ADDIS ABABAUNIVERSITY

SCHOOL OF GRADUATE STUDIES

SCHOOL OF INFORMATION SCIENCE

AND SCHOOL OF PUBLIC HEALTH

PREDICTING INFANT IMMUNIZATION STATUS IN ETHIOPIAN: THE CASE OF ETHIOPIA DEMOGRAPHIC AND HEALTH SURVEY 2011

By: Hiwot Abebe

Advisors: Million Meshesha (PhD)

Wubegzier Mekonnen (PhD)

A THESIS SUBMITTED TO THE SCHOOL OF GRADUATE STUDIES IN PARTIAL FULLFILMENT OF THE REQUIREMENTS FOR THE DEGREE OF MASTERS OF SCIENCE IN HEALTH INFORMATICS

June, 2014

Addis Ababa, Ethiopia

Addis Ababa University

School of Graduate Studies

School of Information Science And

School of Public Health

PREDICTING INFANT IMMUNIZATION STATUS IN ETHIOPIAN: THE CASE OF ETHIOPIA DEMOGRAPHIC AND HEALTH SURVEY 2011

By: Hiwot Abebe

A Thesis Submitted to the School of Graduate Studies of Addis Ababa University in Partial Fulfillment of the Requirements for the Degree of Masters of Science in Health Informatics

Approved by the Examining Board:

Name	Title	Signature	Date
_____	Chairperson	_____	_____
_____	Advisor	_____	_____
_____	Advisor	_____	_____
_____	Examiner	_____	_____
_____	Examiner	_____	_____

ACKNOWLEGEMENT

Above all, I would like to glorify the almighty GOD and St. Virgin Marry for giving me the ability to be where I am. You have done so much for me, O Lord. No wonder I am glad! I sing for joy, Amen!

Secondly, I would like to forward a very much grateful thank to my advisors Dr. Million Meshesha and Dr. Wubegzier Mekonnen for their constructive comments and guidance. But special thanks go to my brother Ato Gizaw Abebe, without whom this research would have not been a success. Gizaw, your helpful personality lasts in my heart forever.

I would also like to thank the FMOH EPI Focal Person, Ato Michael Mesfin, who helped in selecting attributes related to my research topic select the attribute up to constructive comments and Ato Sintaye, who contributed a lot in selecting the best rule based on the algorithm generate. I also thank, Ato Abebe Bekele, Acting Director of Health System for assisting me in giving information I needed.

I am very much grateful to my best friend Ato Misganaw Tadese, for his unreserved advice, guidance and constructive comments from the beginning to the end of the thesis. I am also grateful to Ato Amare Mezmur to assist me morally and materially whenever I needed. Many thanks is also goes to Ato Niguse Shiferaw, 2015 Consultant Demographer of Ethiopian Case Study, for allowing me to carry out this research using the required data from the EDHS 2011 database and for his unreserved constructive comments. Countless thanks goes to Dr. Dinkye Abebe, for his productive comments. Ato Elias, Ato Abiye Walelgn, Ato Binyame Tezera, Ato Tedrose Eshite, W/ro Sesen Terefe and Ato Abrehame were also very crucial in editing the thesis; I thank them very much.

Finally, I would like to express my thankfulness to my mother Mulatua H/Mechael, my elder sister Elizabeth Abebe, and all the rest of my family and friends for giving me love and encouragement since the time of my admission to the end of my study.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ACRONYMS

ANN	Artificial Neural Network
BCG	Bacillus Chalmette-Guerin
DPT	Diphtheria-Pertussis-Tetanus
DPT-HepB-Hib	DPT with Hepatitis B and Homophiles influenza type b
EDHS	Ethiopian Demographic and Health Survey
EHNRI	Ethiopia Health and Nutrition Research Institution
EPI	Expanded Programmed on Immunization
FMOH	Federal Ministry of Health
FPR	False Positive Rate
KDD	Knowledge Data Discovery
MDG4	Millennium Development Goal 4
OPV	Oral Polio Vaccine
PART	Partial Decision Tree
ROC	Receiver Operating Curves
SMO	Sequence Minimal Optimization
SVM	Support Vector Machine
TPR	True Positive Rate
WHO	World Health Organization

ABSTRACT

Background: Immunization is one of the most cost effective and efficient interventions saving the lives of many millions of infants and children from dying of infectious and preventable diseases. In 2007, approximately 27 million infants are not vaccinated against common childhood diseases and 2–3 million children are dying annually from easily preventable diseases and many more fall ill.

Objective: The research has a general objective of construct a predictive model using data mining technology that helps to predict the infants' immunization status in Ethiopia. The result of the study is expected to be important for different parties such as infants, health professionals, policy makers, programmers and researchers.

Methodology: This study is guided by a Hybrid-data mining model which is a six step knowledge discovery process model such as understanding of the problem, understanding of the data, preparation of the data, data mining, and evaluation of the discovered knowledge and use of the discovered knowledge. The study has used 8,210 instances, 12 predicting and one outcome variables to run the experiments. Due to the nature of the problem and attributes contained in the dataset, classification data mining task is selected to build the classifier models. The mining algorithms; J48 decision tree, sequence minimal optimization support vector machine, multilayer perceptron neural network and partial decision tree rule induction are used in all experiment due to their popularity in recent related works. Ten-fold cross validation technique is used to train and test the classifier models. Performance of the models is compared using accuracy, true positive rate, false positive rate, and the area under the Receiver Operating Characteristics curve.

Result: The J48 decision tree has given the best classification and a better predictive accuracy of the infant immunization status in Ethiopia. The experiment has generated a model with accuracy of 62.5%, weighted precision of 62.5% and weighted ROC area of 67.6% for the J48 decision tree. And if place of delivery = home region = Affar AND mother-education-level = no-education AND wealth-status = poor AND listening-to-radio = not-at-all AND mother-age = 25-29 AND parity = 6-7 then Unimmunised (10.0/1.0).Therefore, increase awareness creation among women in pastoralist communities so as to enhance vaccine coverage.

Conclusion: The results achieved from this research indicate that data mining is useful in bringing relevant information from large and complex EDHS dataset, and we can this information for predicting infant immunization status and decision making. The most important attributes that determine infant immunization status were place of delivery, region, mother's educational level, listening to radio, father education level, residence, mother age, wealth status, parity, distance to health facility and marital status.

CHAPTER ONE

INTRODUCTION

1.1. Background

Immunization is one of the most cost effective and efficient intermediations saving in the lives of many millions of infants and children from dying of infectious and preventable diseases [1]. Infants and children are therefore needed to get immunization service not to be affected by communicable, but preventable diseases. Due to this fact, World Health Organization (WHO) launched the Expanded Programmed on Immunization (EPI) in 1974 with the aim of immunizing the world's infant against the six major communicable diseases such as: diphtheria, pertussis, tetanus, tuberculosis, measles and polio myelitis [1]. Over the past decade, immunization programs have added new and underused vaccines to the original six vaccinations. They include vaccines against Hepatitis B (HepB), Haemophilus influenza type b (Hib) disease, mumps, pneumococcal disease, rotavirus, rubella, and in countries where needed yellow fever and Japanese encephalitis [1,2].

Furthermore, the combined effort of WHO, United Nation Development Program (UNDP),United Nation Children's Fund(UNICEF),the World Bank, other developmental agencies, and national programs resulted in the achievement of the global goal of 80% immunization in 1990 [3]. After 1990, global immunization coverage for infants under one year of age was maintained at 80% for the recommended three doses of Diphtheria-Pertussis-Tetanus (DPT) and polio by many countries both from developing and developed nations [3].

However, there is wide difference among regions and DPT3 coverage ranged from 91% in East Asia and pacific to 62% in East Africa and 42% in West and Central Africa [3].In Ethiopia vaccine preventable communicable diseases is also major public health problems. Accordingly, the prevention and control of these communicable diseases have received high priority. EPI was started in Ethiopia in 1980 with the aim of reducing morbidity and mortality of children from vaccine preventable diseases [4, 5]. In 1980, the immunization coverage for infants and children in Ethiopia was estimated to be less than 1% and the plan was to increase coverage by 10% every year in order to attain the goal of universal immunization coverage by 1990 [6]. However, in 1990, immunization coverage for less than one year infants was reported to be 49% for DPT3 among accessible population (about 50% of the total) and the dropout rate from the schedule

was 36%)[6]. Although the immunization coverage has showed significant improvement, due to those dropouts the intended goal for the year 1990 is not achieved.

Of the new vaccines, the pneumococcal vaccine has been shown to be associated with a 39% reduction in hospital admissions due to pneumonia from any cause [2]. Among children who survive an episode of pneumococcal meningitis, a big percentage is left with long term disabilities. Similarly, the rotavirus vaccine has been shown to reduce clinic visits and hospitalizations due to rotavirus diarrhea by 95% [2].

The Ethiopian MOH launched EPI in 1980 according to the recommendations of WHO that states Infants should be fully immunized with the following vaccines [8]: one dose of Bacillus Chalmette-Guerin (BCG) vaccine at birth (or as soon as possible); three doses of DPT with HepB and Hib (DPT-HepB-Hib) at 6, 10 and 14 weeks of age; at least three doses of Oral polio vaccine (OPV) at birth, and at 6, 10 and 14 weeks of age; and one dose of measles vaccine at 9 months of age. Therefore, infants are expected to be fully immunized by 12 months of age. The three doses of pentavalent replaced in DPT by the Ethiopian EPI Program in 2005 [9].

To achieve the millennium development goal four (MDG4) of reducing children's death by two-thirds in 2015, Ethiopia has adopted strategies such as sustainable outreach service (SOS) and reaching every district (RED) that focus on identifying bottlenecks and developing community ownership of the services in order to improve routine immunization services and increasing coverage[10].

The Ethiopian government has also introduced Health Extension Program (HEP) that helps to enhance health condition of people in all parts of the country including children and infants. The core element of the Health Extension Program (HEP) is the Health Extension workers (HEW), applying 16 components of the health extension package at Kebele level. Immunization is one of the preventive health interventions in the health extension package. The catchment population for each Kebele (where two HEWs are assigned) is 5,000 per health post and 25,000 per health centers in an agrarian setup and in the pastorals it is 3,000 per health post and 15,000 per health center respectively [11].

According to health sector development program IV (HSDP IV)(2010-2014), By the end of 2009, 33,819 HEWs were trained and deployed under health sector development program III

(HSDPIII) (2005-2009), covering 26% of households [11].Additional HEW have been trained and deployed since 2010[11].

To understand the immunization status of infants in Ethiopia factors affecting immunization status of infant are various socioeconomic and demographic factors may influence immunization coverage of infant such as parity, place of delivery, region, residence, mother education, father education, marital status, wealth status, distance of health facility and frequency lessening to radio there is a need to apply data mining technology. Data Mining (DM) is a process of discovering various models, summaries, and derived values from a given collection of data [12]. The process must be automatic or (more usually) semiautomatic. The patterns discovered must be meaningful and interesting in that they lead to some advantage, usually an economic one [13].

Data Mining has a potential to in identifying hidden knowledge from huge datasets [14]. It has been seriously used in the medical field, to include diagnosis of patient records to help identify best practices [14].

The term data mining has been mostly used by statisticians, data analysis, and the management information systems (MIS) communities [15]. Data mining is a step in the Knowledge Discovery in Databases (KDD) process consisting of applying computational techniques that, under acceptable computational efficiency limitations, produce a particular enumeration of patterns [16]. And it is emerging as a new active area of research which combines methods and tools from the fields of statistics, machine leaning, database management and data visualization [17]. Data mining techniques have been applied to many real life applications, and new applications continue to drive research in the area. Many statistical models exist for explaining relationships in a data set or for making predictions: cluster analysis, discriminant analysis and nonparametric regression can be used in many data mining problems [18].It is an interdisciplinary field merging ideas from statistics, machine learning, information science, visualization and other disciplines [7]. It is a very useful approach to integrate information and theory for knowledge discovery from any informatics such Bioinformatics, Chemo informatics, Nano informatics, Materials informatics and so on. The impact of DM and knowledge discovery has been evidenced by many successful research experimental results [19, 20, 21, and 22]. Therefore, it can be used to extract non-trivial, hidden, previously unknown, potential useful and ultimately understandable knowledge from massive materials databases [23.24].

1.2. Statement of the problem

Vaccination has been shown to be one of the most effective public health interventions worldwide, through which a number of serious childhood diseases have been successfully eradicated [63].

According to guidelines developed by the WHO [25], infants are considered fully vaccinated when they have received a vaccination against tuberculosis BCG, three doses each of the DPT and polio vaccines, and a measles vaccination by the age of 12 months. The DPT-HepB-Hib, introduced in 2007, has replaced the previous DPT vaccine. This new vaccine protects against DPT-HepB-Hib. Therefore, in Ethiopia, the vaccination policy calls for BCG vaccine given at birth or at first clinical contact, three doses of DPT-HepB-Hib vaccine given at approximately 6, 10, and 14 weeks of age, four doses of OPV given approximately at 0-2, 6, 10, and 14 weeks of age, and measles vaccine given at or soon after reaching 9 months of age [26]. However, many infants and children still die every year from these diseases. It has been shown that in 2007 approximately 27 million children are not vaccinated against common childhood diseases, such as measles or tetanus [25]. As a result, 2–3 million children are dying annually from easily preventable diseases and many more fall ill [25]. But in the same year, 24 million infants are not being reached with vaccines and over 10% of infants under one year old in developing countries were not receiving even one dose of DPT vaccine, compared with 2% in industrialized countries [27].

The 2011 Ethiopian Demographic and Health Survey (EDHS) report has showed that markedly increment in vaccination coverage over the past ten years [26]. The percentage of children aged 12-23 months who were fully vaccinated at the time of the survey increased from 14 % in 2000 to 20 % in 2005 and 24 % in 2011. Consequently, there was a 70 % increase over ten years and a 19 % increase in the five years before 2011 survey [26]. The percentage who had received none of the six basic vaccinations increased from 17 % to 24 % between 2000 and 2005 and then decreased to 15 % in 2011.With the exception of polio 3, the percentage of children who received all the other vaccinations has increased in the five years before the 2011 survey [26].

Accordingly, infants are more likely to be vaccinated the first doses of vaccination than the third and the fourth doses in which 60% of children received BCG and from these only 35% of them

receive measles vaccine which is the last vaccine dose in EPI program of Ethiopian [19]. This shows that there is drop out of from vaccination.

According to the Ethiopia National Immunization Coverage 2012, the weighted Ethiopian EPI coverage by antigen is BCG 79.6%; DPT-HepB-Hib1 80.0%; OPV1 90.1%; adjusted DPT-HepB-Hib3 65.7%; OPV3 70.5%; and measles 68.2%.Coverage for all antigens tends to be higher in infant of caregivers with higher educational attainment, higher wealth, infant of first parity, and those residing in urban areas. This survey revealed high drop-out rates in the immunization program. The total unadjusted drop-out rate (card, verification and history) for DPT-HepB-Hib3 was 25.6% nationally, ranging from 2.6% (Addis Ababa) to 63.8% (Somali) [11]. However, evidences show that Ethiopia has made notable progress in routine immunization coverage with an increase in DPT3 coverage from 52% in 2003 to 69% in 2005 [11].

To be fully protected against vaccine preventable diseases, infants shall have several contacts with the health care system in their first year of life. Certain social and cultural practices may increase the risk preventable by immunization, adversely affecting immunization status. Therefore, the aim of this study is predicting infant's immunization status that enables us to determine factors for dropout.

To this end, this study attempts to explore and answer the following research questions:

- What are the determinant attributes that contribute to the minimal status of infant immunization in each region of Ethiopia?
- Which mining algorithm produces best infant immunization status prediction model?

1.3. Objective of the study

1.3.1. General objective

The general objective of the research is to construct a predictive model using data mining technology that helps to predict the infant immunization status in Ethiopia.

1.3.2. Specific Objectives

The specific objectives of the study are:

- To identify determinant attributes that affects the status of infant immunization in all regions of Ethiopia
- To prepare data for analysis and model building by cleaning, extracting, and transforming it into a format suitable for the data mining algorithm.
- To apply classification algorithms to train, test and build the classifier models
- To evaluate the performance of the predictor model base on which the best model is selected for the prediction of the status of infant immunization
- To develop a graphically user friendly prototype system (interface) of the model to ease usage of the predictor model by domain users.

1.4. Significance of the study

The findings of this research will be used to predict immunization status of infants who attend EPI programs at birth, 6, 10, 14 weeks and 9 months of vaccination. It contributes a great deal of benefit to infants, health professionals, policy makers, programmers and researchers.

Accordingly, the study has the following significances:

- It helps infants to be fully vaccinated in order to protect them preventable diseases
- It helps health professionals to use the model to predict routinely immunized infant from those who are not and hence, it eradicates complication which might happen due to lack of awareness for the routinely immunized ones during the right time.
- Policy makers and programmers can make use of the model to develop new guidelines and policies and/or modify the existing ones in order to improve achievement of EPI programs goals in the country.

- Can also serve as a baseline data reference and initiative other researchers to conduct further studies in the future.

1.5. Scope and Limitation of the Study

The scope of the study is evaluating the potential of data mining technique and predicting the status of infant immunization using EDHS dataset in 2011. Study participants were children who were aged 1-5 year at the time of survey who started vaccination. The socioeconomic and demographic characteristics of mothers are the factors that were considered in this study to develop the model since they are the most prominent reasons causing difference in infant's immunization.

In this study classification algorithms of data mining are considered to create a prediction model so as to know immunization status of an infant. Due to missing values in the dataset on DPT-heB and DPT-hib because of shortage of full data in the dataset, this study is limited on the six vaccines. Lack of relevant related literature on data mining in the area was one of the limitations encountered for experience sharing and comparing the result of this study to undertake the study. Respondents Lack of awareness on infant vaccination.

1.6. Methodology

1.6.1. Research Design

In order to achieve the above stated objectives, the researcher as per the discussion made on model evaluation of DM methodologies in literature review *section 2.3*, Hybrid-DM process model was selected as a better model regarding its design to suit for academic researches. To understand a model that produces best possible classifier of immunization status of infant, a Hybrid DM process model was used. Hybrid-DM model is an intersection of KDD and CRISP-DM models [15].

The Hybrid DM presented in figure 1.1 consists of six-step Knowledge Discovery Process; i.e. understanding the problem domain, understanding data, preparation of data, Data mining, evaluation of the discovered knowledge and use of the discovered knowledge.

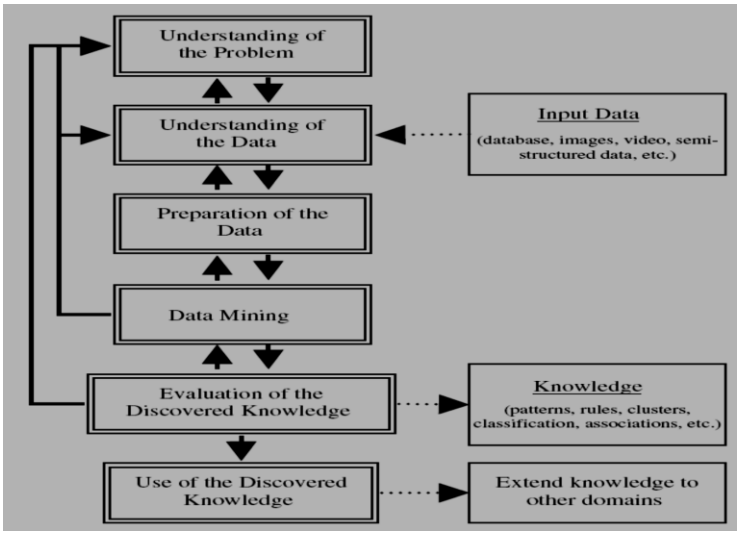


FIGURE 1.1: Hybrid-DM Process Models [79]

1.6.1.1 Understanding of the problem

A model was needed to predict which region most probable would have dropout rate the EPI preventive processes before it causes additional harm to the society. In Ethiopia vaccine preventable communicable disease are major health problem ensure all infant had access to routinely recommended vaccines that is BCG, three doses of DTP, three doses of Polio and measles vaccination by the age of 12 months. To understand the problem domain of infant immunization status in Ethiopia, the researcher used secondary data, the understanding of the detail the problem of infant immunization drop out in EDHS manual report and also discussion with domain expert from Government hospital and Health center into health professionals discuss the drop out of infant immunization.

Three health professionals from Yikatite hospital, Black lion hospital and kasanche Health Centers have been discuss the existing EPI problems. The researcher had 1 hour discussion with each of the professionals regarding the problem of the existing problem of EPI.

When we come to the problem understanding of the discussion held with the domain area expert, one of the health professionals explained her view as follow:

“In this health center do not have methods to control if the child is fully or partially immunized. Previously there is no system of communication with the family to remind about the continuation of the immunization program and a number of infants missed the vaccine, which consequently prevents the infant from fully immunized. Recently, before one year in the health center, manual registration form was prepared to follow and remind the family about their infant status of immunization. For the child who comes for the first time to the health center, the nurse registers and provides the vaccine along with having family’s full address including their telephone number. If the family did not come to the program, the responsible person in the health center will call and remind the family to come and to have the infant get vaccine, with this method the remaining vaccination problem will be minimized.”

The second health professional explained her ideas saying “In this reference hospital (Black lion hospital) there in no control system of the dropout of immunization infant and not having family communication system for the routine of infant immunization indicated so far in the health center

still exists in the hospital level that contributes to the partial immunization of the infant. Due to this reason the health personnel will rely on the card that is carried by the family member or the report given by them for infant vaccination whenever they come to the hospital. If the family member did not bring the vaccination card or did not remember the previous vaccine time, it will be difficult to the health personnel to give the right vaccine at the right time and forced give the vaccine by guess". The third health professional explained her ideas saying "there is no system to control the status of infant vaccination that delivered in the hospital level, once referred to the nearest health center after getting the firs vaccination there in the hospital. The only time that the hospital gets the infants completed card is when the family brings it for the confirmation of successfully completed of the vaccination from the health center and the hospital will stamp on the card for successful completion of the immunization."

In general, an attempt has been made to understand the problem of the infant immunization dropout at one Health center (Kasanchese) and two Government Hospitals (Yekatite hospital and Black lion hospital) with the health professionals. Then the researcher understands the problem of infant immunization dropout problem on the discussion of domain experts.

To achieve the objective of the study 2011 EDHS vaccination data was used. The 2011 EDHS is part of the worldwide monitoring and evaluation to assess and use results (MEASURE) DHS project which is funded by the United States Agency for International Development (USAID). The survey which was funded by the HIV/AIDS Prevention and Control Office (HAPCO), USAID, the United Nations Population Fund (UNFPA), the United Kingdom for International Development (DFID), the UNICEF and the Centers for Disease Control and Prevention (CDC), was implemented by the Ethiopian Central Statistical Agency (CSA) and the ICF International provided technical assistance through the MEASURE DHS project. In the survey, the authors reflected their own personal opinions rather than views of USAID. For this particular study child of 1-5 year age dataset collected on vaccination coverage of EDHS 2011 was used. Finally the researcher translated the goals in to data mining objectives and initial selection of data to be used was performed in the next step.

1.6.1.2. Understanding of the data

High quality data is a prerequisite for and data mining technique. The source of data for this study is 2011 EDHS dataset available from CSA or http://www.measuredhs.com.

