan C, et. al, (2002). *Medical Data Mining: Knowledge Discovery in a Clinical Data Warehouse*. URL: <http://www.amia.org/pubs/symposia/D004394.PDF>
- Raghavan, Vijav V., et. al. (1998). *A Perspective on Data Mining*. Journal of the American Society For Information Science; 49(5)
- Shegaw, Anagaw (2002). *Application of Data Mining Technology To Predict Child Mortality Pattern: The Case Of Butajira Rural Health Project*. Unpublished Masters Thesis, Addis Ababa University, Department of Information Science.
- StatSoft (2003). Data Mining Techniques. URL:<http://www.statsoftinc.com/textbook/stathome.html>
- Sivanandam, S.N.,S.Sumathi,.T.Hamsapriya, Mr.K.Babu, (2004). *Parallel Buddy Prima – A Hybrid Parallel Frequent itemset mining algorithm for very large databases*. www.acadjournal.com
- SMCC, (1997) ``Scalable Data Mining with Sun<sup>TM</sup> Ultra<sup>TM</sup> Enterprize services ``
URL: [www.gca.net/solutions/whitepapers/sun/sun-data mining .htm !](http://www.gca.net/solutions/whitepapers/sun/sun-data%20mining.htm)
- Susan Craw. Application of Data Mining: Association Rules. www.comp.rgu.ac.uk/staff/smc/teaching/datamining/
- TCC (1999). *Introduction to Data Mining and Knowledge Discovery* Two Crows Corporation
. Available at URL: <http://www.twocrows.com>
- Theeuwens, Marc, et. al. (2001). *Neural Networks Analysis to Predict Treatment Outcome In Patients With Ovarian Cancer*. URL: <http://www.mbfys.kun.nl/mbfys/people/bertt/>
- Trybula, Walter J. (1997). *Data Mining and Knowledge Discovery*. Annual Review of Information Science and Technology (ARIST);(32):197-229.
- UNAIDS/WHO (2004), *AIDS epidemic update 2004*. Joint United Nations Programme on HIV/AIDS (UNAIDS) World Health Organization (WHO)
- UNAIDS/WHO (2002), *AIDS epidemic update 2002*. Joint United Nations Programme on HIV/AIDS (UNAIDS) World Health Organization (WHO)
- UNAIDS (2000), *Voluntary Counseling and Testing (VCT) Technical update*.
<http://www.rhrc.org/resources/sti/hivaidsmanual>
- UNAIDS, (2000), *Global HIV/AIDS Epidemic*, at <http://www.who.int/emc-hiv/>
- WHO. (1994). *Counselling for HIV/AIDS: A key to caring*. For policy makers, planners and implementers of counselling activities. Geneva,.

Wiley, J. (1997), *Introduction to Data Mining and Knowledge Discovery*. Two Crows Corporation.

Available at URL: <http://www.twocrows.com>

Witten, H. and Frank, E. (2000). *Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations*. San Francisco: Morgan Kaufmann Publishers.

AppendixA: The Original Data With Sample Records and attributes

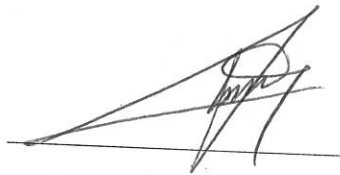
COUNTRY	REGION	WOREDA	SITECODE	STYPE	ORGTPE	RESCAT	CLIENT	VDATE	RETURNV	NEWCODE	AGE	SEX	COUNCODE	PARTCODE	COUPCODE	GROUPEDU	SESSTYPE	MARSTAT	COUPTYPE	EDUEXP	EMPLOYED	OCCUPAT	KNOW1	KNOW2	KNOW3	KNOW4	KNOW5	KNOW6	KNOW7	KNOW8	REFERBY	REASHERE	PRTESTED			
ET	14	4	1	1	1	1	1	13/03/2002	0	0	50	2	C04																							
ET	14	4	1	1	1	1	2	13/03/2002	0	0	30	1	C03			0	1	3	98	1	1	4														
ET	14	4	1	1	1	1	3	03/02/2003	1	0	26	1	C03			0	1	2	98	2	0	99									2	4	0			
ET	14	4	1	1	1	1	4	13/03/2002	0	0	27	1	C02			0	1	2	98	3	1	8				1					1	2	0			
ET	14	4	1	1	1	1	5	13/03/2002	0	0	24	1	C03			0	1	2	98	2	1	8				1										
ET	14	4	1	1	1	1	6	13/03/2002	0	0	27	1	C04			0	1	1	98	3	1	2										1	16	2		
ET	14	4	1	1	1	1	7	13/03/2002	0	0	32	1	C02			0	1	2	98	3	0	10										1	1	0		
ET	14	4	1	1	1	2	8	13/03/2002	0	98	35	1	C05			0	1	2	98	3	1	9										1	2	0		
ET	14	4	1	1	1	1	9	13/03/2002	0	0	22	2	C03			0	1	2	98	7	1	3	1									1	1	2		
ET	14	4	1	1	1	1	10	13/03/2002	1	1	25	2	C04			0	1	2	98	7	1	2										2	1	2		
ET	14	4	1	1	1	1	11	13/03/2002	0	0	37	1	C02			0	1	2	98	3	0	99										1	18	0		
ET	14	4	1	1	1	1	12	13/03/2002	0	0	32	1	C05			0	1	2	98	3	0	10										2	3	2		
ET	14	4	1	1	1	1	13	13/03/2002	0	0	22	1	C03			0	1	3	98	1	1	5										1	1	0		
							14	14/03/2002	1	98	30	2	C05			0	1	3	98	2	1	3				1							1	1	0	
ET	14	4	1	1	1	1	15	14/03/2002	0	0	30	2	ZC03			0	1	4	98	3	1	5				1						1	2	0		
ET	14	4	1	1	1	1	16	14/03/2002	0	0	29	1	C04			0	1	99	98	3	1	99												2		
ET	14	4	1	1	1	1	17	14/03/2002	0	0	21	2	C03			0	1	1	98	2	0	99											1	1	2	
							18	14/03/2002	0	98	19	1	C05			0	1	2	98	2	1	8											2	16	2	
ET	14	4	1	1	1	1	19	14/03/2002	0	98	22	2	ZC03			0	1	2	98	3	0	10				1							1	16	0	
																0	1	2	98	3	0	10											1	2	0	
																																		1	2	2