Descriptive summarization and visualizations of data conducted using statistical software of SPSS is an acronym for Statistical Package for Social Science. This application software was used to create a database for infant immunization dataset.

Thus, for this research, total amounts of 8210 dataset utilized. The datasets for this study have a scale of measurement of nominal ten (10 attributes), categorical two (2 attributes) and one dependent categorical attribute. This datasets partitioned and used for training the model and testing the model accuracy.

1.6.1.3. Preparation of the data

This step is one of the key steps to produce dataset using for modeling by Waikato Environment for Knowledge Analysis (WEKA) software. It is concerned with deciding which data should be used as an input for DM methods in the subsequent step. It involve data cleaning, which includes checking the completeness of data records, removing or correcting for noise and missing values, outliers and to verify the importance of attributes . The cleaned data may be further processed by feature selection and extraction algorithms (to reduce dimensionality) and by summarization of data. The final results to be used in the study are data that meet the specific input requirements for the DM tools selected in Step 1.

1.6. 1.4. Data Mining

To build a predictive model from the cleaned data, WEKA 3.6.10 DM software was used. WEKA is a tool containing numerous machine learning algorithms that can be applied to achieve the objective of this research. Decision tree, support vector machine, rule induction and artificial neural network algorithm were used among acceptance of recent research [76, 79]. For this reason the researcher has employed four classification DM algorithms to develop and compare in section 3, the classification models. Accordingly, among the available algorithms in WEKA machine learning software; J48, PART, MLP, and SMO algorithms were applied on the EDHS 2011 data to come up with the predictive model for predicting the infant immunization status.

1.6.1.5. Evaluation of the discovered knowledge

The result of knowledge discovery process was evaluated to reach at a certain conclusion which is relevant to the problem at hand. Evaluation includes understanding the results, checking whether the discovered knowledge is novel and interesting, interpretation of the results by

domain experts, and checking the impact of discovered knowledge. The researcher used accuracy, sensitivity, specificity and precision to evaluate the performance of each of the models.

1.6.1.6. Use of the discovered knowledge

It is the final stage which consists of planning where and how to use the discovered knowledge. The application area in the current domain may be extended to other domains. A plan to monitor the implementation of the discovered knowledge is created and the entire project has been documented. The results of the study either in soft or hard copy will be disseminated to the following stakeholders and to any interested parties.

- It will be presented to the schools of information science and public health of Addis Ababa University.
- A hardcopy document will be available in the school of information science and public health libraries.
- Maximum effort will be exerted to publish the result on different journals to initiate other interested groups to find gaps and do more research in the area.

1.7. Organization of the Thesis

This research report is organized in six chapters. The first chapter deals with the general overview of the study including, background of the study, Statement of the problem, research objectives, significance of the research, scope and limitations of the research, and thesis organization.

The second chapter focuses on literature review on data mining technology, methodologies in data mining, evaluation of data mining methodologies, data mining functionalities, data mining, statics and machine learning and health care and data mining, technical definition of immunization, coverage of immunization globally and in Ethiopia, role of vaccines and also extensive review of related works are included.

Chapter Three deals with the data mining algorithms decision tree, rule induction, support vector machines and artificial neural network and research methodology incorporated to guide the research work.

The fourth chapter is about data preparation which constitutes of understanding of the problem, understanding of the data and preparation of the data. Therefore, at this stage of modeling quality data is made ready for the next chapter with classification algorithms.

In chapter five, points related to DM model selection, attribute selection, rules from decision making and prototype development are discussed. Results of the experiments are also analyzed, interpreted and presented.

Chapter six is the final chapter which presents concluding remarks and recommendations of the study.

CHAPTER TWO

LITERATURE REVIWE

2.1. Factors Associated With Immunization Status in Ethiopia

A study conducted by Belachew Etana [70], which is targeted on finding factors affecting immunization status of infants aged 12-23 month in Ambo Woreda, West Shewa Zone of Oromia Regional State revealed the poor vaccination coverage in the area. The researcher undertook a cross-sectional community based study during January to February, 2011 in Ambo Woreda in West Shewa Zone of Oromia regional state using modified WHO EPI cluster sampling method. A total of 536 children of aged between 12-23 months from 536 households were selected as sample population from 8 rural and 2 urban Kebeles.

The finding of this study indicated that about 96% of mothers have information about vaccination and vaccine preventable diseases and 79.5% knew correctly the benefit of immunization. About 36% of infant were fully vaccinated by card plus recall, but only 27.7% were fully vaccinated by card alone and 23.7% infant were unvaccinated. The study revealed that infant are more likely to be vaccinated if the child is male (Adjusted Odds Ratio [AOR]=1.8: 95% CI: 1.1, 3.1), health institution born (AOR=2.3, 95% CI, mothers' followed ANC (AOR=2.4 95% CI: 1.2, 5) and mothers' knew the correct age at which begins (AOR=2.5 95% CI: 1.3, 4.7) and finishes (AOR=2.6 95% CI: 1.8, 5.7) the immunization. Similarly, infant whose mothers attended ANC (AOR=2.1 95% CI: 1.03, 4), infant born in the health facility (AOR=2.1, 95% CI: 1.3, 3.5), infant whose mothers knew the age at which the vaccination begins (AOR= 2.4 95% CI: 1.5, 4) and completes (AOR=5.4, 95% CI: 2.6, 10) were significant predictors of full vaccination among infant aged 12-23 months.

Another Study conducted by Teshome Teklu [73] targeted on finding Assessment of Quality of Service delivery in immunization in western Gojam, Amhara Regional State in Ethiopia. The study used a cross-sectional health facility based study to assess the quality of immunization service delivery using qualitative and quantitative methods. It was conducted from March to May 1997 in 10 districts of western Gojam zone. All hospitals and health centers in the zone and randomly selected health stations were included in the study. Exit interview, observations of client provider interaction, document review, inventor of equipment, interview with service

providers, and focus group discussions with service providers and mothers were the techniques utilized.

According to the study, more than 98% satisfaction rate was reported by clients of the service during the exit interview in the dichotomous scale. But the finding was not consistent with the FGDs and the specific items addressed to assess the satisfaction. Therefore, it is concluded that quality of service delivery in immunization was not satisfactory and hence need improvement to have effect on coverage and mortality a morbidity reduction targets.

Another study done by Henok Tadesse, Amare Deribew and Mirkuzier Woldie [92], targeted on finding explorative assessment of factors affecting child immunization Wonago district, Gedeo Zone, South Ethiopia. The study used a cross-sectional qualitative study employed both focus group discussions and in-depth interviews using focus group discussion and in-depth interview guides.a total of 6 focus group discussions and 22 in-depth interviews were carried out with community representatives in wongo district.Thematic analysis of transcripts of focus group discussions and in-depth interviews was made and the data were transcribed verbatim. Also, overall interpretation was performed by relating thematic areas to each another and explaining how the various concepts related to the study questions.

According to the study, most of the in-depth interviewees and focus group discussants were knowledgeable regarding vaccines and vaccination. However, a few of the mothers and traditional healers do not know about vaccines, its benefits and side effects. Some of the focus group discussants explained that fear of side effects of vaccines could prevent mothers from having their children vaccinated. Most of the interviewees of in-depth interviews and focus group discussants raised different obstacles to vaccination in children. Of these, lack of awareness about immunization, misunderstanding of side effects, absence of electricity in the presence of refrigerators, few immunization sites on an outreach basis, misunderstanding of health extension services (more preventive services though the community demand is curative) and seasonal events especially during the time of coffee collection.

Another study done by Isaac Badu [74], the immunization status of 193 infant (aged 12-24 months) and 193 mothers (aged 19-45 years) were indicated by interviewing mothers who were selected by employing cluster probability sampling technique. A review of immunization history

was done for confirmation from immunization cards of infant. A focus group discussion with the health service providers was conducted to find out the preferred immunization strategy with special emphasis on ranking/scoring all the approved strategies.

Overall, default rate for the entire recommended series of vaccines was 15.0% and a coverage level of 92.7 % for infant immunization. Further, partially immunized infant were 14.5% and fully immunized stood at 85.5 %. The major reason for immunization failure for infants, according to the study, was obstacles and prominent amongst the reasons being postponement of immunization session(s) until another time.

2.2. Overview of Data Mining

Data mining which is considered relatively recently developed methodology and technology, coming into prominence in 1994 , is an iterative process with in which progress is defined by discovery, through either automatic or manual methods [30]. It may be defined as the exploration and analysis, by automatic or semiautomatic means of large quantities of data in order to discover meaningful patterns and rules.

DM is the search for new, valuable, and nontrivial information in large volumes of data. It is a cooperative effort of humans and computers and has mostly been used by statisticians, data analysts, and the MIS communities [31]. Best results of DM are achieved by balancing the knowledge of human experts in describing problems and goals with the search capabilities of computers [30].

Finding useful patterns in data has been given a variety of names, including DM, knowledge extraction, information discovery, information gathering, data archaeology, and data pattern processing. Hence, it may be considered as mining knowledge from large amounts of data since it involves knowledge extraction, as well as data/pattern analysis [22].

DM techniques can be broadly classified based on what they can do, namely description and visualization; association and clustering; and classification and estimation, which is predictive modeling.

Description and visualization data can contribute greatly towards understanding a data set, particularly a large one, and detecting hidden patterns in data, especially complicated data containing complex and nonlinear interactions.

In association, besides, the aim is to decide which variables go together [32] For example, market-basket analysis (the most popular form of association analysis) refers to a process that generates probabilistic statements such as, "If patients undergo treatment A, there is a 0.35 probability that they will exhibit symptom Z" [33]. With clustering, the objective is to group objects, such as patients, in such a way that objects belonging to the same cluster are similar and objects belonging to different clusters are dissimilar. In Koh and Leong, [34] Clustering is used to group readmitted patients to better profile and understand such patients.

The most common and important applications in DM probably involve predictive modeling. Classification refers to the prediction of a target variable that is categorical in nature, such as predicting healthcare fraud vs. non fraud.[35] Estimation, on the other hand, refers to the prediction of a target variable that is metric (i.e., interval or ratio) in nature, such as predicting the length of stay or the amount of resource utilization. For predictive modeling, the DM techniques commonly used include traditional statistics, such as multiple discriminate analysis and logistic regression analysis. They also include non-traditional methods developed in the areas of artificial intelligence and machine learning [36].The two for the most part major models of these are neural networks and decision trees.

2.3. Methodologies in Data Mining

Recently, the growth and consolidation of the DM area has been coming into existence. So far, efforts that seek the establishment of standards for DM has been made in the area by both academics and by people in the industry field. The academics efforts are centered in the attempt to formulate a general framework for DM [38]. The bulk of these efforts are centered in the definition of a language for DM that can be accepted as a standard, in the same way that Structured Query Language (SQL) was accepted as a standard for relational databases [39, 40, 41, 42, 43]. The efforts that have been made in the industrial field concern mainly in the definition of processes or methodologies that can guide the implementation of DM applications. To formulate meaningful patterns, rules and other related discoveries from some given attributes, data mining is incredibly important. To implement DM applications, one needs to select the best modeling technique. Therefore, in this study, relative comparison of, KDD, SEMMA, CRISP-DM and Hybrid-DM are made to select the most suitable modeling for this study.

2.3.1. The Knowledge Discovery in Database process

Knowledge Discovery in Database (KDD) is the nontrivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data [44]. The KDD process, as presented in Fayyad et al [45, 46] is the process of using DM methods to extract what is deemed knowledge according to the specification of measures and thresholds, using a database along with any required preprocessing, sub sampling, and transformation of the database. The KDD process may consist of the following steps: data selection, data processing, data transformation, DM, finding interpretation, and finding evaluation. The five stages of KDD, which are taken from Fayyad et al [45,46] that clearly indicate the processes involved in knowledge data discovery are presented in figure 2 .1 below.

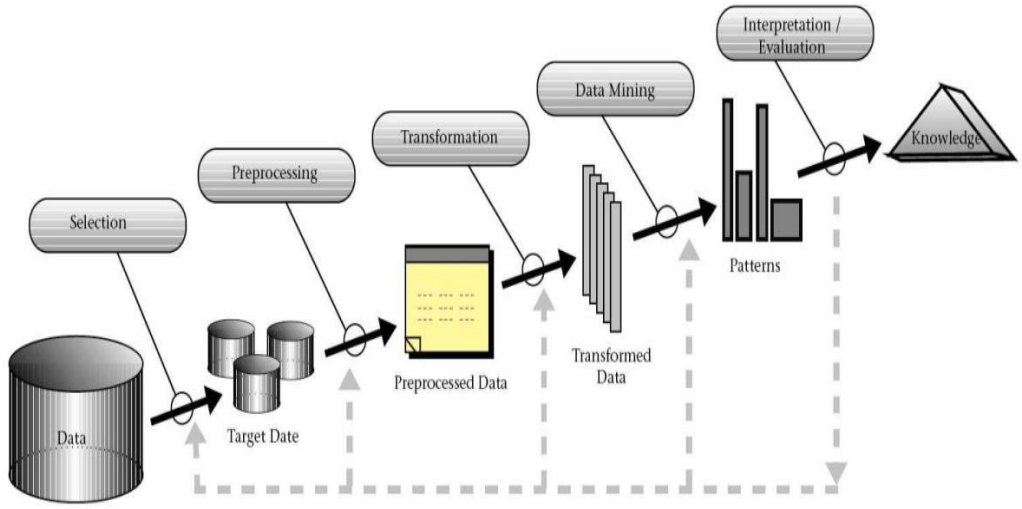


FIGURE2.1: KDD- Process Model

The KDD process is interactive and iterative, involving numerous steps with many decisions being made by the user [47].As figure 2.1 shows several steps from selecting data to providing understandable knowledge to the user are involved in the KDD process. According to Fayyad et al [45, 46] the details of the explanation on the KDD process model is explained in the following steps.

1. Selection: - Before directly going to selection one needs to understand the domain and identify the goal of applying the procedure. Then, select the required data to solve the problem. The main

part of the KDD process is the selection of raw data that is necessary for the discovery. There might be unnecessary data attributes provided, but only few data attributes are needed in the process. Selecting the necessary data attributes for the uncovering places an important part which yields target dataset.

2. Data preprocessing: - Handling missing values, eliminating noise and duplicate records in the data sets in the main part of this process. Missing values in the data lead to loss of useful information, which might not result in discovering useful knowledge. Noise in data and duplicate records mislead the process in obtaining accurate knowledge. Therefore, data cleaning and the preprocessing are necessary to produce better quality results. There are various tools such as Microsoft Excel and SPSS used for these tasks.

3. Data transformation: - In this step data is transformed or consolidated into forms appropriate for DM. Data transformation involves. The first is first smoothing of data, where the noise from data is removed. Second aggregation of data is the other means of data transformation where aggregation operations are applied to data for the analysis of the data. Third generalization is also where raw data is replaced with high level concepts and finally normalization is where the attribute data are scaled to fall within a specified range; and feature Selection, where new attributes are constructed and added from the given set of attributes.

4. Data Mining: - The step that requires analysis of the main problem and decision on which models and parameters are appropriate. Depending on the model, different DM algorithms and methods are chosen that are needed for searching data patterns. DM methods are performed to achieve the goal by finding the interesting patterns in the data. Better results are obtained if the preceding steps are performed properly.

5. Data interpretation/Evaluation: - The step where the mining results are interpreted and even the process can start again from step 1 if there are any errors or for providing further accurate results. It involves the visualization of extracted patterns and results. The knowledge discovery process then takes the raw results from DM and transforms them into useful and understandable information for the users.

Using the Knowledge directly, incorporating the knowledge into another system for further action, or simply documenting it and reporting it to interested parties, checking for and resolving potential conflicts with previously believed or extracted knowledge can be part of this step.

2.3.2. Sample, Explore, Modify, Model and Assess (SEMMA)

The SEMMA process was developed by the Statistical Analysis System (SAS) Institute. The acronym SEMMA stands for Sample, Explore, Modify, Model, Assess, and refers to the process of conducting a DM project. The SAS Institute considers a cycle with five stages for the process [48].

The SEMMA analysis cycle guides the analyst through the process of exploring the data using visual and statistical techniques, transforming data to uncover the most significant predictive variables, modeling the variables to predict outcomes, and assessing the model by testing it with new data.

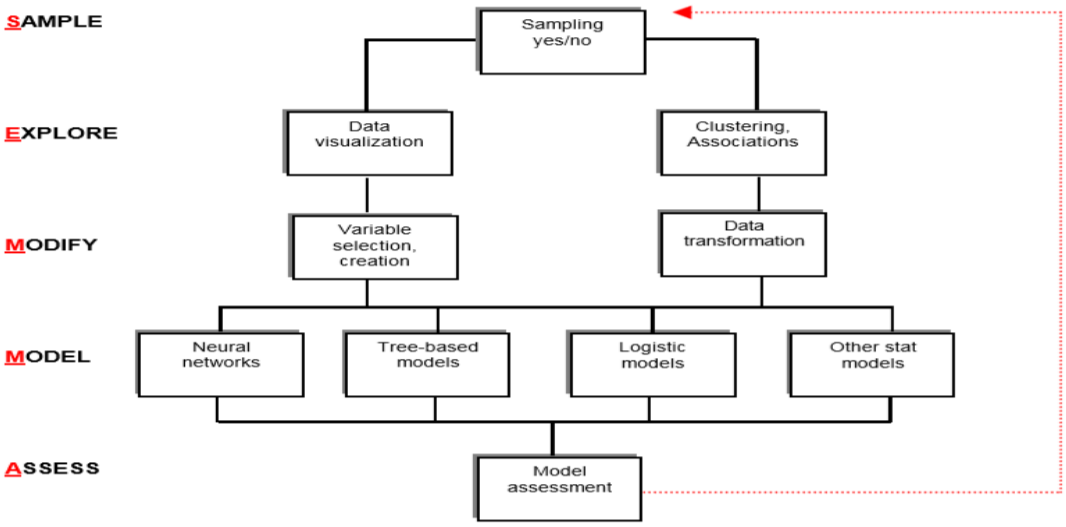


FIGURE 2.2: The SEMMA Analysis Cycle

1. Sample: the first step is to create one or more data tables by sampling data from the data warehouse. Mining a representative sample instead of the entire volume radically reduces the processing time required to obtain business information.

2. Explore: after sampling the data, the next step is explore the data visually or numerically for trends or groupings. Exploration helps to refine the discovery process and techniques such as factor analysis, correlation analysis and clustering are often used in the discovery process.

3. Modify: modifying the data refers to creating, selecting, and transforming one or more variables to focus the model selection process in a particular direction, or to modify the data for clarity or consistence.

4. Model: Creating a data model involves using the DM software to search automatically for combination of data that predicts the desired outcome reliably.

5. Assess: the last step is to assess the model to determine how well it performs. A common means of assessing a model is to set aside a portion of the data during the sampling stage. If the model is valid it should work for both the reserved sample and for the sample that was used to develop the model [48].

Although the SEMMA process is independent from DM chosen tool, it is linked to the SAS enterprise miner software and pretends to guide the user on the implementations of DM applications. SEMMA offers an easy to understand process, allowing an organized and adequate development and maintenance of DM projects. It thus confers a structure for his conception, creation and evolution, helping to present solutions to business problems as well as to find DM business goals [49].

2.3.3. Cross-Industry Standard Process for Data Mining

Cross-Industry Standard Process for DM (CRISP-DM) is a knowledge discovery approach which is widely used by industry members. This model consists of six phases intended as a cyclical process **Figure 2.3**. According to David L and Dursen D all the steps included in this model are explained in the paragraphs below [21].

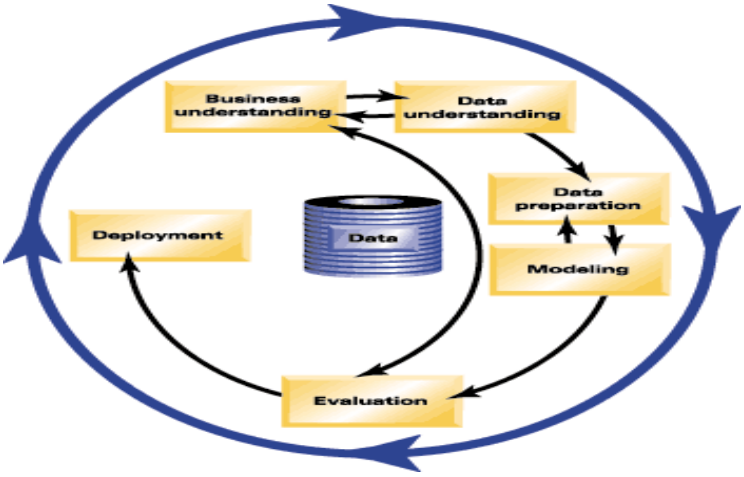


FIGURE 2.3: The CRISP-DM Model

1. **Business understanding:** This initial phase focuses on understanding the project objectives and requirements from a business perspective, then converting this knowledge into a DM problem definition and a preliminary plan designed to achieve the objectives.

2. Data understanding: This phase starts with an initial data collection and proceeds with activities in order to get familiarity with the data, to identify data quality problems, to discover first insights into the data or to detect interesting subsets to/form hypotheses for hidden information.

3. Data preparation: The data preparation phase covers all activities to construct the final dataset (data that will be fed into the modeling tool(s)) from the initial raw data. Data preparation tasks are likely to be performed multiple times and not in any prescribed order. Tasks include table, record and attribute selection as well as transformation and cleaning of data for modeling tools.

4. Modeling: In this phase, various modeling techniques are selected and applied and their parameters are calibrated to optimal values. Typically, there are several techniques for the same DM problem type. Some techniques have specific requirements on the form of data. Therefore, stepping back to the data preparation phase is often necessary.

5. Evaluation: At this stage in the project you have built a model (or models) that appear to have high quality from a data analysis perspective. Before proceeding to final deployment of the model, it is important to more thoroughly evaluate the model and review the steps executed to construct the model to be certain it properly achieves the business objectives. A key objective is to determine if there is some important business issue that has not been sufficiently considered. At the end of this phase, a decision on the use of the DM results should be reached.

6. Deployment: Creation of the model is generally not the end of the project. Even if the purpose of the model is to increase knowledge of the data, the knowledge gained will need to be **organized** and presented in a way that the customer can use it. It often involves applying "live" models within an organization's decision making processes, for example in real-time personalization of web pages or repeated scoring of marketing databases. However, depending on the requirements, the deployment phase can be as simple as generating a report or as complex as implementing a repeatable DM process across the enterprise. In many cases, it is the customer not the data analyst, who carries out the deployment steps. However, the analyst will not carry out the deployment effort it is important for the customer to understand up front what actions need to be carried out in order to actually make use of the created models [21].

2.3.4. Hybrid Data mining Process Model

Hybrid-DM model is developed by Cios et al on the CRISP-DM model by adopting it to academic research. It is a six stage process modeling which constitutes; understanding of the problem domain, understanding of the data, preparation of data, data mining and evaluation of the discovered knowledge and use of the discovered knowledge [50].

According to Cios et al the description of the six steps follows:

1. Understanding of the problem domain: - This initial step involves working closely with domain experts to define the problem and determine the project goals, identifying key people and learning about current solutions to the problem. It also involves learning domain specific terminology.

2. Understanding of the data: - This step includes collecting sample data and deciding which data, including format and size, will be needed. Background knowledge can be used to guide these efforts. Data are checked for completeness, redundancy, missing values, plausibility of attribute values, etc. Finally, the step includes verification of the usefulness of the data with respect to the DM goals.

3. Preparation of the data:-This step concerns deciding which data will be used as input for DM methods in/ the subsequent step. It involves sampling, running correlation and significance tests, and data cleaning, which includes checking the completeness of data records, removing or correcting for noise and missing values, etc. The end results are data that meet the specific input requirements for the DM tools selected in Step 1.

4. Data mining: - here the data miner uses various DM methods to drive knowledge from preprocessed data.

5. Evaluation of the discovery knowledge: - Evaluation includes understanding the results, checking whether the discovery knowledge is novel and interesting, interpretation of the results by domain experts, and checking the impact of the discovered knowledge.

6. Use of the discovered knowledge:-This final step consists of planning where and how to use the discovered knowledge. The application area in the current domain may be extended to other domains. AA plan to monitor the implementation of the discovered knowledge is created and the entire project documented. Finally, the discovered knowledge is deployed.

2.4. Evaluation of Data Mining Methodologies

The comparison among the four methodologies has done using the steps they have as a criterion, their applicability to academic researchers in common and its applicability to solve problems.

The first comparison was made by comparing KDD and SEMMA DM methodologies using the steps they have as criteria to oversee the DM task. The result of the comparison of the two methodologies (KDD and SEMMA) indicates that they are equivalent [51].

- Sample can be identified with Selection;
- Explore can be identified with Preprocessing;
- Modify can be identified with Transformation;
- Model can be identified with DM;
- Assess can be identified with Interpretation/Evaluation.

Examining it thoroughly, we may affirm that the five stages of the SEMMA process can be seen as a practical implementation of the five stages of the KDD process, since it is directly linked to the SAS Enterprise Miner software.

Comparing the KDD stages with the CRISP-DM stages is not as straightforward as in the SEMMA situation. Nevertheless, we can first of all observe that the CRISP-DM methodology incorporates the steps that, as referred above, must precede and follow the KDD process that is to say:

The Business Understanding phase can be identified with the development of an understanding of the application domain, the relevant prior knowledge and the goals of the end-user. The deployment phase can be also identified with the consolidation by incorporating this knowledge into the system.

The remaining stages are data understanding, data preparation, and modeling and evaluation phases. Data Understanding phase can be identified as the combination of Selection and Preprocessing whereas data Preparation phase can be identified with transformation. The other is modeling phase which can be identified with DM. Finally, evaluation phase can be identified with Interpretation/Evaluation as it is summarized in table 2.1 below.

Based on the above discussions, CRISP is better than KDD as well as SEMMA by including additional stages for business understanding and deploying the final outcome or discovered knowledge in to the existing system.

When we look at CRISP-DM and Hybrid-DM they are almost equivalent with the development stages contained but it can be true if there are some adjustments done for Hybrid-DM. Then comparing the two DMs, we can conclude that Hybrid-DM Provides more general, research-oriented description of the steps. They also introduced a DM step instead of the modeling step. In addition, Hybrid DM introduced several new explicit feedback mechanisms, (the CRISP-DM model has only three major feedback sources, while the hybrid model has more detailed feedback mechanisms) and in Hybrid DM, the knowledge discovered for a particular domain may be applied in other domains.

Considering the above details on both CRISP and Hybrid-DM process modeling, the one which suits this research is Hybrid-DM modeling due to its nature to be used for academic environments and the opportunity of using the discovered knowledge into other domains.

The other important reason for the selection of hybrid-DM modeling is; it is initially tested with many health related works [80, 76].The summary of data mining methodologies is presented table 2.1 below.

TABLE 2.1: Comparison data mining KDD, SEMMA, CRISP-DM and Hybrid-DM

KDD	SEMMA	CRISP-DM	Hybrid
Pre KDD	----------	Business Understanding	Understanding of the problem
Selection	Sample	Data Understanding	Understanding of the data
Preprocessing	Explore		
Transformation	Modify	Data Preparation	Preparation of the data
Data mining	Model	Modeling	Data Mining
Interpretation/Evaluation	Assessment	Evaluation	Evaluation of the discovered knowledge
Post KDD	----------	Deployment	Use of the Discovered Knowledge

2.5. Data Mining Functionalities

Data mining functionalities are used to specify the kind of patterns to be found in DM tasks. In general, DM tasks can be classified into two categories: descriptive and predictive. Descriptive mining tasks characterize the general properties of the data in the database. Predictive mining tasks perform inference on the current data in order to make predictions [30]. Different data mining functionalities, and the kinds of patterns they can discover, are described below.

2.5.1. Class Description: Characterization and Discrimination

Data can be associated with classes or concepts. It can be useful to describe individual classes and concepts in summarized, concise, and yet precise terms [30], these concepts and class descriptions can be derived using data characterization and data discrimination. Data characterization is a summarization of the general characteristics of a target class of data and data discrimination is a comparison of the general features of a target class data objects with the general features of objects from one or a set of contrasting classes [30].

2.5.2. Mining Frequent Patterns, Associations, and Correlations

Frequent patterns, as the name suggests, are patterns that occur frequently in data. There are many kinds of frequent patterns, including item sets, subsequences, and substructures [30]. A frequent item set typically refers to a set of items that frequently appear together in a transactional data set. A frequently occurring subsequence, such as the pattern that customers tend to purchase first a PC, followed by a digital camera, and then a memory card, is a (frequent) sequential pattern [30]. A substructure can refer to different structural forms, such as graphs, trees, or lattices, which may be combined with item sets or subsequences [30]. If a substructure occurs frequently, it is called a (frequent) structured pattern. Mining frequent patterns leads to the discovery of interesting associations and correlations within data.

2.5.3. Classification and Prediction

Classification is the process of finding a model or function that describes and distinguishes data classes or concepts, for the purpose of being able to use the model to predict the class of objects whose class label is unknown. The derived model is based on the analysis of a set of training data (i.e., data objects whose class label is known) [30].

The derived model can be represented in various forms, such as classification (IF-THEN) rules, decision trees, mathematical formulae, or neural networks [30]. A decision tree is a flow-chart-like tree structure, where each node denotes a test on an attribute value, each branch represents an outcome of the test, and tree leaves represent classes or class distributions [30]. Decision trees can easily be converted to classification rules. A neural network, when used for classification, is typically a collection of neuron-like processing units with weighted connections between the units [30]. There are many other methods for constructing classification models, such as naïve Bayesian classification, support vector machines, and k-nearest neighbor classification [30].

Whereas classification predicts categorical (discrete, unordered) labels, prediction models continuous-valued functions [30]. That is, it is used to predict missing or unavailable numerical data values rather than class labels. Regression analysis is a statistical methodology that is most often used for numeric prediction, although other methods exist as well.

2.5.4. Cluster Analysis

Unlike classification, in clustering class labels are not present in the training data simply because they are not known to begin with. Clustering can be used to generate such labels. The objects are clustered or grouped based on the principle of *maximizing the intra class* similarity and minimizing the interclass similarity [30]. That is, clusters of objects are formed so that objects within a cluster have high similarity in comparison to one another, but are very dissimilar to objects in other clusters. Each cluster that is formed can be viewed as a class of objects, from which rules can be derived. Clustering can also facilitate classification formation, that is, the organization of observations into a hierarchy of classes that group similar events together.

2.6. Data Mining, Statistics, Machine learning and database Systems

As a multi-disciplinary field, DM adopted its techniques from many research areas, including statistics, machine learning, database systems , neural networks, rough sets, and Visualization [50].

2.6.1. Statistical approaches

Many statistical tools including Bayesian network, regression analysis, correlation analysis, and cluster analysis have been used for DM. Usually statistical models are built from a set of training

data [50].An optimal model, based on a predefined statistical measure, is searched among the hypothesis space. Rules, patterns, and regularities are then drawn from the model.

A Bayesian network is a directed graph which represents the casual relationships among the variables, computed using the Bayesian probability theorem. Regression is the other statistical tool which is the derivation of a function which maps a set of attributes of objects to an output variable. Correlation analysis studies the correspondence of variables to each other, such as the x^2 and Cluster analysis finds groups from a set of objects based on distance measures.

2.6.2. Machine learning approaches

Like statistical methods, machine learning techniques are used to search for a best model that matches the testing data. Unlike statistical techniques, the searching space in machine learning techniques is a cognitive space of n attributes instead of a vector space of n dimensions. Besides, most machine learning methods use heuristics in the searching. The most common machine learning methods used for DM include decision tree, induction, inductive concept learning, and conceptual clustering [50].

A decision tree is a classification tree which determines an object's class by following the path from the root to a leaf node, choosing the branches according to the attribute values of the object. Decision trees are induced from the training set and classification rules can be extracted from the decision trees.

Inductive concept learning derives a brief, logical description of a concept from a set of examples.

Conceptual clustering finds groups or clusters in a set of objects, based on conceptual closeness among objects.

2.6.3. Database-oriented approaches.

Database-oriented methods do not search for a best model as the previous two kinds of methods. Instead, data model or database specific heuristics are used to exploit the characteristics of the data in hand. The attribute-oriented induction, the iterative database scanning for frequent item sets, and the attribute focusing, are representatives of the database-oriented methods [50].

- In attribute-oriented induction, primitive, low-level data are generalized into high-level concepts using conceptual hierarchies.

- The iterative database scanning method is employed to search for frequent item sets in a transactional database.
- The association rules are then derived from these frequent item sets.
- The attribute focusing method looks for patterns with unusual probabilities by adding attributes selectively in to patterns

2.7. Health care and Data Mining

Data can be a great asset to healthcare organizations, but they have to be first transformed into information .Data mining, as it is already defined in chapter two, is the analysis of (often large) observational data sets to find unsuspected relationships and to summarize the data in novel ways that are both understandable and useful to the data owner [24].

DM is not a new approach of changing dataset to a meaningful pattern; it has been used intensively and extensively by financial institutions, for credit scoring and fraud detection; marketers, for direct marketing and cross-selling or up-selling; retailers, for market segmentation and store layout; and manufacturers, for quality control and maintenance scheduling.

In addition to its importance for financial institutions, it is also used in health care centers; DM is becoming step by step more well-liked, if not ever more essential. Several factors have motivated the use of DM applications in healthcare[53].The existence of medical insurance fraud and abuse, for example, has led many healthcare insurers to attempt to reduce their losses by using DM tools to help them find and track offenders [54]. DM applications can help healthcare insurers detect fraud and abuse, and healthcare providers can gain assistance in making decisions, for example, in customer relationship management using this application [55], recently, there have been reports of successful fraud and abuse detection in healthcare centers [56].

Furthermore, the huge amounts of data generated by healthcare transactions that are too complex and voluminous to be processed and analyzed by traditional methods can be simplified by using DM technique so that decision-making can be improved by discovering patterns and trends in large amounts of complex data [57]. Such analysis has become increasingly essential as financial pressures have heightened the need for healthcare organizations to make decisions based on the analysis of clinical and financial data [58].

Another factor motivating the use of DM applications in healthcare is the realization that DM can generate information that is very useful to all parties involved in the healthcare industry. DM applications also can benefit healthcare providers, such as hospitals, clinics and physicians, and patients, for example, by identifying effective treatments and best practices [61].

Insights gained from DM can influence cost, revenue, and operating efficiency while maintaining a high level of care[59].Healthcare organizations that perform DM are better positioned to meet their long-term needs, Benko and Wilson argue [60].

There are also other factors boosting DM's popularity. For instance, as a result of the Balanced Budget Act of 1997, the Centers for Medicare and Medicaid Services must implement a prospective payment system based on classifying patients into case-mix groups, using empirical evidence that resource use within each case-mix group is relatively constant. CMS has used DM to develop a prospective payment system for inpatient rehabilitation [60].

2.8. Review of Related Works

2.8.1. Previous Attempts in Using Data Mining Techniques on Immunization Data

Adebayo et al [75] used Mathematical Model (MM) for predicting immunize-able diseases that affect infant between ages 0 - 5 years. The model was adapted and deployed for use in six selected localized areas within Osun State in Nigeria.

These DM techniques provided the means by which hidden information were discovered for detecting trends within databases, and thus facilitate the prediction of future disease occurrence in the tested locations. Results obtained showed that diseases have peak periods depending on their epdemicity, hence the need to adequately administer immunization to the right places at the right time. Therefore, this paper argues that using this model would enhance the effectiveness of routine immunization in Nigeria.

A study was conducted by Selam Assamnew [76] on predicting the occurrence of measles outbreak in Ethiopia using DM techniques for the prediction the following key attributes are used Name of the reporting health facility, Id number, Date district sent record, Date record received

at national level, name of the patient, Date of birth of the patient, Age in years of the patient, Vaccine status, whether the patient is vaccinated, unvaccinated or unknown, and Record status.

The methodology used to achieve the goal of building predictive model using DM technique for this research was a hybrid six-step Cios KDP. The required data was collected from WHO measles surveillance database covering the period 2006-2011. Naïve Bayes and decision tree DM techniques were employed to build and test the models. Models were built and tested by using a dataset of 15,631 records.

To get a better insight in choosing which model produced sound prediction and higher accuracy, 12 experiments were done with J48 algorithm and naïve Bayes classifier, by inputting all the records with a 10-fold cross-validation mode, and percentage split (70%) for training and then remaining 30% of the record for testing the performance of the model. Experimental results show that J48 decision tree algorithm register better prediction accuracy.

Hemalatha and Megale [78] briefly examine the potential use of classification based DM techniques such as decision tree, Artificial Neural Network to massive volume of Immunization data. In their study data analysis of infant with Immunization and vaccination have been used as an upstream, from protecting infant, against such infections and infectious diseases as BCG, DPG, Polio and Measles. After preliminary results were analyzed, the program projected that over three million cases deaths would be prevented and it has been resulted in a statistically significant in table survey. There is still, however, much that can be done. Through the use of DM algorithms in order to verify the improvement of quality

Up to the knowledge of the researcher, no previous researches have been done to predict the routine infant immunization status by applying DM techniques in Ethiopia. Hence, this research contributes a lot to generate patterns that help in planning a better strategy for routine infant immunization status using DM technique.

The Related Works is presented table 2.2.below.

TABLE 2.2 Related Works

SN.	Name (year)	Title	Objective	Methodology	Key finding
1	Selam Assamnew (2011)	Predicting The Occurrence Of Measles Outbreak In Ethiopia Using Data Mining Technology	To design a predictive model using data mining technology that can help predict the occurrence of measles outbreaks in Ethiopia.	Hybrid six-step Cios KDP	Results show that J48 decision tree algorithm register better prediction accuracy
2	Adebayo Peter Idowu, Bernard Ijesunor Akhigbe and Olajide Olusegugun Adeosun et.al. (2013)	Data Mining Techniques for Predicting immunize-able diseases : Nigeria as a case study	To predicting immunize-able diseases that affect infant between ages 0 - 5 years	CRISP-DM	Results obtained showed that diseases have peak periods depending on their epidemicity, hence the need to adequately administer immunization to the right places at the right time
3	Hemalatha, M., and Megala, S. (2011).	Mining Techniques In Health Care: A Survey Of Immunization	To compare immunization uptake by district to disease levels in those same areas, as immunization areas with higher disease rates may be potential targets for future efforts and to obtain with higher accuracies in their prediction capabilities.	Classification method.	Polio is opportune disease for data mining technology for a number of factors, the huge amount of data polio virus invades the central nervous system the spinal cord and the brain and may cause weakness, paralysis, serious breathing problems or death.

CHAPTER THREE

DATA MINING ALGORITHMS

There are different DM techniques presented with their appropriateness to be applied in different health care areas. This study has incorporated classification algorithms to build the prediction model. Then J48 from decision tree, Sequence Minimum Optimization (SMO) from Support Vector Machine (SVM), PART from rule induction, and MLP from artificial neural network are selected to run the experiments due to their acceptance in recent research [76, 79].

3.1. Decision Trees

Decision tree can produce a model with rules that are human-readable and interpretable. The classification task using decision tree technique can be performed without complicated computations and the technique can be used for both continuous and categorical variables. This technique is suitable for predicting categorical outcomes [80]. Decision tree classifiers are quite popular techniques because the construction of tree does not require any domain expert knowledge or parameter setting, and is proper for exploratory knowledge discovery. [80].

Decision tree is a kind of classifying and prediction DM technology, belonging to inductive learning and supervised knowledge mining technology. As decision tree is advantageous in fast construction and generating easy-to-interpret If-Then decision rule, it has become the most widely applied technique among numerous classifications [81].

Decision tree is a king of tree diagram based method, the node on the top of its tree structure is root node, nodes in the bottom are leaf nodes, and one target class attribute is given to each leaf node. From root node to every leaf node, there is a path made of multiple internal nodes with attributes. This path generates rule required for classifying unknown data. Moreover, most of decision tree algorithms contain two-stage task, i.e., tree building and tree pruning. In tree building stage, a decision tree algorithm can use its unique approach (function) to select the best attribute, so as to split training data set. The final situation of this stage will be that data contained in the split training subset belong to only one certain target class. Recursion and repetition upon attribute selecting and set splitting will fulfill the construction of decision tree root node and internal nodes. On the other hand, some special data in training data set may lead to improper branch on decision tree structure, which is called over fitting. Therefore, after

building a decision tree, it is has to be pruned to remove improper branches so as to enhance decision tree model accuracy in predicting new data [82].

Decision trees are produced by algorithms that identify various ways of splitting a data set into branch-like segments. These segments form an inverted decision tree that originates with a root node at the top of the tree. The object of analysis is reflected in this root node as a simple, one dimensional display in the decision tree interface. The name of the field of data that is the object of analysis is usually displayed, along with the spread or distribution of the values that are contained in that field. The discovery of the decision rule to form the branches or segments underneath the root node is based on a method that extracts the relationship between the object of analysis (that serves as the target field in the data) and one or more fields that serve as input fields to create the branches or segments. The values in the input field are used to estimate the likely value in the target field. The target field is also called an outcome, response, or dependent field or variable [83].

3.1.1. J48 Decision tree algorithm

J48 is an open source Java implementation of the C4.5 algorithm in the Weka DM tool. C4.5 is a program that creates a decision tree based on a set of labeled input data. J48 classifier is among the most popular and powerful decision tree classifiers [84]. J48 classifier is a simple C4.5 decision tree for classification. It creates a binary tree. The decision tree approach is most useful in classification problem. With this technique, a tree is constructed to model the classification process. Once the tree is built, it is applied to each tuple in the database and results in classification for that tuple [85, 86].

The basic algorithms for decision tree induction is a greedy algorithm which constructs decision trees in a top down approach dividing each node recursively until a leaf node is encountered. The following algorithm shows how decision tree algorithms generate a decision tree from the given training data [87].

Input:

Data partition, D, which is a set of training tuples and their associated class labels;

Attribute list, the set of candidate attributes;

Attribute selection method, a procedure to determine the splitting criterion that "best"

This criterion consists of a splitting

Attribute, possibly, and either a split point or splitting subset.

***Output**: A decision tree.*

Method:

(1) Create a node N;

(2) If tuples in D are all of the same class, C then

(3) Return N as a leaf node labeled with the class C;

(4) If attribute list is empty then

(5) Return N as a leaf node labeled with the majority class in D

(6) Apply Attribute selection method (D, attribute list) to find the "best" splitting criterion

(7) Label node N with splitting criterion;

(8) If splitting attribute is discrete-valued and multi way splits allowed then

(9) Attribute list←attribute list−splitting attribute;

(10) For each outcome j of splitting criterion

(11) Let D j be the set of data tuples in D satisfying outcome j;

(12) If D j is empty then

(13)Attach a leaf labeled with the majority class in D to node N;

(14) Else attach the node returned by Generate decision tree (D j, attribute list) to node N; End for

(15) Return N; [88.89].

J48 generates decision trees, the nodes of which evaluate the existence or significance of individual features, C4.5 algorithm uses the perception of information gain or entropy reduction to select the optimal split. Suppose that we have a variable x whose k possible values have probabilities p1, p2 and . . . .pk what is the smallest number of bits, on average per symbol,

needed to transmit a stream of symbols representing the values of X observed? The answer is called the entropy of X and is defined as

$$H(X) = -\sum_{j} P\,j \log_2 (pj)$$

For an event with probability p, the average amount of information in bits required to transmit the result is log2 (p). For example, the result of a fair coin toss, with probability 0.5, can be transmitted using log2 (0.5) = 1 bit, which is a zero or 1, depending on the result of the toss. For variables with several outcomes, we simply use a weighted sum of the log2 (p j)'s, with weights equal to the outcome probabilities, resulting in the formula

$$H(X) = -\sum_{j} P\,j \log_2 (pj)$$

C4.5 uses this concept of entropy as follows. Suppose that we have a candidate split S, which partitions the training dataset T into several subsets, T1, T2,…Tk. The mean information requirement can then be calculated as the weighted sum of the entropies for the individual subsets, as follows

$$H\,(T) = \sum_{i=1}^{k} PiHs\,(Ti)$$

Where Pi represents the proportion of records in subset i. We may then define our information gain to be gain (S) = H (T) − HS (T), that is, the increase in information produced by partitioning the training data T according to this candidate split S. At each decision node, C4.5 chooses the optimal split to be the split that has the greatest information gain, gain (S). J48 uses the same concept to construct the decision tree.

3.1.2. Rule Induction

There is an alternative approach to rule induction that avoids global optimization nevertheless produces accurate, compact rule sets. The method combines the divide-and- conquers strategy for decision tree learning with the divide-and-conquer one for rule learning. It adopts the divide-and-conquer strategy in that it builds a rule, removes the instances it covers, and continues creating rules recursively for the remaining instances until none are left [87].

3.1.2.1. PART

PART (Partial Decision Tree) is a rule induction algorithm which grabs rule from a decision tree. A partial decision tree is an ordinary decision tree that contains branches to undefined sub trees [87]. To generate such a tree, the construction and pruning operations are integrated in order to find a "stable" sub tree that can be simplified no further [87]. Once this sub tree has been found, tree building ceases and a single rule is read off. The following algorithm depicts the steps and procedures followed in implementing PART rule induction.

Initialize **E** to the instance set

For each class C, from smallest to largest

BUILD:

Split E into Growing and Pruning sets in the ratio 2:1

Repeat until (a) there are no more uncovered examples of C; or (b) the description length (DL) of rule set and examples is 64 bits greater than the smallest DL found so far, or (c) the error rate exceeds 50%:

GROW phase: Grow a rule by greedily adding conditions until the rule is 100% accurate by testing every possible value of each attribute and selecting the condition with greatest information gain G

PRUNE phase: Prune conditions in last-to-first order. Continue as long as the worth W of the rule increases

OPTIMIZE:

GENERATE VARIANTS:

For each rule R for class C,

Split E afresh into Growing and Pruning sets

Remove all instances from the Pruning set that are covered by other rules for C

Use **GROW** and **PRUNE** to generate and prune two competing rules from the newly split data:

R1 is a new rule, rebuilt from scratch;

R2 is generated by greedily adding antecedents to R.

Prune using the metric A (instead of W) on this reduced data

SELECT REPRESENTATIVE:

Replace R by whichever of R, R1 and R2 has the smallest DL.