Appendix A: Continued

TMONTH	TYEAR	PREVLOC	EVERHAD	EXPTIME	1SECOND	LASTSEN	STHIST	CASUAL	STEADY	SEXWORK	SEXEXCH	PREG	EXPECT	NOTIFPRE	HIV	SYPHILIS	DISCOR	RERES	CONDACC	NUMCOND	RISKPLAN	NOTIFPOS	REFER1	REFER2	REFER3	REFER4	REFER5	REFER6	REFUSE	COUNSEL	GAVETEST	GAVESUSU	REFERRED	CONDEMO
		86	1	1	2	0	0	2	0	0	0	0	98	2	1	95	98	0	0	0	0	1	2	1						1	1	1	0	
3	2002	86	1	4	0	0	1	0	0	98	0	98	97	98	1	95	98	0	96	0	1	98	1							1	1	1	0	
		98	1	4	0	0	0	0	1	98	0	98	0	1	0	95	98	0	96	0	1	1								1	1	1	1	
		98	1	2	1	1	1	0	0	98	0	98	97	1	0	95	98	0	96	0	1	1								1	1	1	0	
0		98	1	4	0	0	0	0	0	98	0	98	0	1	0	95	98	0	96	0	1	1								1	1	1	1	
8	2001	98	0	98	98	98	98	0	0	98	0	98	0	1	0	95	98	0	96	0	1	1								1	1	1	0	
1	2002	2	1	4	1	1	1	0	0	0	0	98	97	1	0	95	98	0	96	0	1	1								1	1	1	0	
1		98	0	98	98	98	98	0	1	98	0	98	98	98	0	95	98	0	98	0	1	98								1	1	1	0	
1	1997	2	1	1	2	0	0	0	0	98	0	98	97	3	0	95	98	0	96	0	0	1	1							1	1	1	0	
		98	1	3	2	0	1	0	0	0	0	98	98	1	0	95	98	0	96	0	1	1								1	1	1	0	
		98	1	4	2	0	1	0	0	98	0	98	97	98	1		98	0	96	0	1	1								1	1	1	0	
		98	1	4	2	0	1	0	0	98	0	98	1	98	1	95	98	0	96	0	1	98	1							1	1	1	0	
8	2001	3	1	2	2	0	0	1	1	98	0	98	1	1	0	95	98	0	96	0	1	98	1							1	1	1	1	
9	2001	3	1	1	2	0	0	0	1	0	0	0	0	2	1	95	98	1	96	0	1	2								1	1	1	1	
1	2001	3	1	1	0	98	97	0	2	0	0	0	1	1	0	95	98	0	96	0	1	3								1	1	1	0	
		98	1	4	0	0	0	0	0	98	0	98	0	2	0	95	98	0	96	0	1	1								1	1	1	0	
		98	1	4	0	0	0	0	0	98	0	98	0	2	0	95	98	0	96	0	1	2								1	1	1	0	
2	1996	3	1	4	98	0	0	0	0	98	0	98	98	1	0	95	98	0	96	0	1	1								1	1	1	0	

Declaration

I, the undersigned, declare that this thesis is my original work and has not been presented as a partial degree requirement for a degree in any other university and that all sources of materials used for the thesis have been duly acknowledged.



Abraham Tesso Nedjo

June 2005

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

Dr. Bandaru Rama Krishna Rao

June 2005