MOP UP:

If there are residual uncovered instances of class C, return to the BUILD stage to generate more rules based on these

CLEAN UP:

Calculte DL for the whole rule set and for the rule set with each rule in turn omitted; delete any rule that increases the DL

Remove instances covered by the rules just generated

Continue

3.1.3. Support Vector Machine

In today's machine learning applications, SVM [68] are considered a must try it offers one of the most robust and accurate methods among all well-known algorithms. SVM have been promising methods for data classification and regression [91, 92].Their success in practice is drawn by its solid mathematical foundations which convey the following two salient properties:

Margin maximization: The classification boundary functions of SVMs maximize the margin, which in machine learning theory, corresponds to maximizing the generalization performance given a set of training data.

Nonlinear transformation of the feature space using the kernel trick: SVMs handle a nonlinear classification efficiently using the kernel trick, which implicitly transforms the input space into another high dimensional feature space.

In a two-class learning task, the aim of SVM is to find the best classification function to distinguish between members of the two classes in the training data. The metric for the concept of the "best" classification function can be realized geometrically. For a linearly separable dataset, a linear classification function corresponds to a separating hyper plane **f(x)** that passes through the middle of the two classes, separating the two. Once this function is determined, new data instance $\mathbf{x_n}$ can be classified by simply testing the sign of the function $\mathbf{f(x_n)}$; $\mathbf{x_n}$ belongs to the positive class if $\mathbf{f(x_n)>0}$.

Because there are many such linear hyper planes, what SVM additionally guarantee is that the best such function is found by maximizing the margin between the two classes. Intuitively, the margin is defined as the amount of space, or separation between the two classes as defined by the hyper plane. Geometrically, the margin corresponds to the shortest distance between the closest data points to a point on the hyper plane. Having this geometric definition allows us to explore how to maximize the margin, so that even though there are an infinite number of hyper planes, only a few qualify as the solution to SVM.

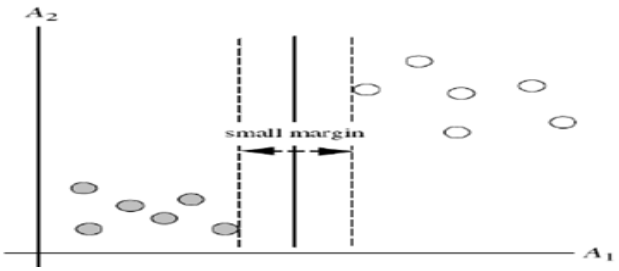


FIGURE 3.1: Linearly Separable 2D Training Data

SVM insists on finding the maximum margin hyper planes are that it offers the best generalization ability. It allows not only the best classification performance (e.g., accuracy) on the training data, but also leaves much room for the correct classification of the future data. To ensure that the maximum margin hyper planes are actually found, an SVM classifier attempts to maximize the following function with respect to w and b:

$$LP = 1/2\ w - \sum_{i=1}^{t} \alpha_i y_i (w.x_i + b) + \sum_{i=1}^{t} \alpha_i$$

Since 1992 SVM were largely unnoticed due to widespread belief in the statistical and/or machine learning community, despite being theoretically appealing. A due attention towards

SVM has come in to prominence when excellent results achieved in numeral recognition, text categorization and computer vision; today SVM show better results than neural network and other statistical models [67].

The success of SVMs in machine learning naturally leads to its possible extension to the classification or regression problems for mining a huge amount of data. However, despite the prominent properties of SVMs, they are not as favored for large-scale DM as for pattern recognition or machine learning because the training complexity of SVMs is highly dependent on the size of dataset. Many real-world DM applications involve millions or billions of data records. [52].

3.1.3.1. Sequential Minimal Optimization

Sequential Minimal Optimization algorithm, due to John Platt's, gives an efficient way of solving the dual problem arising from the derivation of the SVM. This implementation globally replaces all missing values and transforms nominal attributes into binary ones [87]. It also normalizes all attributes by default. In that case the coefficients in the output are based on the normalized data, not the original data, this is important for interpreting the classifier [87]

3.1.4. Artificial Neural Network

An Artificial Neural Network (ANN) commonly referred as neural network is an information processing paradigm that is inspired by the way biological nervous systems such as the brain, process information. Structure of Neural network is composed of a large number of highly interconnected processing elements called as neurons. This structure basically consist of inputs, which are multiplied by weights (strength of the respective signals), and then computed by a mathematical function which determines the activation of the neuron and one more function computes the output of the artificial neuron (sometimes in dependence of a certain threshold). Training or learning is the process of adjusting weights on input connections to get the specific desired output. Resemblance of ANN with biological neural network gives opportunity to user to apply parallel processing concept on each neuron at each layer [66].

3.1.4.1. A Multi-layer Feed-Forward Neural Network

The back propagation algorithm performs learning on a *multilayer feed-forward* neural network. It iteratively learns a set of weights for prediction of the class label of tuples. A multilayer feed-forward neural network consists of an *input layer*, one or more *hidden layers*, and an *output layer*. An example of a multilayer feed-forward network is shown in figure 3.2 [65].

Each layer is made up of units. The inputs to the network correspond to the attributes measured for each training tuple. The inputs are fed simultaneously into the units making up the **input layer**. These inputs pass through the input layer and are then weighted and fed simultaneously to a second layer of "neuron like" units, known as a hidden layer. The outputs of the hidden layer units can be input to another hidden layer, and so on [65]. The number of hidden layers is arbitrary, although in practice, usually only one is used. The weighted outputs of the last hidden layer are input to units making up the output layer, which emits the network's prediction for given tuples [65]. The network is feed-forward in that none of the weights cycles back to an input unit or to an output unit of a previous layer. It is fully connected in that each unit provides input to each unit in the next forward layer [65].

The steps for computing the output of a single neuron are as follows:

1. Compute the weighted sum of inputs to the neuron and add the bias to the sum

$$(Ij) = \sum_i W_{ij} O_i + \theta_j$$

Where Wij is the weight of the connection from unit i in the previous layer to unit j; Oi is the output of unit i from the previous layer, θ j is the bias of unit which acts as a threshold.

2. Each unit in the hidden and output layers takes its net input and then applies an activation function [25]. The output of the activation function is defined to be the output of the neuron.

$$O\,j = \frac{1}{1+e^{-1j}}$$

This function is called logistic or sigmoid function or also is referred to as a squashing function, because it maps a large input domain onto the smaller range of 0 to 1 [25].

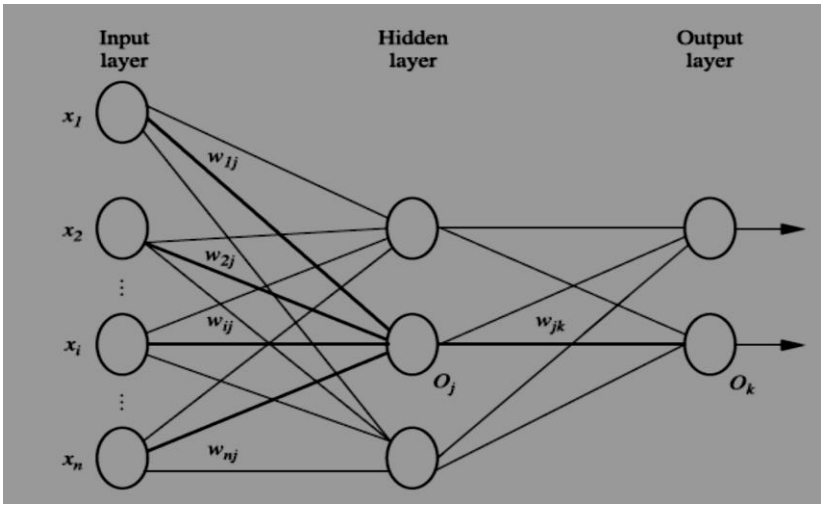


FIGURE 3.2: A multilayer feed-forward neural network.

3.2. Performance Evaluation for Predictive Modeling

Once a predictive model is developed using the EDHS 2011 data, the model should be checked as to how it will accomplish for the coming data that it has not seen during the model building process. The researcher has used different classifiers to build the predictive model and in order to evaluate the performance of the model, confusion matrix is used. Confusion matrix is a useful tool for analyzing how well a classifier can recognize tuples of different classes [24]. A confusion matrix a binary classification model classifies each instance into one of two classes; say a true and false class. This gives rise to four possible classifications for each instance; a true positive, a true negative, a false positive, or a false negative. This situation can be depicted as a confusion matrix (also called contingency table) given in figure 3.3.The confusion matrix juxtaposes the observed classifications for a phenomenon (columns) with the predicted classifications of a model (row) [52].

In figure 3.3, the classifications that lie along the major diagonal of the table are the correct classifications, that is, the true positives and the true negatives [37]. The other fields signify model errors. For a perfect model we would only see the true positive and true negative fields filled out, the other fields would be set to zero. It is common to call true positives hits, true negatives correct rejections, false positive false alarms, and false negatives misses. A number of model performance metrics can be derived from the confusion matrix. Perhaps, the most common metric is accuracy defined by the following formula [52].

The **accuracy** of a classifier is estimated by dividing the total correctly classified positives and negatives instance by the total number of samples.

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

Other performance metrics include precision and recall defined as follows:

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

$$\text{Precision} = TP/TP+FP$$

Recall is what percent of positive tuples the classifier labeled as positive for both True and False classes.

$$\text{Recall} = TP/\ TP + FN$$

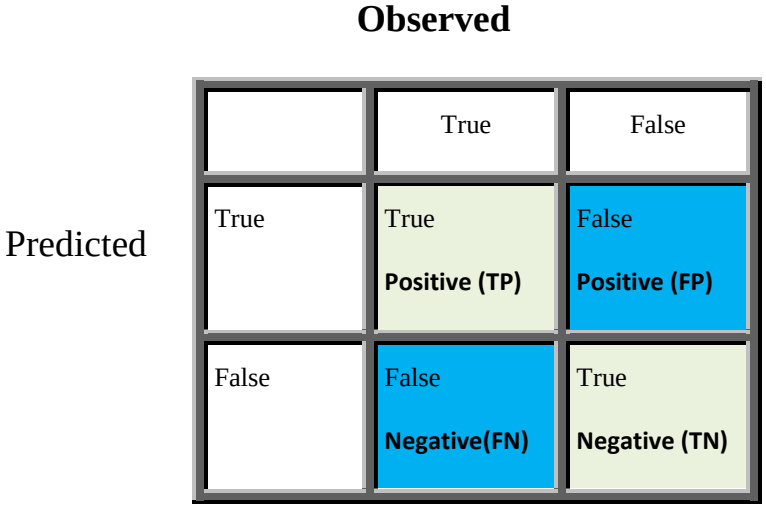


FIGURE 3.3: Confusion matrix

3.2.1. Receiver Operating Curves (ROC)

Receiver operating curves (ROC) curves visually convey the same information as the confusion matrix in a much more intuitive and robust fashion. ROC curve are two-dimensional graphs that visually depict the performance and performance trade-off of a classification model [65]. ROC curves were originally designed as tools in communication theory to visually determine optimal operating points of signal discriminators. Two new performance metrics have to be introduced

here in order to construct ROC curves (they have been defined here in terms of the confusion matrix), the true positive rate (TPR) and false positive rate (FPR): [65]

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

$$\text{False Positive Rate} = \frac{FP}{FP+TN}$$

ROC graphs are constructed by plotting the true positive rate against the false positive rate

Graph, as shown in figure 3.4, the value of FP Rate is plotted on the horizontal axis, with TP Rate plotted on the vertical axis. Each point on the graph can be written as a pair of values (x, y) indicating that the FP Rate has value x and the TP Rate has value y [65]. The performance of different types of classifier with different parameters can be compared by inspecting their ROC curves**.**

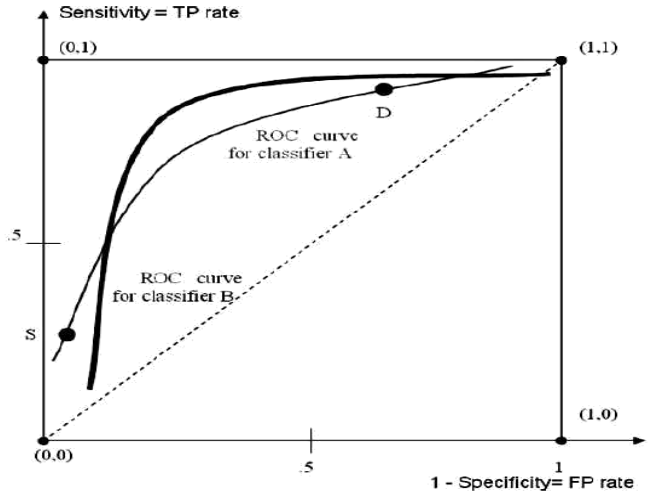


FIGURE.3.4: ROC Curve for two classifiers

In order to decide which of the two classifiers in figure .3.4 constitutes a better model/classifier of the data, visual analysis could be performed, that is the curve more to the upper left would indicate a better classifier. However, the curves often overlap, as shown in figure 3.4 for two classifiers, in this case the popular method called area under curve (AUC) is used [65]. Since the AUC is a portion of the area of the unit square, its value will always be between 0 and 1.0. This method chooses a model/classifier that has maximum area under its corresponding ROC curve: the larger the area, the better performing the model/classifier [65].

CHAPTER FOUR

DATA PREPARATION

According to, a hybrid data mining process model, before applying classification, which consists of six step algorithm these is a need to problem understanding, business understanding, data understanding and data preprocessing. There steps enable to select and prepare the data for DM tasks [30]. Data understanding and preprocessing is the most challenging part in the preparation of entire DM process as it consumes the majority of time and effort.

4.1 Business Understanding

Business understanding is a stage where the researcher is acquainted with the overall business process or working conditions of the identified work process. To this end, the researcher has made efforts to understand what the business is and it's in and out as much as possible. As it has been discussed in the domain expert, ***section 1.6.1.1***, and understands the problem of infant immunization dropout.

Business understanding is mainly concerned with the identifying business objectives and determination of data mining goals

4.1.1. Identifying Business objective

Immunization is a process of inducing immunity against a specific disease and it can be induced actively by administration of vaccine or toxoid to produce hum oral or cell mediated immune response. Vaccines are not only used in preventing disease, they are useful in the mitigation of the severity of disease, prevention of infections, prevention of cancers (for example cancer of the cervix, cancer of the liver), and reduction in the complications associated with the infections [23].

Immunization is reported to be second to clean water in reducing the burden of infectious diseases [23]. Therefore, employing immunization is one of the mechanisms by which diseases can be eradicated. When a sufficient proportion of the population is immune, there is an indirect effect on the whole population, called 'herd immunity' [23]. This causes a reduction in the spread of the infective agent by blocking its transmission from one person to another.

Since the time of discovery of smallpox vaccination in 1796 by Edward Jenner, vaccinations have a major impact on global health [71]. Vaccines may contain live attenuated organisms, inactivated or killed organisms, toxoids or subunits of antigen. It may be derived from cell culture techniques.

Vaccines are available for tuberculosis, diphtheria, measles, tetanus, HpB, poliomyelitis, Hib, pertussis, yellow fever, mumps, rubella, pneumococcal infections, rotavirus, and cholera. Therefore, immunization which encourages timely vaccination is said to be the single most efficient and cost effective means of controlling these diseases [63]. The drastic decline, and in some cases elimination, of certain infectious diseases since the introduction of vaccines in the 20^{th} century can be taken as an evidence for the importance of immunization [64].

The EPI is a program basically established to control six vaccines preventable diseases including diphtheria, polio, pertussis, tetanus, tuberculosis and measles. The EPI schedule recommends that children should be vaccinated with the following vaccines [8]. BCG vaccine at birth; three doses of DPT at 6, 10 and 14 weeks of age; at least three doses of OPV, at birth, 6, 10 and 14 weeks of age; and one dose of measles vaccine at 9 months of age [9]. Therefore, to reduce morbidity and mortality of infants, they are expected to be fully immunized by 12 months of age. From this conclusion, the business objective of the study is to minimize the morbidity and mortality rate of infants as much possible.

4.1.2. Determination of Data Mining

The Main purpose of the study is to apply DM techniques on immunization dataset to detect dropout rate or unimmunized infants. The result of study helps to improve EPI to control faire distribution of infant vaccination among different regions of Ethiopia after understanding immunization dropout rate and unimmunized infants using attributes of socio-economic and demographic information. It also contributes a lot in predicting routine infants' immunization dropout rate and unimmunized infants by identified predicting variables; determine those having a better prediction performance.

4.2 Data Understanding

Data is a basic element in DM projects. To be successful in preparing DM project, data understanding is a mandatory stage to bring about a valid result. In order to meet the general

objective of this research, secondary representative data was found to be necessity. Data understanding begins with EDHS dataset.

4.2.1 Data Source and data collection

The data source for this research is taken from 2011 EDHS dataset and the children immunization coverage data. The data was collected on vaccination coverage in two ways: first from vaccination in cards shown to the interviewer and from mothers' verbal reports. The interviewer copied the vaccination dates directly to the questionnaire if the cards were available. By the time the interviewer came across a child without vaccination card, or if a vaccine had not been recorded on the card when it was given, the respondent was asked to recall the date of the vaccines given to her child. Vaccinations reported on the vaccination card represent vaccines given by routine immunization services.

4.2.2 Description of Data

Advantageous consultations were made with domain area experts working in FMOH, EHNRI and physicians working on Internal medicine Department in Black lion hospital in order to select the last dataset or attribute to be used for the research and also, literatures conducted so far on related areas were analyzed to validate the selected attributes. According to the literature were founded attributes sex of child, parity, residence, region, mother education level, mother age, marital status, wealth status and place of delivery in the domain area expert dissection were found distance to health facility, frequency of listening to radio and father education. Therefore, twelve (12) independent variables i.e. Mother's age, Marital Status, Residence, Place of delivery, Region, Wealth status, Mother's level of education, Father's level of education, Parity, Distance to health facility, Frequency of listening to radio**,** and Sex of the child are the major variables that affect mothers of the child from using the healthcare service as indicated in literature survey and discussions made with the domain area experts. Immunization status of data was selected from the huge dataset having over 929 attributes from EDHS 2011 dataset. These variables provide the socio-economic and demographic information for each respondent. The description of the selected attributes is presented table 4.1 below.

TABLE 4.1: Selected Attributes with their Description from EDHS2011 dataset

No	Variable Name	Description	Data Types	Values
1	Age	Mother's age Group	Categorical	1.15-19, 2. 20-24, 3. 25-29, 4. 30-34, 5. 35-39, 6. 40-44, 7. 45-49
2	Marital status	Current Marital status of the mother	Nominal	1. Single, 2.Married, /Living to gather, 3.Widowed, and 4.Divorced /Separated
3	Mother Education	Mother's level of education	Nominal	0. No education, 1. primary, 2 Secondary, and 3. Higher
4	Place of delivery	Access of health facility	Nominal	1. Home, 2. Government health facility, 3. Private health facility, 4.NGO health facility, 5. Other health facility (96, 12,26 ...etc)
5	Distance from health facility	Problem for getting medical help for self:	Nominal	0.No problem, 1. Big problem, 2. and Not a big problem
6	Father Education	Father level of education	Nominal	**0.**No education, 1. Primary, 3, Secondary, and 4. Higher
7	Wealth status	The living standard of the mother	Nominal	**1.** Poor, 2. Middle and 3. Rich
8	Residence	Place of Residence	Nominal	1. Urban, 2. Rural
9	Region	The 11 administrative region of the country where live	Nominal	1. Tigray 2. Affar, 3. Amhara 4. Or Omiya 5. Somali, 6. Benishangul Gumz 7. Southern. National Nationality People (SNNP), 12. Gambela 13. Harare 14. Addis Ababa 15. Dire Dawa
10	Frequency of listening to radio	How often do a mother listen to radio	Categorical	0. Not at all, 1 less than once a week, And 2. At least once a week
11	Sex	Sex of child	Nominal	1. Male, 2. Female
12	Parity	Number of children of a mother including the child	Nominal	1, 1, 2. 2- 3. 3, 4-5, 4. 6-7 5,7-8 , 6..9-10 , 7.11-12
Dependent variable				
	Immunization Status	The extent to which the child received all vaccines or not	Categorical	Full, Partial , No

The dependent variable in this study has three classes (full, partial or not immunized) to express immunization states of the infants if the child had received all for which he/she is eligible by age, the infant is fully immunized. Otherwise if the child had received some vaccine it is partially immunized. More detail describe below.

Infant immunization status can be categorized into three [72]:

1. **Fully immunized**: When the child had received BCG, three doses DPT, and three doses of OPV and measles vaccine for which he/she was eligible by age.
2. **Partially immunized**: When the child had received some but not all vaccines for his/her age as per schedule.
3. **Not immunized**: When the child had not received any of the vaccine for the age, though eligible age.

4.3. Data Preprocessing

Data preprocessing is a DM technique that involves transforming raw data into an understandable format [28]. For the data preparation of the study, there are two aspects mentioned

1. Real-world data may be incomplete, noisy, and inconsistent, which can disguise useful Patterns. This is due to the fact that noisy data contains errors or outliers. Inconsistent data contains discrepancies in codes or names. Incomplete data also lacks attribute values, lacks certain attributes of interest, or contains only aggregate data. Data preparation generates a dataset smaller than the original one, which can significantly improve the efficiency of DM.

This task includes: Selecting relevant data which can be carried out by attribute selection (filtering and wrapper methods) removing anomalies, or eliminating duplicate records and reducing data which incorporates sampling or instance selection.

2. Data preparation generates quality data, which leads to quality patterns. For example, we can recover incomplete data by filling the values missed, or by reducing ambiguity. And we can purify data, correct errors, or removes outliers (unusual or exceptional values). The other is resolving data conflicts using domain knowledge or expert decision to settle discrepancy

While data-mining technology can support the data-analysis applications in using to identify infant immunization rate**,** it must be possible to prepare quality data from the raw data to enable efficient and quality knowledge discovery from the data given [17].

Different data preprocessing tasks were involved in this research including, selecting relevant data which can be done by attribute selection, reducing data, sampling or instance selection and, data cleaning, fill in missing values. The other data processing task is carried out by data transformation such as normalization and aggregation. Data discretization which is part of data reduction, replaces numerical attributes with nominal ones is also a data processing task. Data reduction is also another important task which is meant to reduce the volume but producing the same or similar analytical results.

The general aim of data preprocessing is mandatory to obtain quality data and expose as much information as possible for data modeling by the DM algorithms.

4.3.1. Exploratory Data Analysis

Descriptive data summarization techniques can be used to classify the representative properties of data and highlight which data values should be treated as noise or outliers [17]. Moreover, missing values can easily be known through this system which in turn assists data preparation.

To understand the nature of the data values in the selected EDHS 2011 dataset. An exploratory data analysis was done by categorical attributes. Categorical and nominal attributes were explored to expose the valid and missing instances and the percentage distribution of each data instances.

Three (3) categorical variables and ten nominal variables were used to reach at the final goal of the study. The frequency of data values in terms of number and percentage with the number of valid and missing values is presented in the table below.

Table.4.2. show below frequency distribution of region in the dataset

TABLE: 4. 2 Frequency distribution of Region

Attribute Name		Variables	Frequency	Percent (%)
Region	Valid	Tigray	891	10.9
		Affar	798	9.7
		Amhara	954	11.6
		Oromiya	1211	14.8
		Somali	683	8.3
		Benishangul-Gumuz	706	8.6
		SNNP	1135	13.8
		Gambela	574	7.0
		Harari	473	5.8
		Addis Ababa	284	3.5
		Dire Dawa	501	6.1
		Total	8210	100.0
	Missing		-	-
	Total		**8210**	**100.0**

Region: It is a nominal valued attribute so there was no found missing value in the dataset.

Table.4.3. Frequency distribution of Residence

Attribute Name		Variables	Frequency	Percent (%)
Residence	Valid	Urban	1381	16.8
		Rural	6829	83.2
		Total	8210	100.0
	Missing		-	-
	Total		**8210**	**100.0**

Residence: It is a nominal valued attribute and missing value in the table 4.3 was not found.

The above table shows that frequency distribution of residence in the dataset

Table 4.4 Frequency distribution of Parity

Attribute Name		Variables	Frequency	Percent (%)
Parity	Valid	1	1576	19.2
		2-3	2626	32.0
		4-5	1899	23.1
		6-7	1278	15.6
		8-9	565	6.9
		10-11	217	2.6
		12-13	45	.5
		Total	8210	100.0
	Missing		**8210**	**100.0**
	Total		**8210**	**100.0**

Parity:-It is categorical valued attribute there was no missing value in the dataset found.

Table 4.5 Frequency distribution of Sex of child

Attribute Name		Variables	Frequency	Percent (%)
Sex of child	Valid	Male	4193	51.1
		Female	4017	48.9
		Total	8210	100.0
	Missing		-	-
	Total		**8210**	**100.0**

Sex of Child: - It is nominal valued attribute. We can observe that value in the dataset was not found.

Table 4.6 Frequency distribution of Place of delivery

Attribute Name		Variables	Frequency	Percent (%)
Place of delivery	Valid	Home	7051	85.9
		Government health facility	905	11.0
		Private health facility	165	2-0
		Other health facility	89	1.1
		Total	8210	100..0
	Missing		-	-
	Total		**8210**	**100.0**

Place of delivery: - It is nominal valued attribute. Missing value in the dataset was not found.

Table 4.7 Frequency distribution of Wealth index

Attribute Name		Variables	Frequency	Percent (%)
Wealth index	Valid	Poor	4017	48.9
		Middle	1316	16.0
		Rich	2877	35.0
		Total	8210	100.0
	Missing		-	-
	Total		**8201**	**100.0**

Wealth index: - It is nominal valued attribute. Missing value in the dataset was not found.

Table 4.8 Frequency distribution of Frequency of listening to radio

Attribute Name		Variables	Frequency	Percent (%)
Frequency of listening to radio	Valid	Not at all	4670	56.9
		Less than once a week	2259	27.5
		At least once a week	1281	15.6
		Total	8203	99.9
	Missing	9	7	1.1
	Total		**8210**	**100.0**

Frequency of listening to radio: - It is categorical valued attribute. Missing value in the dataset was found to be 7(1.1%) respectively.

Table 4.9 Frequency distribution of Distance to health facility

Attribute Name		Variables	Frequency	Percent (%)
Distance to health facility	Valid	Big problem	6067	73.9
		Not a big problem	2143	26.1
		Total	8201	99.9
	Missing	9	9	.1
	Total		**8210**	**100.0**

Distance to health facility: - It is nominal valued attribute. Missing value in the dataset was found to be 9(1.1%) respectively.

Table4.10. Frequency distribution of Age

Attribute Name		Variables	Frequency	Percent (%)
Age of mother	Valid	15-19	241	2.9
		20-24	1501	18.3
		25-29	2607	31.8
		30-34	1774	21.6
		35-39	1314	16.0
		40-44	568	6.9
		45-49	205	2.5
	Missing	Total	8210	100.0
			-	-
	Total		**8210**	**100.0**

Age: - It is categorical valued attribute. Missing value in the dataset was not found.

Table 4.11 Frequency distribution of Mother's level of educational

Attribute Name		Variables	Frequency	Percent (%)
Mother's level of educational	Valid	No education	5830	71.0
		Primary	1989	24.2
		Secondary	251	3.1
		Higher	140	1.7
		Total	8210	100.0
	Missing		-	-
	Total		**8210**	**100.0**

Mother's level of educational: - It is nominal valued attribute. Missing value in the dataset was not found. Table 4.10 frequency distribution of Mother's level of educational in the dataset

Table4.12. Frequency distribution of Father Education level

Attribute Name		Variables	Frequency	Percent (%)
Father education level	Valid	No education	4392	53.5
		Primary	2974	36.2
		Secondary	526	6.4
		Higher	318	3.9
		Total	8165	98.5
	Missing	9	5	.0
		System	40	.5
		Total	45	.5
	Total		**8210**	**100.0**

Father level of educational: - It is categorical valued attribute. Missing value in the dataset was not found.

Table 4.13 Frequency distribution of Marital Status

Attribute Name		Variables	Frequency	Percent (%)
Marital status	Valid	Single	42	.5
		Married / Living with together	7601	92.6
		Widowed	170	2.1
		Divorced / Separate	397	4.8
		Total	8210	100.0
	Missing		-	-
	Total		**8210**	**100.0**

Marital status: - It is nominal valued attribute. Missing value in the dataset was not found.

Therefore, looking the above frequency distribution tables, there are very clear to define the characteristics of background variable.

4.3.2. Data Cleaning

Data cleaning routines work to "clean" the data by filling in missing values, smoothing noisy data, identifying or removing outliers, and resolving inconsistencies [30].Generally data cleaning decreases errors and increases the data quality. The next sections clearly discuss the tasks done in cleaning the dataset in order to gain a "clean" data for the mining algorithms.

4.3.2.1. Handling missing values

Missing Values and its problems are very common in the data cleaning process. Many datasets are plagued by the problem of missing values. The missing values problem may happen for various reasons such as incomplete manual data entry, incorrect measurements, equipment errors, and lack of consistency with other recorded data and thus deleted [30].

To handle the missing value the various methods are describe below.

- Deleting the missing attribute or deleting a record until missing values
- Replacing the deleted missing attribute by the most common value of an attribute.
- Replacing the missing value with a global constant to fill in the missing value

- Replacing the missing value with some constant, specified by the analyst
- Replacing the missing value with the field mean(for numerical attributes) or the mode (for categorical variables)
- Replacing the missing values with a value generated at random from the variable distribution observed.

In this research dataset, missing values were observed in categorical and nominal variables as it has been presented in table 4.2 above. Fortunately, three attributes (which Husband Educational Status, Frequency of listening to radio, and distance to health facility) has missing value from the categorical variables. Hence, the percentage of missing values in the mentioned attributes is not considerably significant. Thus, the missing values were replaced by the most common value of an attribute and replaced by cross checking with attribute. Using SPSS package version 22 frequency distribution shows the missing value and valid variable and using Microsoft Excel version 2010 shows the null or missing value. The above mentioned versions handling the missing value used and replace or deleting the variable in which more clear data are prepared before applying DM algorithms.

Table 4.14 shows attributes with the number of missing values, replace by the most common value of an attribute and replaced by cross checking with value.

TABLE: 4.14 Handling Missing Values

No	Attribute	Valid	Missing value	Handling Mechanism
1	Frequency of listening to radio	8203 (99.9%)	7 (1.1%)	Replace by the most common value
2	Distance to health facility	8201 (99.9%)	9(0. 1%)	Replace by the most common value
3	Father education level	8165(99.5)	45 (5%)	Replace by the most common value
4	Received BCG	8206 (99.9%)	4(0.1%)	Replace by the most common value
5	Received DPT1	8197(99.8%)	13(0.2%)	Replace by the most common value and Cross checking with value
6	Received POLIO 1	8205(99.9)	5(0.1%)	Replace by the most common value and Cross checking with value
7	Received DPT 2	8195(99.8%)	15(0.2%)	Replace by the most common value and Cross checking with value
8	Received POLIO 2	8195(99.8%)	15(0.2%)	Replace by the most common value and Cross checking with value
9	Received DPT 3	8195(99.8%)	15(0.2%)	Replace by the most common value and Cross checking with value
10	Received POLIO 3	8195(99.8%)	15(0.2%)	Replace by the most common value and Cross checking with value
11	Received MEASLE	8194(99.8%)	16(0.2%)	Replace by the most common value

4.3.3. Data transformation

Data transformation is about transforming the data to make it appropriate for mining. Therefore the researcher based on the dataset encoding data variable. Table 4.15 shows below encoding attribute.

TABLE 4.15: Data Encoding

No	Attribute's Name	Old value	New Value
1	**Wealth index**	1= Poorest 2= Poorer 3= Middle 4= Richest 5= Richer	1= Poor 2 = Middle 3 = Rich **Note** Poorest, Poorer change in to **poor** also Richest and Richer change in to **Riche**
2	**Current marital status**	0= Never in union 1= Married 2= Living with partner 3= widowed 4=Divorced 5= No longer living together/Separated	1= Single 2= Married 3= Widowed 4= Divorced **Note**: Never in union in to **single**, Married and living with partner in to **Married** Divorced and no longer living together/Separated in to **Divorced**
3	**Parity**	1,2,3,4,5,6,7,8,9,10,11, 12,,13,14,15,16,17 and 18	1=1 2= 2-3 3= 4-5 4= 6-7 5= 8-9 6= 10-11 7= 12-13
4	**Place of delivery**	10=Home, 11=Resident home 12=Other home, 20=Public Sector 21= Government hospital 22= Government health center, 23= Government health station/Clinic 24= Government health post 26= Other public sector 30=Private Sector 31=Private hospital 32=Private Clinic , 33=NGO health facility,	1=Home 2=Government health facility 3=Private health facility 4=Other Note: Home, resident home and other home change in to **Home** Public sector, Government hospital, Government health center, Government health station/Clinic and Government health post change in to **Government health facility** Private sector, Private hospital, Private Clinic in to Private health facility change in to **private health facility** NGO health facility, Other private sector, and Other home change in to **other**
5	**Received BCG**	0= No , 1= Vaccination date on, 2 = Reported by mother, 3= Vaccination marked on card and 8= don't know	1= No and 2= vaccinated **Note**: The reaming five vaccines were encoded data by this method. Vaccination on card, report by mother and vaccination marked on card in to encoded vaccinated and also don't know and no in to **No**

Finally, after pre-processing the original dataset assumed to be relevant to the target variable, which consists of 12 independent variables and one dependent variable with 3 classes and 8,210 instances, was used for constructing the model.

CHAPTER FIVE

Experimentation and Discussion

In the study, four DM algorithms were used to achieve the objective of developing predictive model using children immunization records taken from EDHS 2011 dataset. Decision tree (J48), rule induction (PART), artificial neural network (Multilayer Perceptron) and support vector machine (Simple Minimal Optimization) were used to run the experiments. 10-fold cross validation technique was also used to train and test classifiers in each experimental setting.

5.1. Experimental Setup

As has already mentioned, the data which was collected from EDHS 2011 dataset is in SPSS format. The dataset initially had 929 attributes and 11, 654 instances, but after preprocessing, it was reduced to twelve (12). Almost all from 12 attributes are independent variables from which only one is dependent variable. There are three classes and 1-5years age children 8,210 records for building the predictive model for infant immunization status in Ethiopia. Preprocessing was computed on SPSS version, 22 and Excel software version, 2010 that were extracted from EDHS 2011 database. The preprocessed dataset converted to Comma Separated Values (.CSV) file format so as to make it ready for WEKA machine learning software. Moreover, these experiments are done by using WEKA, version 3.6.10. The final dataset with its selected attributes are shown in Figure 5.1 below.

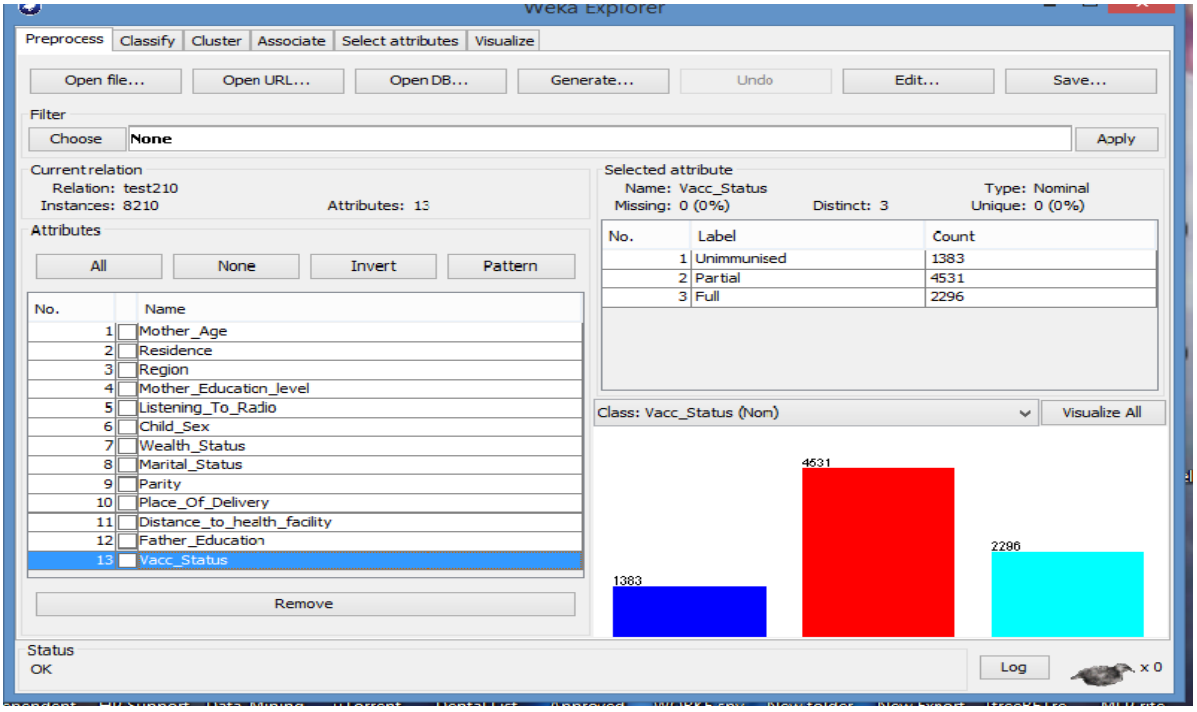


FIGURE5.1: WEKA View of the Final Dataset

The modeling stage in the DM process of the research was carried out in phase one and two sub-phases. The phase one experimentation was conducted using the one datasets. The two sub phases carried out in the study were by using all the 12 attribute of the datasets and by using 3 best attributes of the datasets selected by the DM attribute much algorithm.

The test option used by this study was 10 fold cross validation for partition of the datasets into training and test set. Data was divided into 10 folds, and each in turn is used for testing and the remainder is used for training. Each part is held out in turn and the learning scheme trained on the remaining nine-tenths; then its error rate is calculated on the holdout set. Thus, the learning procedure is executed a total of 10 times on different training sets [13]. The researcher decided to use 10 folds because extensive tests on numerous datasets, with different learning techniques, have shown that 10 is about the right number of folds to get the best estimate of error, and there is also some theoretical evidence that backs this [13]. Before presenting the details of each of the experiments conducted in this study, it is reasonable to describe first how the best attributes are selected

5.2 Attribute Selection

One important feature of WEKA, which may be crucial for some learning schemes, is the opportunity of choosing a smaller subset from all attributes. One reason could be that some algorithms work slower when the instances have lots of attributes. Another reason could be that some attributes might not be relevant. Both reasons lead to better classifiers. Determining the relevance of the attributes is searching in all possible subsets of attributes and finding the one subset that works best for classifying. Two operators are needed - subset evaluator and search method. The search method traverses that attribute subset space and uses the evaluator for quality measure. Both of them can be chosen and configured similar to the filters and classifiers [37]. After an attribute selection is performed, a list of all attributes and their relevance rank is shown in the table 5.1 below.

To select the best attributes subset selector, the researcher compared two attributes subset selections: CFS (correlation-based feature subset selection) subset evaluator with best first search method attribute evaluator with ranker search method.

CFS subset selection evaluator select best attribute by evaluation the value of a subset of attributes by considering the individual predictive attribute. CFS subset evaluator selected 3 best attributes, such as Region, Wealth Status and Place of delivery.

TABLE 5.1 Best attributes by CFS Subset evaluator.

Rank	**Attribute Name**	**Data Type**
1	Region	Nominal
2	Wealth Status	Nominal
3	Place of Delivery	Nominal

The best selected by CFS Subset evaluator are used for further experimentation.

5.3. Experimentations to Model Immunization Status

In this study, there are many experiments done using different classification algorithms. The experiment starts by default value at the beginning and then continued using different parameters. The results of the experiment were then evaluated by changing the parameters. The setup of the experiment details of the findings for each of the selected mining algorithm is presented in the following paragraphs.

5.3.1. J48 decision tree

Four experiments are conducted using J48 by the parameter with default value and changing the parameter value of unprunned value in to True value with all and best selected attributes.

- Setting # 1: J48 Experiments pruned with all attributes
- Setting # 2: J48 Experiments unpruned True with all attributes
- Setting #3: J48 Experiments pruned with best selected attribute
- Setting #4: J48 Experiment unpruned with best selected attribute

In the first setting of this experiment, the 12 attributes and 8,210 records were used by taking the default parameter value with unpruned value. The result showed that the experiment has generated a model with a tree size of **690** and **529** leaves.

In the second setting, the same number of attributes and records were used but there was a change of pruned into unpruned True to run the experiment. But relatively larger tree having a size of **4,013** and **3,046** leaves was generated.

In the third setting of the experiment, the three best selected attributes and 8,210 records are used by taking the default parameter value with pruned. The result showed that, the experiment has generated a model with a tree size of **36** and **28** leaves.

In the fourth setting of the experiment, the same records and attributes were used by taking the unpruned value True to run the experiment. But relatively larger tree having a size of **83** and **62** leaves were generated.

The performance of the four experiments is presented in the table 5.2 below.

TABLE 5.2: Performance result of J48 Decision tree

Experiment	**Accuracy**	**Leaf Size**	**Tree Size**	**Time Taken**	**WTP Rate**	**WFP Rate**	**W Precision**	**WRecall**	**WF-Measure**	**WROC Area**
J48 pruned with all attributes	**62.497**	**529**	**690**	**0.08**	**0.625**	**0.355**	**0.621**	0.625	**0.608**	**0.676**
J48 unpruned True with all attributes	56.4434	3046	4013	0.16	0.564	0.348	0.558	0.564	0.559	0.64
J48 pruned with best selected attributes	62.0585	28	36	0.14	0.621	0.345	0.618	0.621	0.607	0.692
J48 unpruned True with beast selected attributes	62.1559	62	83	0.04	0.622	0.342	0.619	0.622	0.609	0.698

As shown in the above table, 12 attributes, and 8,210 records are used by taking the default parameter value with J48 pruned. The result showed that, the experiment has generated a model with accuracy of 62.5%, weighted precision of 62.1% and weighted ROC area of 67.6% for the first setting. In the second setting, the same number of attributes and records are used but due to the change of pruned to unpruned true value, experiment has constructed a model with accuracy, weighted precision and weighted ROC area of the model were 56.44%, 55.8% and 64% respectively.

In the third setting J48 pruned with best selection attribute of result showed that the experiment has constructed a model with accuracy of **62.06%,** weighted precision of **61.8% and** weighted ROC area of **69.2%**. The fourth setting used the same number of attributes and record, but due to the change of unpruned value into True, the experiment constructed a model with accuracy, weighted precision and weighted ROC area of the model are **62.16, 61.9%** and **69.8%** respectively.

The false positive rate expression in each model shows the percentage of records which are wrongly classified into any of the four classes. Accordingly, the first setting model has wrongly classified **37.5%** of the record.

Receiver Operating Characteristic (ROC) area also shows the area under the axis of true positive and false positive rates. Therefore, as the area under the ROC curve gets larger, it shows that the classifier is putting more true positives than false positives as indicated in table 5.2 above.

Based on the above experiment, pruned J48 decision tree has scored a better accuracy than J48 unprund J48. Therefore, the pruned J48 decision tree been selected after comparing with other classifiers generated under the experiment above.

5.3.2. Sequential Minimal Optimization (SMO) Experiments

There are four experiments conducted using SMO algorithms. The first experiment was carried out by taking the default parameter values with 8,210 total records and 12 attributes and the second experiment was run using SMO algorithm by changing the build logistic models value in to True, with the same record and number of attributes. And also the third and fourth experiments conducted using SMO algorithm. The third experiment was conducted by taking the default parameter value within the best selected attributes of the three attributes and 8,210 records. The fourth experiment carried out by SMO algorithm changing the buildlogosticModels value in to True with the same record and number attributes.

TABLE 5.3 Performance of SMO for the all and best attributes

Experiment	**Accuracy**	**WTP Rate**	**WFP Rate**	**W Precision**	**W Recall**	**WF-Measure**	**WROC Area**
SMO with all attribute by default parameters	60.475	0.605	0.377	0.593	0.605	0.582	0.658
SMO with all attribute by buildlogisticModels True	**61.2911**	**0.613**	**0.372**	**0.605**	**0.613**	**0.591**	**0.685**
SMO with best select attributes by default parameters	60.4629	0.605	0.356	0.597	0.605	0.589	0.663
SMO with best selected attribute by buildlogisticModels True	60.6577	0.607	0.361	0.599	0.607	0.589	0.676

Based on the above experiment SMO table 5.3, SMO with all attributes in build logistic Models is true scored a better accuracy compared with SMO with all attributes by default parameters .

Taking the performance parameters indicated above, SMO with all attributes build logistic Model True was selected compared with other classifiers under this section

5.3.3. PART Rule induction

Four experiments were conducted using PART by using the default variable and changing the parameter with unpruned to True. The experimental settings are shown as follows:

- Setting # 1: PART Experiments with default value
- Setting # 2: PART Experiments un pruned True
- Setting # 3: PART Experiments best selected attribute with default value
- Setting # 4: PART Experiments best selected attribute with un pruned True

The performance of the four setting experiments are presented in 5.4 below

TABLE5. 4. Experiment at result of PART rule induction

Experiment	**Accuracy**	**WTP Rate**	**WFP Rate**	**W Precision**	**W Recall**	**WF-Measure**	**WROC Area**	**Number of rule**
PART with default parameters and all attribute	57.4909	0.575	0.354	0.566	0.575	0.567	0.653	740
PART with unprunned True and all attribute	55.0183	0.55	0.336	0.553	0.55	0.549	0.632	2084
PART best attribute selection with default parameters	**62.0097**	**0.62**	**0.345**	**0.618**	**0.62**	**0.607**	**0.694**	**20**
PART best attribute selection with unprunned True	62.0041	0.62	0.342	0.618	0.62	0.608	0.7	82

In the first setting of PART experiment a classifier with accuracy of 57.49%, weighted TP rate of 57.5%, and weighted FP rate of 35.4% was generated by taking default value of unprunned. In addition to these performance parameters, the model has generated a total of **740** rules to represent the patterns found within the dataset.

In the second setting the same number of attribute and record were used by changing the default parameter value unpruned in to True. The second setting produced a classifier with an accuracy of 55.02%, weighted TP rate of 55%, and weighted FP rate of 33.6%.

In the third setting of PART experiment a classifier with accuracy of 62.0%, weighted TP rate of 62%, and weighted FP rate of 34.5% was generated by taking default value of unpruned. In addition to these performance parameters, the model has generated a total of **20** rules to represent the patterns found within the dataset.

In the fourth setting the same number of attribute and record were used by changing the default parameter value unpruned in to True. The fourth setting produced a classifier with an accuracy of 62.0%%, weighted TP rate of 62%, and weighted FP rate of 34.2%.

Based on the experiment PART table 5.4, with best attribute by default has scored a better accuracy, taking the performance parameters indicated above.

Therefore, this result has been selected compared with other classifiers in the experiments shown above.

5.3.4. Multilayer Perception (MLP) Neural Network

This is the last experiment conducted using MLP with all attributes and best selected attributes taking default parameter values.

Table 5.5 below presented summary of the performance of model created by MLP neural network

TABLE 5.5 MLP Experimental result using MLP Neural Network

Experiment	**Accuracy**	**WTP Rate**	**WFP Rate**	**W Precision**	**W Recall**	**WF-Measure**	**WROC Area**
MLP with all attribute by default parameters	58.4531	0.585	0.354	0.576	0.585	0.576	0.659
MLP with best selected attribute by default parameters	**61.0962**	**0.611**	**0.337**	**0.607**	**0.611**	**0.601**	**0.695**

The result showed that, with all record and attributes MLP has generated a model with accuracy; weighted precision and weighted recall of 58.45%, 57.6% and 58.5% .The experiment result of the second on the other hand, experimental result of the second setting indicate an accuracy 61.1%, weighted precision of 60.7% and weighted recall 61.1%.

The above table shows, MLP with best selected attributes by default parameters has better result than the MLP with all attribute by default parameter. So the result of MLP with best selected attributes by default parameters was selected.

5.3.5. Model Evaluation

One of the most important objectives of the study is identifying which DM algorithm performs best in predicting the infant immunization status. Therefore, the experiments in the study were carried out with J48 decision tree, SMO support vector machine; MLP artificial neural network and PART rule induction algorithms. The dataset used for each experiment in the study either have all attributes and best selected attributes. Model comparison was performed using performance evaluation matrix like true prediction rate, false prediction rate, recall, precision, ROC area and accuracy of the model.

The detail result of the best model selected from each classification category is shown in table 5.6 below.

TABLE 5.6 Summary of the performance of the beast model created by classification algorithm

Experiment	**Accuracy**	**WTP Rate**	**WFP Rate**	**W Precision**	**WF-Measure**	**W Recall**	**WROC Area**
J48 pruned with all attributes	62.497	0.625	0.355	0.608	0.605	0.625	0.676
SMO with all attribute by buildlogisticModels is True	61.2911	0.613	0.372	0.605	0.591	0.613	0.685
PART best attribute selected with default parameters	62.0097	0.62	0.345	0.618	0.607	0.62	0.694
MLP with best selected attribute by default parameters	61.0962	0.611	0.337	0.607	0.601	0.611	0.695

As it is shown in the Table 5.6 J48 decision tree classification has achieve the best as compared to SMO, PART and MLP algorithms specifically, J48 pruned with all attributes provides the best accuracy of 62.5%.Accordingly, the model created by this algorithm is selected as the best model that can predict the infant immunization status. Table 5.7 depicts details of classification result of pruned J48 decision tree.

5.7. Confusion Matrix for J48 decision tree classifier

The confusion matrix is a useful tool for analyzing how well your classifier can recognize tuples of different classes. The confusion matrix for decision tree shown in table 5.7 illustrate that out of the total 8,210 records 505 records were correctly classified as “unimmunized”, 3,653 records were correctly classified as “partial” and 973 records were correctly classified as “full”. The classifier incorrectly classified 36 records as “unimmunized” and 3,653 records as “partial” and 46 records as “full”. In general the classifier correctly classified 5,131 records and incorrectly classified 3,079 records out of a possible 8,210. The accuracy of the classifier to correctly predict the class value as “unimmunized”, “Partial” and “full” is 62.5%.

Table 5.7 Confusion Matrix for J48 Decision Tree Model

===Confusion Matrix===			
A	**B**	**c**	**Classified as**
505	842	36	**a = Unimmnised**
339	3653	539	b = **Partial**
46	1277	973	c= **Full**

5.4. Generated Rules from Decision Trees

In this study, J48 classifier has achieved relatively the highest in most of performance evaluation criteria compared to MLP, SMO and PART algorithms. Therefore, the model generated by J48 classifier with all (12) attributes was selected as the model that can predict the infant immunization status. As can be seen in Appendix 3, J48 decision tree generated 470 rules for predicting infant immunization status from which the researcher considered. Based on the discussion with the domain expert have selected highest accuracy and also show the most important attributes, 11 creating rules were selected from decision tree. These rules are listed.

1. If place of delivery = Home and Region = Amhara and Listening-To-Radio = Not-At-All then Partial (537.0/160.0)
2. If please of delivery = Home and Region = Amhara AND Listening-To-Radio = Less-Than-Once-A-Week then Partial (227.0/75.0)
3. If Place of delivery = Home and Region = Tigray and Mother Age = 30-34 and Listening radio= Not at all and Mother-Education-level = No-Education and Parity = 4-5 and Marital-Status = Married and Fother-Education-level = No-Education: Partial (36.0/15.0)
4. If Place of delivery = Home Region = Affar and Mother-Education-level = No-Education and Wealth-Status = Poor and Listening-To-Radio = Not-At-All AND Mother_Age = 25-29 and Parity = 6-7 then Unimmunised (10.0/1.0)
5. If Place of delivery = Home and Region = Affar and Mother-Education-level = No-Education and Wealth-Status = Poor and Listening-To-Radio = Not-At-All and Mother_Age = 25-29 and Parity = 4- 5 and Distance-to-health-facility = Big-Problem: Unimmunised (39.0/11.0)
6. If Place of delivery = Home and Region = Affar and Mother-Education-level = No-Education and Mother-Age = 25-29 AND Wealth-Status = Rich and Father-Education = No-Education and Listening-To-Radio = Not-At-All then Unimmunised (17.0/7.0)
7. If Place of delivery = Home and Region = Somali and Wealth-Status = Poor and Mother-Education-level = No-Education and Listening-To-Radio = Not-At-All and Father-Education = No-Education: Unimmunised (214.0/83.0)
8. If Place-Of-Delivery = Government-Health-Facility and Residence = Rural and Region = Amhara and Parity = 2-3 and Distance-to-health-facility = Big-Problem then Partial (4.0/1.0)

9. Place-Of-Delivery = Government-Health-Facility and Residence = Rural and Region = Somali and Wealth-Status = Poor and Marital-Status = Widowed: Partial (2.0)
10. Place-Of-Delivery = Government-Health-Facility and Residence = Rural and Region = SNNP and Father-Education = Primary and Listening-To-Radio = Not-At-All and Mother-Education-level = No-Education then Partial (2.0)
11. If Place-Of-Delivery = Private-Health-Facility and Residence = Urban and Region = SNNP then Partial (3.0/1.0)

5.5. Discussion on Major Findings

The rules generated from the decision tree in section 5.6 predict the infant immunization. The rule considers different conditions of the attributes age, parity, mother's education, father's education, wealth status, distance of health facility, marital status, region, place of delivery and residences were identified as having a higher statistical significance in classifying the predicted value for infant immunization.

Place of delivery in Afar and Somalia was found to be at home due to lack of awareness about immunization. The other factors are low income, uneducated family and absence of health facility in the nearby area. For the above mentioned reasons infants are Unimmunized.

In governmental health facility there is no better means of follow up mechanism for drop out, like capturing individual data profile, poor linkage of health extension professionals and shortage of awareness about the valid type of vaccination like polio, BCG, DPT and measles.

Data from SNNP urban area indicated that private clinic has also the same result like that of government health facilities. Since the vaccinations given from private health organization is for free they only collect for the service rendered with no extra profit. Therefore they become uninterested in the awareness making in the society to immune their children.

In Amhara region infants were found to be partially immunized due to long distance and lack of confortable geographical and natural condition to the health facilities. The illiteracy of parents is significant reason for unimmunized infants in Affar, Tigray & Somalia regions. Moreover they do not listen to radio or less than once a week. Due to this there is less information and awareness about the immunization program and the use of vaccine.

5.6 Prototype Development

The development of graphical user interface in this study was done using Microsoft visual studio 2010. This prototype graphical user interface was developed based on the model generated by J48 decision tree classifier with pruned parameter and all attributes. The rules used by the researcher to design the graphical user interface for predicting the immunization status of infants are the 11 rules listed and briefly discussed in section 5.6. Figure 5.2, this prototype prediction model can be used for predicting infant immunization status based on the rules generated by J48 classifier.

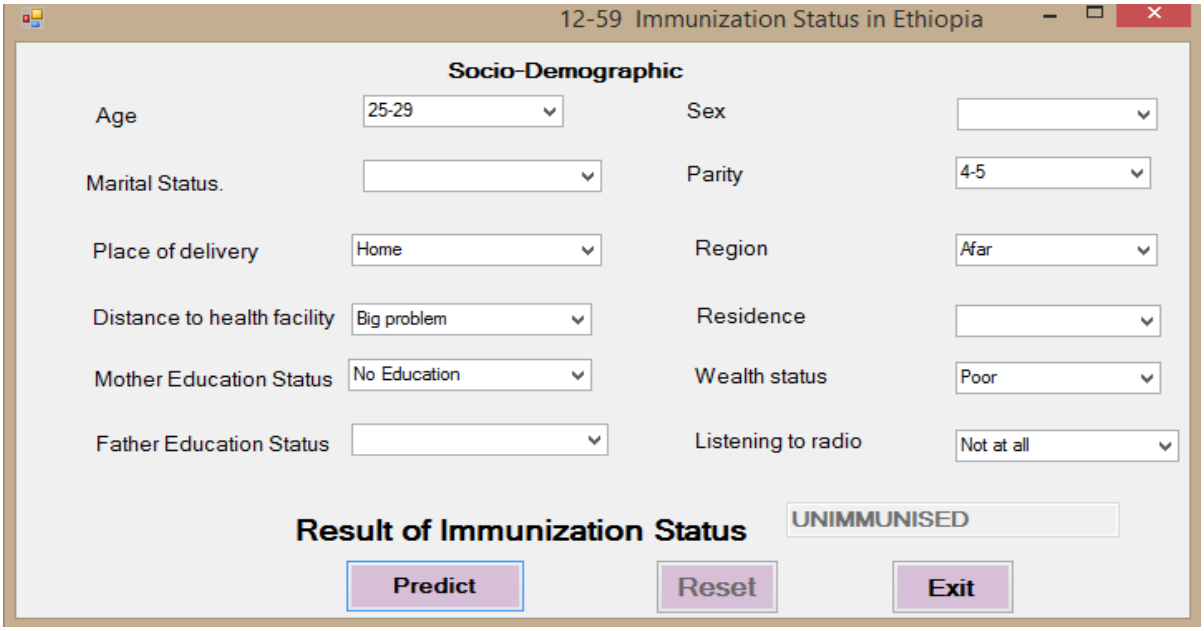


Figure 5.2: Infant immunization status prediction prototype user interface with sample result

5.7. The validity of the designed User Interface (Prototype)

In order to control the drop out of infant immunization, the researcher developed user interface prototype. The questionnaire was developed based on the objective to check the validity of the interface. The questionnaire consisted of five items on five point scale ranging from excellent to poor.

Four health professionals from Shiro Meda and Jal Meda Health Centers have been asked to respond to the questionnaire items. After they responded to the questionnaire, the researcher had 30 minutes discussion with each of the health professionals regarding the interface prediction of infant immunization status.

TABLE 5.10 Validity on Infant immunization status user interface

Questions	Variables	No of Respondents	Frequency	Percent (%)
The features of interface in terms of ***easy to use***	Very Good (4)	3	12	70.59
	Excellent (5)	1	5	29.41
	Total	**4**	**17**	**100.0**
The features of interface in terms of ***easy to understand***	Very Good (4)	3	12	70.59
	Excellent	1	5	29.41
	Total	**4**	**17**	**100.0**
The features of interface in ***facility decision making***	Very Good	1	4	21.05
	Excellent	3	15	78.95
	Total	**4**	**19**	**100.0**
Time Taking The features of interface in terms of ***time taking***	Very Good	4	16	100.0
	Total	**4**	**16**	**100.0**
The features of interface in ***formulating Prediction***	Very Good	1	4	21.05
	Excellent	3	15	78.95
	Total	**4**	**19**	**100.0**

As illustrated in the Table above, the result of the questionnaire shows almost all the respondents rated “excellent” and “very good” on the validity of the interface.

When we come to the result of the discussion held with the respondents, one of the respondents explained her view as follow:

“I found the interface so impressive because when we do our job, we don’t use any tool that will help us to follow and support infant to be fully immunized. I believe, this prototype will help the health professionals to control the drop out of the infant immunization”.

The other respondent explained her ideas saying “I have a lot of things to do in EPI program. There are routine and manual activities that I need to accomplish. Even if this prototype helps to follow up the infant immunization, I believe it is time consuming to inter the necessary data to predict infant immunization”. The third respondent appreciated the interface but she complained that her computer skill is limited. The forth respondent was impressed with the function of the interface and she recommended the researcher to implement the interface in a very short time. She further revealed her view by saying “I think this is very important for me and other health professionals to properly implement EPI program and to predict the infant immunization status. I personally suggest the researcher to further discuss on the implementation of the interface with the concerned stakeholders so that the health professionals will use the prototype”.

In general, an attempt has been made to check the validity of the interface at two health centers (Shiro Meda and Jal Meda) with the health professionals. Thus, the interface is believed to be valid enough to predict (using the selected rules) the infant immunization status based on the collected data from the two health centers.

CHAPTER SIX
CONCLUSION AND RECOMMENDATIONS

6.1. Conclusion

Vaccination has been shown to be one of the most effective public health interventions worldwide, through which a number of serious childhood diseases have been successfully eradicated. In Ethiopia, vaccines are available for tuberculosis, diphtheria, pertussis, tetanus, polio and measles.

According to the EDHS 2011, as for coverage for specific vaccines, 66% of infant had received the BCG vaccine, and 56 % had received the measles vaccine [26]. A relatively high percentage of infant received the first DPT dose. However, only 37% went on to receive the third dose 64% of DPT, reflecting a dropout rate of 43%. More than 82% infant received the first dose of polio, but only about 44% received the third dose, reflecting a dropout rate of 46%. Even though DPT and polio vaccines are often routinely administered at the same time, polio coverage is higher than DPT coverage [26]. The above result indicated that the infants are no full vaccinated.

This research applies data mining to explore hidden pattern from electronic medical record in the aim of predicting the infant immunization status in Ethiopia so that it helps to understand immunization status of infants in different region of Ethiopia. This in turn is important to fairly distribute immunization service for infants in all regions of Ethiopia regardless of socioeconomic and demographic factors that affects immunization.

The goal of the research is attempting to identify the determinants factors that affect the status of infant immunization in each region. When conducting this research, data were prepared for analysis and model building by cleaning, extracting, and transforming in to a format suitable for the data mining tool. This followed by the application of classification algorithm in order to construct best prediction model. The best model that can be used to predict the routine immunization was selected based on its performance compared to other models.

Hybrid data mining process model, which incorporates the six steps, has been used as a methodology. The study considered a total of 8210 records and 12 attributes to predict the infant immunization status in Ethiopia. The algorithms used are decision tree (J48), rule induction (PART), artificial neural network (MLP) and support vector machine (SMO). The mining

algorithms were used in order to build the models that can predict the infant immunization status in Ethiopia. Data cleaning and data transformation was done to prepare the dataset for experiments.

The evaluation of best performed algorithms compared based on accuracy, true positive, false positive, ROC area and accuracy to build model. Based on performance evaluator, the best performing algorithms are J48 classifier followed by PART classifier and SMO classifier. The best selected model in this study is generating by J48 decision tree with all attributes. The accuracy of this model is 62.5%.

The selected classifier, J48 decision tree generated a total of 470 rules from which the researcher has selected only 11 after discussion made with domain experts. The rules that were generated show different status of immunization of infants in different regions of Ethiopia given the 12 attributes.

To rese use of hidden knowledge extracted with the help of J48 decision tree classification algorithm a graphical user interface was designed using selected model. The development of graphical user interface in this study was done using Microsoft visual studio.

The most important attributes that determine infant immunization status were place of delivery, region, mother's educational level, listening to radio, father education level, residence, mother age, wealth status, parity, distance to health facility and marital status. This findings are approved by domain experts and literature cited on the success and challenges of EPI program in Ethiopia [11, 26]

The results found from the research indicated that data mining is advantageous in bringing relevant information from large and complex dataset so that anybody can use the information for decision making.

6.2 Recommendation

The researcher suggests the following major recommendations for MOH, Zonal, and District, health professionals, and researchers on the findings of the study:

- MOH, Zonal and District health officials at different levels are recommended to create awareness on infant immunization for the pastoralists for effectiveness of vaccine and its coverage.
- Health professionals need give due emphasis on immunization and children during Antenatal care visit and outreach programs related to mother and children.
- Health professionals need to intensively teach families about the vaccine type and their uses and side effects.
- The update the system will be on the domain expert
- The application of the prototype is believed to minimize the infant immunization drop out.
- The knowledge of data mining is also recommended for public health professional because it extracts hidden knowledge
- There is a need for the development of knowledge based system for infant immunization status with domain experts
- This research attempted to develop prediction model and prototype graphical interface for infant immunization status. There is, however, a need for the development of knowledge based system for infant immunization status with domain experts. This should be a further research direction.
- In this study the prediction of infant immunization status was done at regional level. The researcher proposed to continue other researchers at district level in order to get a clear alertness about the locations of the dropout rate.

Finally and most important, this research can be used for academic purposes being a reference for those looking to work in the identified problem domain or generally on data mining.

Reference

1. Grant J.P. The state of the world children: Oxford University Press. UNICEF, New York, 1991
2. Bradshaw, York W., Rita Noonan, Laura Gash, and Claudia Buchmann Sershen. "Borrowing against the future: Children and third world indebtedness. Social forces 71, no. 3 (1993): 629-656.
3. Maurice, John M., and Sheila Davey.State of the World's Vaccines and Immunization. World Health Organization, 2009.
4. UNICF. UNICEF annual report, Ethiopia, Addis-Ababa, 1996.
5. Ministry of health. Communicable diseases control Division report, Addis Ababa,1976
6. Ministry of health. Health policy of socialist Ethiopia, Addis Ababa. 1976.
7. Ministry of health Annual report of the expanded program on immunization (EPI). 1990.
8. Ministry of health Guidelines of EPI in Ethiopian, Addis Ababa. 1981.
9. Addisie M, Feleke A, Edris M, Mengistu D, Eredie A, Woreta K, et al: Expanded Program in Immunization. 2002.
10. World Health Organization.Expanded program on immunization: Immunization practice module 1and 2. 1998.
11. Expanded program of immunization, Ethiopia, November 8, 2008
12. Federal Democratic Republic of Ethiopia Ministry of Health & Ethiopia Health and Nutrition Researches Institution. Ethiopia National Immunization Coverage survey 2012.
13. Mehmed Kantardzic John Wiley & Sons. Data mining concept, Models, Methods and Algorithm. 2003.
14. Witten, Ian H., and Eibe Frank. Data Mining: Practical machine learning tools and techniques. Morgan Kaufmann, 2005.
15. T.G. Roche. Expect increased adoption rates of certain types of EHRs, EMRs, Managed Healthcare Executive 16:4, 58, (2006)
16. Fayyad, Usama, and Paul Stolorz. "Data mining and KDD: promise and challenges." Future generation computer systems 13, no. 2 (1997): 99-115.
17. U.Fayyad, G.Piatetsky-Shapiro, P.Smyth, R.Uthurusamy. Advances in Knowledge Discovery and Data Mining. Massachusetts: MIT Press (1996): pp.2-4.

18. Zhang, Shichao, Chengqi Zhang, and Qiang Yang. Data preparation for data mining. Applied Artificial Intelligence 17, no. 5-6 (2003): 375-381.
19. Yan, Xiaowei, Chengqi Zhang, and Shichao Zhang. "Toward databases mining: Pre-processing collected data." Applied Artificial Intelligence 17, no. 5-6 (2003): 545-561.
20. Tadesse, Henoke, Amare Deribew, and Mirkuzie Woldie. "Predictors of defaulting from completion of child immunization in south Ethiopia, May 2008–A case control study." BMC Public Health 9, no. 1 (2009): 150.
21. Trybula, Walter J. "Data mining and knowledge discovery." Annual review of information science and technology 32 (1997): 197-229..
22. Berry, Michael J., and Gordon S. Linoff. Data mining techniques: for marketing, sales, and customer relationship management. John Wiley & Sons, 2004.
23. Han, Jiawei. "Kamber., Micheline.: Data Mining: Concepts and Techniques." (2001): 550.
24. Andre, F. E., R. Booy, H. L. Bock, J. Clemens, S. K. Datta, T. J. John, B. W. Lee et al. "Vaccination greatly reduces disease, disability, death and inequity worldwide." Bulletin of the World Health Organization 86, no. 2 (2008): 140-146.
25. Hand, David J., Heikki Mannila, and Padhraic Smyth. "Principles of data mining (adaptive computation and machine learning)." (2001).
26. Angela Gentile. Pediatric disease burden and vaccination recommendations: Understanding local differences. International Journal of infectious diseases. [Review].30(30):1019- 29, 2010.
27. Ethiopia Demographic and Health Survey Addis Ababa, Ethiopia 2011.
28. WHO, UNICEF, World Bank. States of the world's vaccines and immunization. 3rd ed. Geneva: world health organization 2009.
29. Abio, A., Mark, J. P., Rita, F. H., Rafael, H., Laura W., Russell S. William J. Lack of Evidence of Measles Virus Shedding in People with In apparent Measles Virus Infections.
30. Li, Jing-song, Hai-yan Yu, and Xiao-guang Zhang. Data Mining in Hospital Information System.
31. Han, Jiawei, Micheline Kamber, and Jian Pei. Data mining: concepts and techniques. Morgan kaufmann, 2006.

32. Fayyad, Usama M., Gregory Piatetsky-Shapiro, Padhraic Smyth, and Ramasamy Uthurusamy. Advances in knowledge discovery and data mining. 1996.
33. Margahny, M. H., and A. A. Mitwaly. Fast algorithm for mining association rules. In the conference proceedings of AIML, CICC, pp (36-40) Cairo, Egypt, pp. 19-21. 2005.
34. Sunita Soni, O.P "Using Associative Classifiers for Predictive Analysis in Health Care DM", International Journal of computer Application, 2010, Vol.4.
35. Koh, H.C. & Leong, S. data mining applications in the context of casemix, academy of medicine, 2001, Vol.30, pp 41- 49.
36. Tan AC, Gilbert D, "Ensemble Machine Learning On Gene Expression Data For Cancer Classification", Appl Bioinformatics, 2003, p p 75-78.
37. Joseph L Beult, Colin R. Good all, Peter J" data mining a diabetic data Warehouse, Foes, Artifical Intelligence in medicine", 2002, Elsevier vole 26, pp37-54
38. Dimov, Rossen, Michael Feld, Dr Michael Kipp, Dr Alassane Ndiaye, and Dr Dominik Heckmann. "Weka: Practical machine learning tools and techniques with Java implementations." AI Tools SeminarUniversity of Saarland, WS 6, no. 07 (2007). http://www.dfki.de/~kipp/seminar_ws0607/reports/RossenDimov.pdf
39. Dzeroski S.Towards a General Framework for Data Mining. In: Dzeroski, S and Struyf, J (Eds.), Knowledge Discovery in Inductive Databases. LNCS 47474. Springer; 2006.
40. Han, J. data mining QL: A Data Mining Query Language for Relational Databases. In proceedings of DMKD-96 (SIGMOD-96 Workshop on KDD). Montreal: Canada; 1996.
41. Meo, R. An Extension to SQL for Mining Association Rules: Data Mining and Knowledge Discovery. Kluwer Academic Publishers; 1998, Vol (2), pp (195-224).
42. Imielinski T and Virmani A. MSQL: A Query Language for Database Mining. Data Mining and Knowledge Discovery. Kluwer Academic Publishers; 1999, Vol. 3, pp 373-408.
43. Sarawagi S. Integrating Association Rule Mining with Relational Database Systems: Alternatives and Implications. Data Mining and Knowledge Discovery; 2000, Vol. 4, pp 89–125.

44. Botta, Marco, Jean-François Boulicaut, Cyrille Masson, and Rosa Meo. "Query languages supporting descriptive rule mining: a comparative study." InDatabase Support for Data Mining Applications, pp. 24-51. Springer Berlin Heidelberg, 2004.
45. Fayyad, Usama, Gregory Piatetsky-Shapiro, and Padhraic Smyth. "From data mining to knowledge discovery in databases." AI magazine 17, no. 3 (1996): 37.
46. Fayyad, Usama M., Gregory Piatetsky-Shapiro, Padhraic Smyth, and Ramasamy Uthurusamy. "Advances in knowledge discovery and data mining." (1996).
47. Vickery, Brian. "Knowledge discovery from databases: an introductory review."Journal of Documentation 53, no. 2 (1997): 107-122.
48. Piatetsky-Shapiro, Gregory, Ronald J. Brachman, Tom Khabaza, Willi Kloesgen, and Evangelos Simoudis. An Overview of Issues in Developing Industrial Data Mining and Knowledge Discovery Applications. In KDD, vol. 96, pp. 89-95. 1996.
49. Jackson, Joyce. Data mining: A conceptual overview. Communications of the Association for Information Systems 8, no. 19 (2002): 267-296
50. Azevedo, Ana Isabel Rojão Lourenço. "KDD, SEMMA and CRISP-DM: a parallel overview." (2008).
51. Cios, K, Witold, P, Roman, S abd Kurgan, A. Data Mining: A Knowledge Discover Approach. New Yourk, USa: Springer, Jiawei; 2007
52. Azevedo, Ana Isabel Rojão Lourenço. "KDD, SEMMA and CRISP-DM: a parallel overview." (2008).
53. Yu, Hwan-Jo. "Data Mining via Support Vector Machines: Scalability, Applicability, and Interpretability." PhD diss., University of Illinois, 2004.
54. Data mining in Healthcare: Current Applications and Issues by Ruben D. Canlas Jr.Aug-2009.
55. Christy, T. Analytical tools help health firms fight fraud. Insurance & Technology,1997,Vol .22(3), pp 22-26
56. Biafore, S. Predictive solutions bring more power to decision makers. Health Management Technology, 1999, Vol.20 (10), pp 12-14.

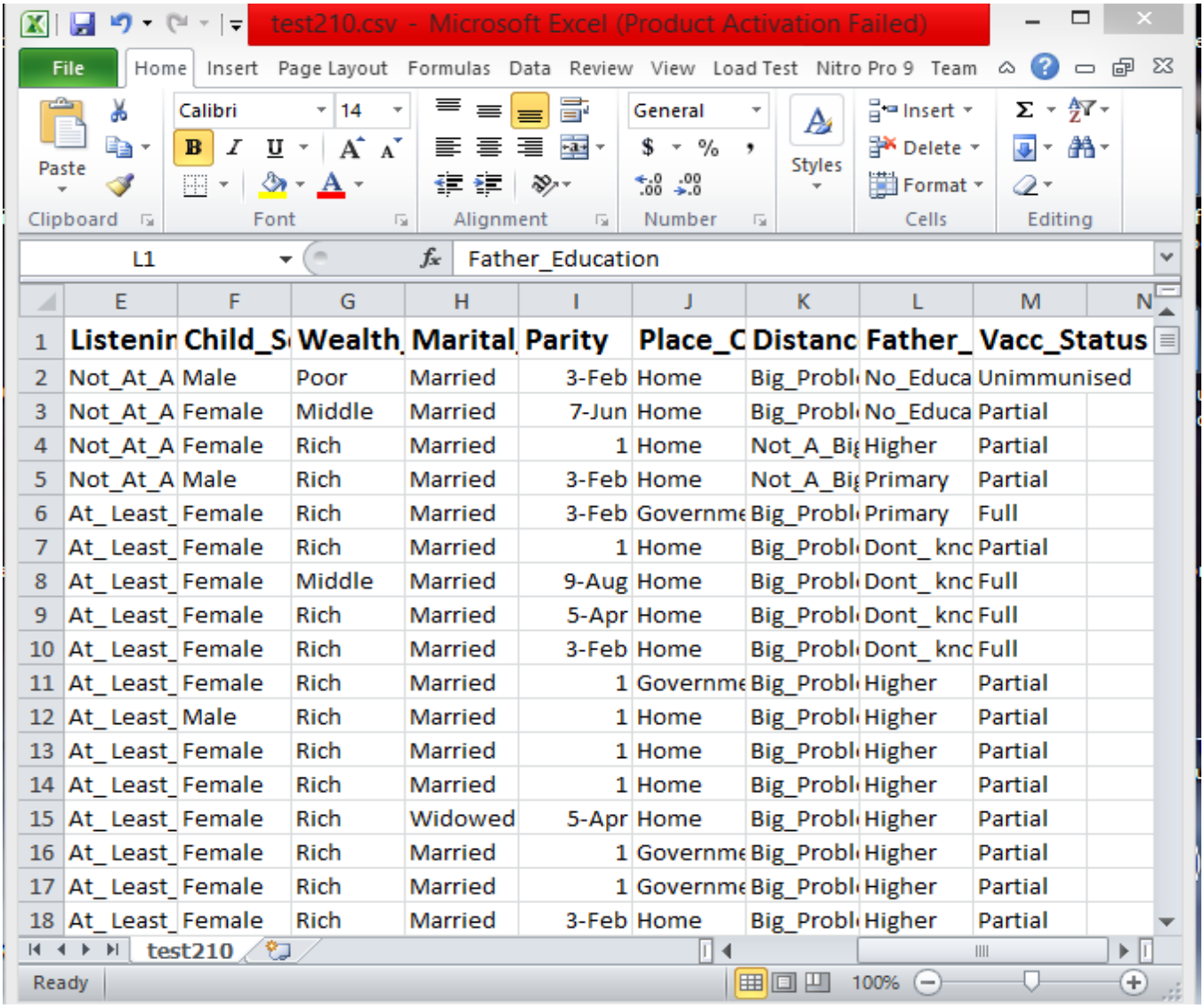
57. Wanqing, Li, Ma Lihua, and Wei Dong. "Data mining based on rough sets in risk decision-making: foundation and application." WSEAS Transactions on Computers 9, no. 2 (2010): 113-123.
58. Jans, Mieke, Nadine Lybaert, and Koen Vanhoof. "A Framework for Internal Fraud Risk Reduction at IT Integrating Business Processes: The IFR 2 Framework. " International Journal of Digital Accounting Research 9, no. 1 (2009).
59. Silver, Michael, Taiki Sakata, Hua-Ching Su, Charles Herman, Steven B. Dolins, and Michael J. O Shea. "Case study: how to apply data mining techniques in a healthcare data warehouse." Journal of healthcare information management 15, no. 2 (2001): 155-164.
60. A fuzzy neural network for assessing the risk of fraudulent financial reporting", Managerial Auditing Journal, (2003) Vol. 188, pp 657-665.
61. Major, John A., and Dan R. Riedinger. "EFD: A Hybrid Knowledge/Statistical- Based System for the Detection of Fraud." Journal of Risk and Insurance 69, no. 3 (2002): 309-324.
62. Benko, A. and Wilson, B. , "Online Decision Support Gives Plans An Edge", Managed Healthcare Executive, 2003, Vol.13 (5), pp 20-25
63. Fabio, A., Mark, J. P., Rita, F. H., Rafael, H., Laura W., Russell S. William J. Lack of Evidence of Measles Virus Shedding in People with In apparent Measles Virus Infections.
64. JAMA. Vaccine Preventable Deaths and the Global Immunization Vision and Strategy, 2006-2015. JAMA 2006; 295:2840–2
65. Centers for Disease Control and Prevention. Vaccine preventable diseases: improving vaccination coverage in children, adolescents, and adults. Morbidity and Mortality Weekly Report 1999; Vol. 48: (RR–8).
66. C.-C. Chang and C.-J. Lin. Training nu-support vector classifiers: Thoery and algorithms. Neural Computation, 13:2119–2147, 2001,
67. Vapnik, Vladimir. The nature of statistical learning theory. springer, New York 2000.
68. Kohavi, Ron. "Scaling up the accuracy of naive-Bayes classifiers: A decision-tree hybrid." In KDD, pp. 202-207. 1996.

69. Kidane T, Yigzaw A, Sahilemariam Y, Bulto T, Mengistu H, Belay T, et al. National EPI coverage survey report Ethiopian Journal of Health Development 2006, 2008;22(2):148-57.
70. Etana, Belachew, and Wakgari Deressa. "Factors associated with complete immunization coverage in children aged 12–23 months in Ambo Woreda, Central Ethiopia." BMC public health 12, no. 1 (2012): 566.
71. Reasons for failure of immunization: A Cross-Selection Study among 12-23 months old children of Lucknow, India http://www.advbiores.net/tem/AdvBiomedRes2171-107309-025855-pdf February 2014
72. Gayathri, R., and A. Malathi. "Exploration of Data Mining Techniques in Record Deduplication." International Journal 2013. http://www.techopedia.com/definition/14650/data-preprocessing April 2014
73. Isaac B. Assessment Of Expanded Program On Immunization Service Utilization In Sekyere West District Of Ashanti Region, Ghana Kwame Nkrumah University Of Science And Technology College Of Health Sciences School Of Medical Sciences Department Of Community Health; March, 2014.
74. Cios, K, Witold, P, Roman, S and Kurgan, A. Data Mining: A Knowledge Discovery Approach. New York, USA: Springer, Jiawei; 2007.
75. Adebayo Peter Idowu, Bernard Ijesunor Akhigbe, Olajide Olusegun Adeosun, Aderonke Anthonia Kayode, and Adekemi Faidat Osungbade. Data Mining Techniques for Predicting Immunize-able Diseases: Nigeria as a Case Study, Volume 5– No.7,May 2013
76. Assamnew, Selam. "Predicting the occurrence of Measles Outbreak in Ethiopia Using data mining Technology." Addis Ababa University, 2011.
77. Anagaw, Shegaw. Application of data mining technology to predict child mortality patterns: The case of (Butajira Rural Health Project)BRHP, M.Sc. Thesis, Addis Ababa University, Addis Ababa, 2002.
78. Hemalatha, M., and S. Megala. "MINING TECHNIQUES IN HEALTH CARE: A SURVEY OF IMMUNIZATION." Journal of Theoretical & Applied Information Technology 25, no. 2 (2011).

79. Tadesse Misganaw. Mining ART data set to predict CD4 cells count the case of Jimma, Bonga and Aman Hospitals. Addis Ababa University School of graduate students school of Information science and school of Public Health. June, 2013.
80. Jantan, Hamidah, Abdul Razak Hamdan, and Zulaiha Ali Othman. "Human Talent Prediction in HRM using C4. 5 Classification Algorithm." International Journal on Computer Science & Engineering (2010).
81. Yeh, Duen-Yian, Ching-Hsue Cheng, and Yen-Wen Chen. "A predictive model for cerebrovascular disease using data mining." Expert Systems with Applications 38, no. 7 (2011): 8970-8977.
82. Shin, Kyung-Shik, Taik Soo Lee, and Hyun-jung Kim. "An application of support vector machines in bankruptcy prediction model." Expert Systems with Applications 28, no. 1 (2005): 127-135.
83. Quinlan, J. Ross. "Induction of decision trees." Machine learning 1, no. 1 (1986): 81-106.
84. Rajput, Anil, Ramesh Prasad Aharwal, Meghna Dubey, S. P. Saxena, and Manmohan Raghuvanshi. "J48 and JRIP rules for e-governance data."International Journal of Computer Science and Security (IJCSS) 5, no. 2 (2011): 201.
85. Margaret H. Danham,S. Sridhar, ” Data mining, Introductory and Advanced Topics”, Person education , 1st ed., 2006
86. Aman Kumar Sharma, Suruchi Sahni, “A Comparative Study of Classification Algorithms for Spam Email Data Analysis”, IJCSE, Vol. 3 No. 5, 2011, pp. 1890-1895
87. Witten, Ian H., and Eibe Frank. Data Mining: Practical machine learning tools and techniques. Morgan Kaufmann, 2005.
88. William E. Spangler, Jerrold H. May and Luis G. Vargas Journal of Management system. Page [37] of 37-62 1999 http://www.jstor.org/discover/10.2307/40398417?uid=3738256&uid=2134&uid=368470121&uid=2&uid=70&uid=3&uid=368470111&uid=60&sid=21101751936641 march 2014
89. Vapnik, Vladimir. The nature of statistical learning theory. springer, New York 2000.
90. Burges, Christopher JC. "A tutorial on support vector machines for pattern recognition." Data mining and knowledge discovery 2, no. 2 (1998): 121-167.

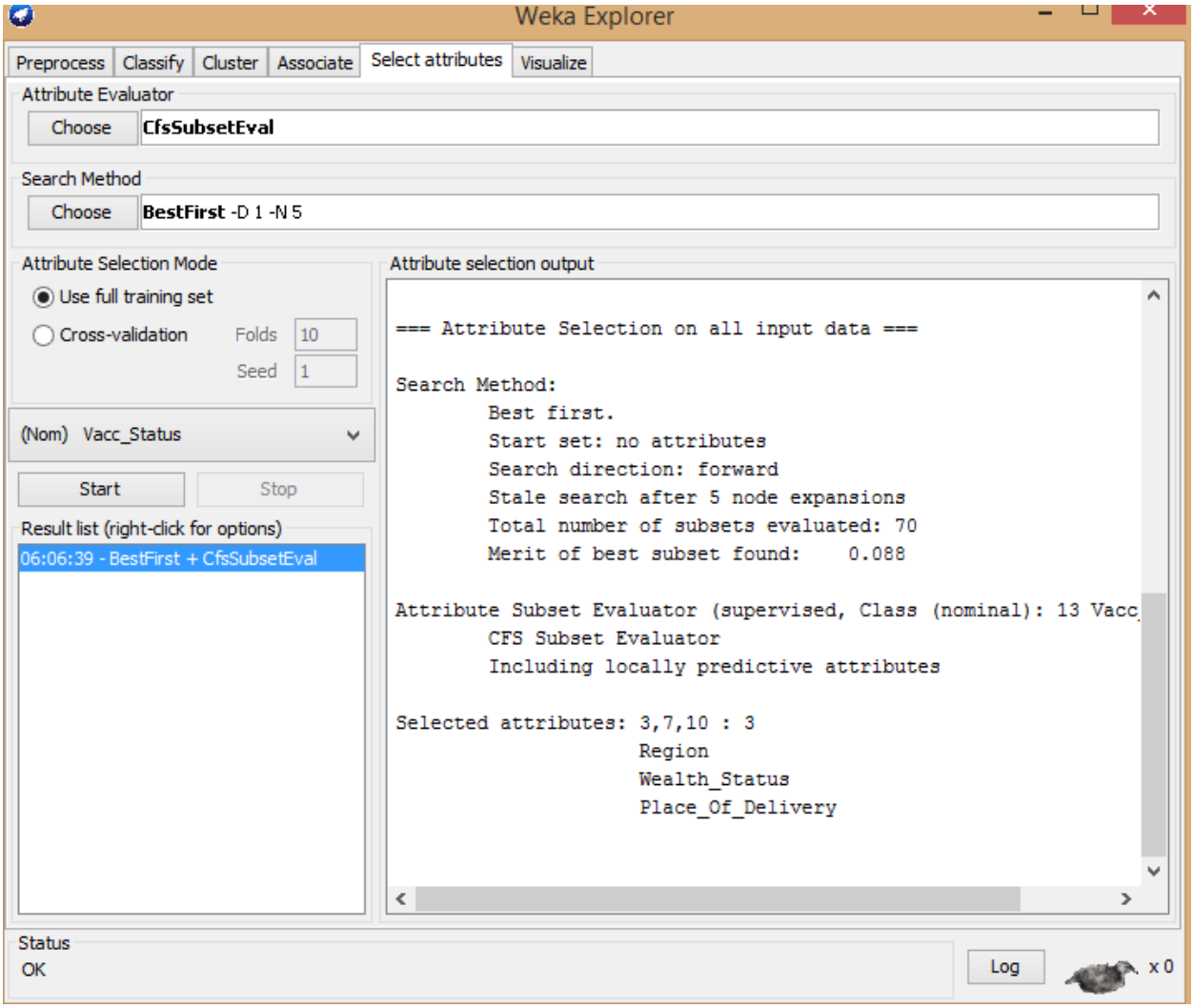
91. H. Yu, J. Han, and K. C. Chang. PEBL: Positive-example based learning for Web page classification using SVM. In Proc. 8th Int. Conf. Knowledge Discovery and Data Mining, Edmonton, Canada, 2002.
92. Henok Tadesse, Amare Deribew and Mirkuzie Woldie. “Explorative assessment of factors affecting child immunization in Wonago district, Gedeo zone, South Ethiopia” Termedia & Banach 2009.

Appendix 1: Dataset Sample with CSV (comma delimited) File Format



	E	F	G	H	I	J	K	L	M	N
1	Listenir	Child_S	Wealth_	Marital_	Parity	Place_C	Distanc	Father_	Vacc_Status	
2	Not_At_A	Male	Poor	Married	3-Feb	Home	Big_Probl	No_Educa	Unimmunised	
3	Not_At_A	Female	Middle	Married	7-Jun	Home	Big_Probl	No_Educa	Partial	
4	Not_At_A	Female	Rich	Married	1	Home	Not_A_Big	Higher	Partial	
5	Not_At_A	Male	Rich	Married	3-Feb	Home	Not_A_Big	Primary	Partial	
6	At_Least_	Female	Rich	Married	3-Feb	Governme	Big_Probl	Primary	Full	
7	At_Least_	Female	Rich	Married	1	Home	Big_Probl	Dont_knc	Partial	
8	At_Least_	Female	Middle	Married	9-Aug	Home	Big_Probl	Dont_knc	Full	
9	At_Least_	Female	Rich	Married	5-Apr	Home	Big_Probl	Dont_knc	Full	
10	At_Least_	Female	Rich	Married	3-Feb	Home	Big_Probl	Dont_knc	Full	
11	At_Least_	Female	Rich	Married	1	Governme	Big_Probl	Higher	Partial	
12	At_Least_	Male	Rich	Married	1	Home	Big_Probl	Higher	Partial	
13	At_Least_	Female	Rich	Married	1	Home	Big_Probl	Higher	Partial	
14	At_Least_	Female	Rich	Married	1	Home	Big_Probl	Higher	Partial	
15	At_Least_	Female	Rich	Widowed	5-Apr	Home	Big_Probl	Higher	Partial	
16	At_Least_	Female	Rich	Married	1	Governme	Big_Probl	Higher	Partial	
17	At_Least_	Female	Rich	Married	1	Governme	Big_Probl	Higher	Partial	
18	At_Least_	Female	Rich	Married	3-Feb	Home	Big_Probl	Higher	Partial	

Appendix 2: Result of CFS Attributes Subset Evaluator



Appendix 3: Rules generated by J48 decision tree

=== Run information ===

Scheme:weka.classifiers.trees.J48 -C 0.25 -M 2

Relation: test210

Instances: 8210

Attributes: 13

Mother_Age

Residence

Region

Mother_Education_level

Listening_To_Radio

Child_Sex

Wealth_Status

Marital_Status

Parity

Place_Of_Delivery

Distance_to_health_facility

Father_Education

Vacc_Status

Test mode: 10-fold cross-validation

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

J48 pruned tree

Place_Of_Delivery = Home

| Region = Amhara

| | Listening_To_Radio = Not_At_All: Partial (537.0/160.0)

| | Listening_To_Radio = At_ Least_Once_A_Week

| | | Mother_Age = 25-29

| | | | Wealth_Status = Poor

| | | | | Distance_to_health_facility = Big_Problem: Full (6.0/2.0)

| | | | | Distance_to_health_facility = Not_A_Big_Problem: Partial (2.0)

| | | | Wealth_Status = Middle: Partial (3.0)

| | | | Wealth_Status = Rich: Full (13.0/4.0)

| | | Mother_Age = 40-44: Partial (11.0/5.0)

| | | Mother_Age = 15-19

| | | | Mother_Education_level = No_Education: Partial (3.0)

| | | | Mother_Education_level = Primary: Full (3.0/1.0)

| | | | Mother_Education_level = Secondary: Partial (0.0)

| | | | Mother_Education_level = Higher: Partial (0.0)

| | | Mother_Age = 30-34

| | | | Father_Education = No_Education

| | | | | Wealth_Status = Poor: Full (4.0/1.0)

| | | | | Wealth_Status = Middle: Partial (4.0)

| | | | | Wealth_Status = Rich: Full (3.0/1.0)

| | | | Father_Education = Higher: Full (0.0)

| | | | Father_Education = Primary: Full (6.0)

| | | | Father_Education = Dont_ know: Full (0.0)

| | | | Father_Education = Secondary: Full (0.0)

| | | Mother_Age = 20-24: Partial (22.0/8.0)

| | | Mother_Age = 35-39

| | | | Distance_to_health_facility = Big_Problem: Partial (19.0/1.0)

| | | | Distance_to_health_facility = Not_A_Big_Problem: Full (5.0/1.0)

| | | Mother_Age = 45-49: Full (5.0/1.0)

| | Listening_To_Radio = Less_Than_Once_A_Week: Partial (227.0/75.0)

| Region = SNNP: Partial (1077.0/374.0)

| Region = Harari: Partial (332.0/127.0)

| Region = Oromiya: Partial (1113.0/378.0)

| Region = Benishangul_Gumuz: Partial (599.0/208.0)

| Region = Tigray

| | Mother_Age = 25-29

| | | Marital_Status = Married: Full (180.0/73.0)

| | | Marital_Status = Widowed: Full (1.0)

| | | Marital_Status = Single: Partial (2.0/1.0)

| | | Marital_Status = Divorced

| | | | Child_Sex = Male: Full (2.0)

| | | | Child_Sex = Female: Partial (6.0/1.0)

| | Mother_Age = 40-44: Full (75.0/22.0)

| | Mother_Age = 15-19

| | | Father_Education = No_Education

| | | | Listening_To_Radio = Not_At_All: Partial (7.0/2.0)

| | | | Listening_To_Radio = At_ Least_Once_A_Week: Partial (4.0/2.0)

| | | | Listening_To_Radio = Less_Than_Once_A_Week: Full (7.0/1.0)

| | | Father_Education = Higher: Full (1.0)

| | | Father_Education = Primary: Partial (12.0/2.0)

| | | Father_Education = Dont_ know: Partial (0.0)

| | | Father_Education = Secondary: Partial (0.0)

| | Mother_Age = 30-34

| | | Listening_To_Radio = Not_At_All

| | | | Parity = 3-Feb: Full (21.0/7.0)

| | | | Parity = 7-8

| | | | | Mother_Education_level = No_Education

| | | | | | Child_Sex = Male

| | | | | | | Distance_to_health_facility = Big_Problem: Full (10.0/3.0)

| | | | | | | Distance_to_health_facility = Not_A_Big_Problem: Partial (2.0)

| | | | | | Child_Sex = Female: Partial (6.0/2.0)

| | | | | Mother_Education_level = Primary: Partial (4.0)

| | | | | Mother_Education_level = Secondary: Partial (0.0)

| | | | | Mother_Education_level = Higher: Partial (0.0)

| | | | Parity = 1.0: Partial (2.0/1.0)

| | | | Parity = 9-10: Partial (4.0/1.0)

| | | | Parity = 5-6

| | | | | Marital_Status = Married

| | | | | | Mother_Education_level = No_Education: Partial (36.0/15.0)

| | | | | | Mother_Education_level = Primary: Full (10.0/3.0)

| | | | | | Mother_Education_level = Secondary: Partial (0.0)

| | | | | | Mother_Education_level = Higher: Partial (0.0)

| | | | | Marital_Status = Widowed: Full (1.0)

| | | | | Marital_Status = Single: Full (0.0)

| | | | | Marital_Status = Divorced: Full (4.0/1.0)

| | | | Parity = 11-12: Full (0.0)

| | | | Parity = 13-14: Full (0.0)

| | | Listening_To_Radio = At_ Least_Once_A_Week: Full (25.0/5.0)

| | | Listening_To_Radio = Less_Than_Once_A_Week

| | | | Marital_Status = Married

| | | | | Wealth_Status = Poor: Full (24.0/7.0)

| | | | | Wealth_Status = Middle: Full (4.0)

| | | | | Wealth_Status = Rich

| | | | | | Distance_to_health_facility = Big_Problem: Partial (4.0)

| | | | | | Distance_to_health_facility = Not_A_Big_Problem: Full (2.0)

| | | | Marital_Status = Widowed: Full (4.0/1.0)

| | | | Marital_Status = Single: Full (0.0)

| | | | Marital_Status = Divorced: Partial (3.0)

| | Mother_Age = 20-24

| | | Parity = 3-4

| | | | Child_Sex = Male

| | | | | Listening_To_Radio = Not_At_All

| | | | | | Wealth_Status = Poor: Full (15.0/5.0)

| | | | | | Wealth_Status = Middle: Full (5.0/1.0)

| | | | | | Wealth_Status = Rich: Partial (5.0/1.0)

| | | | | Listening_To_Radio = At_ Least_Once_A_Week: Full (4.0)

| | | | | Listening_To_Radio = Less_Than_Once_A_Week

| | | | | | Father_Education = No_Education: Partial (6.0)

| | | | | | Father_Education = Higher: Partial (0.0)

| | | | | | Father_Education = Primary: Full (3.0/1.0)

| | | | | | Father_Education = Dont_ know: Partial (0.0)

| | | | | | Father_Education = Secondary: Partial (0.0)

| | | | Child_Sex = Female

| | | | | Father_Education = No_Education

| | | | | | Mother_Education_level = No_Education: Full (9.0/1.0)

| | | | | | Mother_Education_level = Primary

| | | | | | | Distance_to_health_facility = Big_Problem: Partial (4.0/1.0)

| | | | | | | Distance_to_health_facility = Not_A_Big_Problem: Full (2.0)

| | | | | | Mother_Education_level = Secondary: Full (0.0)

| | | | | | Mother_Education_level = Higher: Full (0.0)

| | | | | Father_Education = Higher: Full (0.0)

| | | | | Father_Education = Primary: Partial (12.0/5.0)

| | | | | Father_Education = Dont_ know: Full (0.0)

| | | | | Father_Education = Secondary: Partial (2.0/1.0)

| | | Parity = 7-8: Full (0.0)

| | | Parity = 1.0: Full (84.0/31.0)

| | | Parity = 9-10: Full (0.0)

| | | Parity = 5-6: Partial (7.0/3.0)

| | | Parity = 11-12: Full (0.0)

| | | Parity = 13-14: Full (0.0)

| | Mother_Age = 35-39

| | | Mother_Education_level = No_Education

| | | | Wealth_Status = Poor

| | | | | Marital_Status = Married: Full (70.0/19.0)

| | | | | Marital_Status = Widowed: Partial (4.0/2.0)

| | | | | Marital_Status = Single: Full (0.0)

| | | | | Marital_Status = Divorced

| | | | | | Father_Education = No_Education: Full (9.0/2.0)

| | | | | | Father_Education = Higher: Full (0.0)

| | | | | | Father_Education = Primary: Partial (4.0)

| | | | | | Father_Education = Dont_ know: Full (0.0)

| | | | | | Father_Education = Secondary: Full (0.0)

| | | | Wealth_Status = Middle

| | | | | Parity = 3-4: Full (0.0)

| | | | | Parity = 7-8: Full (10.0)

| | | | | Parity = 1.0: Full (0.0)

| | | | | Parity = 9-10

| | | | | | Father_Education = No_Education: Partial (3.0)

| | | | | | Father_Education = Higher: Partial (0.0)

| | | | | | Father_Education = Primary: Full (3.0/1.0)

| | | | | | Father_Education = Dont_ know: Partial (0.0)

| | | | | | Father_Education = Secondary: Partial (0.0)

| | | | | Parity = 5-6

| | | | | | Father_Education = No_Education: Partial (6.0/1.0)

| | | | | | Father_Education = Higher: Partial (0.0)

| | | | | | Father_Education = Primary: Full (2.0)

| | | | | | Father_Education = Dont_ know: Partial (0.0)

| | | | | | Father_Education = Secondary: Partial (0.0)

| | | | | Parity = 11-12: Full (0.0)

| | | | | Parity = 13-14: Full (0.0)

| | | | Wealth_Status = Rich: Full (25.0/3.0)

| | | Mother_Education_level = Primary

| | | | Parity = 3-4: Full (3.0)

| | | | Parity = 7-8: Full (10.0/2.0)

| | | | Parity = 1.0: Partial (1.0)

| | | | Parity = 9-10: Partial (2.0)

| | | | Parity = 5-6

| | | | | Listening_To_Radio = Not_At_All: Partial (4.0)

| | | | | Listening_To_Radio = At_ Least_Once_A_Week: Full (2.0)

| | | | | Listening_To_Radio = Less_Than_Once_A_Week: Partial (4.0/1.0)

| | | | Parity = 11-12: Full (0.0)

| | | | Parity = 13-14: Partial (2.0/1.0)

| | | Mother_Education_level = Secondary: Full (1.0)

| | | Mother_Education_level = Higher: Full (0.0)

| | Mother_Age = 45-49

| | | Listening_To_Radio = Not_At_All

| | | | Child_Sex = Male: Partial (6.0/1.0)

| | | | Child_Sex = Female: Full (5.0/1.0)

| | | Listening_To_Radio = At_ Least_Once_A_Week: Partial (1.0)

| | | Listening_To_Radio = Less_Than_Once_A_Week: Full (7.0/2.0)

| Region = Affar

| | Mother_Education_level = No_Education

| | | Wealth_Status = Poor

| | | | Listening_To_Radio = Not_At_All

| | | | | Mother_Age = 25-29

| | | | | | Parity = 3-4: Partial (54.0/25.0)

| | | | | | Parity = 7-8 Unimmunised (10.0/1.0)

| | | | | | Parity = 1.0: Partial (19.0/7.0)

| | | | | | Parity = 9-10: Unimmunised (3.0)

| | | | | | Parity = 5-6

| | | | | | | Distance_to_health_facility = Big_Problem: Unimmunised (39.0/11.0)

| | | | | | | Distance_to_health_facility = Not_A_Big_Problem: Partial (3.0)

| | | | | | Parity = 11-12: Unimmunised (0.0)

| | | | | | Parity = 13-14: Unimmunised (0.0)

| | | | | Mother_Age = 40-44: Unimmunised (31.0/12.0)

| | | | | Mother_Age = 15-19: Partial (5.0/1.0)

| | | | | Mother_Age = 30-34: Unimmunised (87.0/34.0)

| | | | | Mother_Age = 20-24: Unimmunised (81.0/30.0)

| | | | | Mother_Age = 35-39

| | | | | | Marital_Status = Married

| | | | | | | Distance_to_health_facility = Big_Problem: Unimmunised (57.0/20.0)

| | | | | | | Distance_to_health_facility = Not_A_Big_Problem: Partial (4.0/1.0)

| | | | | | Marital_Status = Widowed: Partial (3.0)

| | | | | | Marital_Status = Single: Unimmunised (1.0)

| | | | | | Marital_Status = Divorced: Unimmunised (1.0)

| | | | | Mother_Age = 45-49: Partial (15.0/7.0)

| | | | Listening_To_Radio = At_ Least_Once_A_Week

| | | | | Distance_to_health_facility = Big_Problem: Partial (49.0/10.0)

| | | | | Distance_to_health_facility = Not_A_Big_Problem: Unimmunised (2.0)

| | | | Listening_To_Radio = Less_Than_Once_A_Week

| | | | | Father_Education = No_Education

| | | | | | Mother_Age = 25-29

| | | | | | | Parity = 3-Feb: Partial (17.0/6.0)

| | | | | | | Parity = 7-Jun: Unimmunised (7.0/2.0)

| | | | | | | Parity = 1.0: Partial (4.0/1.0)

| | | | | | | Parity = 9-Aug: Partial (0.0)

| | | | | | | Parity = 5-Apr: Unimmunised (13.0/5.0)

| | | | | | | Parity = 11-Oct: Partial (0.0)

| | | | | | | Parity = 13-Dec: Partial (0.0)

| | | | | | Mother_Age = 40-44: Unimmunised (11.0/2.0)

| | | | | | Mother_Age = 15-19: Partial (1.0)

| | | | | | Mother_Age = 30-34: Unimmunised (31.0/11.0)

| | | | | | Mother_Age = 20-24: Unimmunised (21.0/7.0)

| | | | | | Mother_Age = 35-39: Partial (22.0/9.0)

| | | | | | Mother_Age = 45-49: Partial (6.0/1.0)

| | | | | Father_Education = Higher: Unimmunised (2.0)

| | | | | Father_Education = Primary: Partial (17.0/6.0)

| | | | | Father_Education = Dont_ know: Unimmunised (3.0/1.0)

| | | | | Father_Education = Secondary: Full (1.0)

| | | Wealth_Status = Middle

| | | | Listening_To_Radio = Not_At_All

| | | | | Parity = 3-Feb: Partial (5.0/1.0)

| | | | | Parity = 7-Jun: Partial (2.0)

| | | | | Parity = 1.0: Unimmunised (2.0/1.0)

| | | | | Parity = 9-Aug: Unimmunised (1.0)

| | | | | Parity = 5-Apr: Unimmunised (4.0)

| | | | | Parity = 11-Oct: Unimmunised (2.0)

| | | | | Parity = 13-Dec: Unimmunised (0.0)

| | | | Listening_To_Radio = At_ Least_Once_A_Week

| | | | | Parity = 3-Feb: Partial (6.0)

| | | | | Parity = 7-Jun: Partial (0.0)

| | | | | Parity = 1.0: Unimmunised (2.0/1.0)

| | | | | Parity = 9-Aug: Partial (0.0)

| | | | | Parity = 5-Apr: Unimmunised (3.0/1.0)

| | | | | Parity = 11-Oct: Partial (0.0)

| | | | | Parity = 13-Dec: Partial (0.0)

| | | | Listening_To_Radio = Less_Than_Once_A_Week

| | | | | Parity = 3-4: Full (5.0/1.0)

| | | | | Parity = 7-Jun: Partial (1.0)

| | | | | Parity = 1.0: Partial (0.0)

| | | | | Parity = 9-Aug: Partial (0.0)

| | | | | Parity = 5-Apr: Partial (3.0)

| | | | | Parity = 11-Oct: Partial (1.0)

| | | | | Parity = 13-Dec: Partial (0.0)

| | | Wealth_Status = Rich

| | | | Father_Education = No_Education

| | | | | Listening_To_Radio = Not_At_All: Unimmunised (17.0/7.0)

| | | | | Listening_To_Radio = At_ Least_Once_A_Week: Partial (7.0)

| | | | | Listening_To_Radio = Less_Than_Once_A_Week: Partial (7.0/1.0)

| | | | Father_Education = Higher: Partial (0.0)

| | | | Father_Education = Primary

| | | | | Residence = Rural: Unimmunised (10.0/4.0)

| | | | | Residence = Urban

| | | | | | Listening_To_Radio = Not_At_All: Partial (5.0)

| | | | | | Listening_To_Radio = At_ Least_Once_A_Week: Full (2.0)

| | | | | | Listening_To_Radio = Less_Than_Once_A_Week: Partial (5.0/2.0)

| | | | Father_Education = Dont_ know: Partial (0.0)

| | | | Father_Education = Secondary: Unimmunised (7.0/2.0)

| | Mother_Education_level = Primary: Partial (31.0/10.0)

| | Mother_Education_level = Secondary: Partial (8.0/4.0)

| | Mother_Education_level = Higher: Partial (2.0/1.0)

| Region = Gambela: Partial (480.0/178.0)

| Region = Addis_Ababa: Full (44.0/14.0)

| Region = Somali

| | Wealth_Status = Poor

| | | Mother_Education_level = No_Education

| | | | Listening_To_Radio = Not_At_All

| | | | | Father_Education = No_Education: Unimmunised (214.0/83.0)

| | | | | Father_Education = Higher: Partial (1.0)

| | | | | Father_Education = Primary

| | | | | | Distance_to_health_facility = Big_Problem: Partial (32.0/12.0)

| | | | | | Distance_to_health_facility = Not_A_Big_Problem: Unimmunised (11.0/1.0)

| | | | | Father_Education = Dont_ know: Unimmunised (5.0)

| | | | | Father_Education = Secondary: Unimmunised (2.0/1.0)

| | | | Listening_To_Radio = At_ Least_Once_A_Week

| | | | | Distance_to_health_facility = Big_Problem

| | | | | | Mother_Age = 25-29: Partial (6.0)

| | | | | | Mother_Age = 40-44: Partial (1.0)

| | | | | | Mother_Age = 15-19: Unimmunised (2.0)

| | | | | | Mother_Age = 30-34: Partial (3.0/1.0)

| | | | | | Mother_Age = 20-24: Unimmunised (6.0/2.0)

| | | | | | Mother_Age = 35-39: Unimmunised (1.0)

| | | | | | Mother_Age = 45-49: Partial (1.0)

| | | | | Distance_to_health_facility = Not_A_Big_Problem: Unimmunised (6.0)

| | | | Listening_To_Radio = Less_Than_Once_A_Week

| | | | | Mother_Age = 25-29: Partial (16.0/4.0)

| | | | | Mother_Age = 40-44

| | | | | | Child_Sex = Male: Unimmunised (5.0/2.0)

| | | | | | Child_Sex = Female: Partial (3.0)

| | | | | Mother_Age = 15-19

| | | | | | Parity = 3-Feb: Unimmunised (2.0)

| | | | | | Parity = 7-Jun: Unimmunised (0.0)

| | | | | | Parity = 1.0: Partial (2.0)

| | | | | | Parity = 9-Aug: Unimmunised (0.0)

| | | | | | Parity = 5-Apr: Unimmunised (0.0)

| | | | | | Parity = 11-Oct: Unimmunised (0.0)

| | | | | | Parity = 13-Dec: Unimmunised (0.0)

| | | | | Mother_Age = 30-34: Partial (8.0/1.0)

| | | | | Mother_Age = 20-24

| | | | | | Parity = 3-Feb: Partial (6.0)

| | | | | | Parity = 7-Jun: Partial (0.0)

| | | | | | Parity = 1.0: Unimmunised (4.0)

| | | | | | Parity = 9-Aug: Partial (0.0)

| | | | | | Parity = 5-Apr: Partial (1.0)

| | | | | | Parity = 11-Oct: Partial (0.0)

| | | | | | Parity = 13-Dec: Partial (0.0)

| | | | | Mother_Age = 35-39: Unimmunised (11.0)

| | | | | Mother_Age = 45-49: Partial (3.0)

| | | Mother_Education_level = Primary

| | | | Marital_Status = Married: Partial (25.0/7.0)

| | | | Marital_Status = Widowed: Partial (0.0)

| | | | Marital_Status = Single: Partial (0.0)

| | | | Marital_Status = Divorced: Unimmunised (2.0)

| | | Mother_Education_level = Secondary: Unimmunised (0.0)

| | | Mother_Education_level = Higher: Unimmunised (0.0)

| | Wealth_Status = Middle

| | | Distance_to_health_facility = Big_Problem

| | | | Father_Education = No_Education

| | | | | Listening_To_Radio = Not_At_All: Partial (28.0/12.0)

| | | | | Listening_To_Radio = At_ Least_Once_A_Week: Partial (2.0)

| | | | | Listening_To_Radio = Less_Than_Once_A_Week: Unimmunised (3.0/1.0)

| | | | Father_Education = Higher: Unimmunised (1.0)

| | | | Father_Education = Primary

| | | | | Listening_To_Radio = Not_At_All: Unimmunised (4.0)

| | | | | Listening_To_Radio = At_ Least_Once_A_Week: Unimmunised (0.0)

| | | | | Listening_To_Radio = Less_Than_Once_A_Week: Partial (3.0)

| | | | Father_Education = Dont_ know

| | | | | Listening_To_Radio = Not_At_All: Partial (3.0)

| | | | | Listening_To_Radio = At_ Least_Once_A_Week: Partial (0.0)

| | | | | Listening_To_Radio = Less_Than_Once_A_Week: Unimmunised (2.0)

| | | | Father_Education = Secondary: Unimmunised (1.0)

| | | Distance_to_health_facility = Not_A_Big_Problem: Partial (13.0/6.0)

| | Wealth_Status = Rich

| | | Residence = Rural

| | | | Listening_To_Radio = Not_At_All: Partial (38.0/14.0)

| | | | Listening_To_Radio = At_ Least_Once_A_Week: Partial (17.0/4.0)

| | | | Listening_To_Radio = Less_Than_Once_A_Week

| | | | | Marital_Status = Married

| | | | | | Parity = 3-Feb

| | | | | | | Father_Education = No_Education: Full (9.0/3.0)

| | | | | | | Father_Education = Higher: Full (0.0)

| | | | | | | Father_Education = Primary: Unimmunised (3.0/1.0)

| | | | | | | Father_Education = Dont_ know: Full (0.0)

| | | | | | | Father_Education = Secondary: Full (0.0)

| | | | | | Parity = 7-Jun: Partial (8.0/3.0)

| | | | | | Parity = 1.0: Full (1.0)

| | | | | | Parity = 9-Aug: Partial (2.0)

| | | | | | Parity = 5-Apr

| | | | | | | Father_Education = No_Education: Unimmunised (6.0/1.0)

| | | | | | | Father_Education = Higher: Unimmunised (0.0)

| | | | | | | Father_Education = Primary: Full (2.0)

| | | | | | | Father_Education = Dont_ know: Partial (2.0/1.0)

| | | | | | | Father_Education = Secondary: Unimmunised (0.0)

| | | | | | Parity = 11-Oct: Full (1.0)

| | | | | | Parity = 13-Dec: Full (0.0)

| | | | | Marital_Status = Widowed: Partial (0.0)

| | | | | Marital_Status = Single: Partial (0.0)

| | | | | Marital_Status = Divorced: Partial (3.0)

| | | Residence = Urban

| | | | Father_Education = No_Education

| | | | | Mother_Age = 25-29

| | | | | | Mother_Education_level = No_Education: Unimmunised (12.0/1.0)

| | | | | | Mother_Education_level = Primary: Full (4.0/1.0)

| | | | | | Mother_Education_level = Secondary: Unimmunised (0.0)

| | | | | | Mother_Education_level = Higher: Unimmunised (0.0)

| | | | | Mother_Age = 40-44: Partial (9.0/1.0)

| | | | | Mother_Age = 15-19: Unimmunised (2.0)

| | | | | Mother_Age = 30-34

| | | | | | Child_Sex = Male

| | | | | | | Listening_To_Radio = Not_At_All: Unimmunised (5.0)

| | | | | | | Listening_To_Radio = At_ Least_Once_A_Week: Unimmunised (0.0)

| | | | | | | Listening_To_Radio = Less_Than_Once_A_Week: Full (3.0/1.0)

| | | | | | Child_Sex = Female: Unimmunised (8.0/5.0)

| | | | | Mother_Age = 20-24: Unimmunised (1.0)

| | | | | Mother_Age = 35-39

| | | | | | Listening_To_Radio = Not_At_All

| | | | | | | Parity = 3-Feb: Unimmunised (1.0)

| | | | | | | Parity = 7-Jun: Full (2.0)

| | | | | | | Parity = 1.0: Unimmunised (0.0)

| | | | | | | Parity = 9-Aug: Unimmunised (0.0)

| | | | | | | Parity = 5-Apr: Unimmunised (2.0)

| | | | | | | Parity = 11-Oct: Full (1.0)

| | | | | | | Parity = 13-Dec: Unimmunised (1.0)

| | | | | | Listening_To_Radio = At_ Least_Once_A_Week: Partial (3.0/1.0)

| | | | | | Listening_To_Radio = Less_Than_Once_A_Week: Unimmunised (0.0)

| | | | | Mother_Age = 45-49: Unimmunised (0.0)

| | | | Father_Education = Higher: Partial (6.0/2.0)

| | | | Father_Education = Primary: Partial (32.0/10.0)

| | | | Father_Education = Dont_ know: Partial (2.0/1.0)

| | | | Father_Education = Secondary

| | | | | Marital_Status = Married: Partial (9.0/3.0)

| | | | | Marital_Status = Widowed: Partial (0.0)

| | | | | Marital_Status = Single: Partial (0.0)

| | | | | Marital_Status = Divorced: Unimmunised (2.0)

| Region = Dire_Dawa

| | Father_Education = No_Education

| | | Mother_Age = 25-29

| | | | Child_Sex = Male

| | | | | Wealth_Status = Poor: Partial (17.0/7.0)

| | | | | Wealth_Status = Middle: Partial (12.0/2.0)

| | | | | Wealth_Status = Rich: Full (7.0/2.0)

| | | | Child_Sex = Female

| | | | | Distance_to_health_facility = Big_Problem: Full (28.0/10.0)

| | | | | Distance_to_health_facility = Not_A_Big_Problem: Partial (9.0/3.0)

| | | Mother_Age = 40-44: Partial (7.0/3.0)

| | | Mother_Age = 15-19: Unimmunised (1.0)

| | | Mother_Age = 30-34

| | | | Mother_Education_level = No_Education

| | | | | Residence = Rural: Full (47.0/16.0)

| | | | | Residence = Urban: Partial (6.0/1.0)

| | | | Mother_Education_level = Primary: Partial (4.0)

| | | | Mother_Education_level = Secondary: Full (0.0)

| | | | Mother_Education_level = Higher: Full (0.0)

| | | Mother_Age = 20-24

| | | | Mother_Education_level = No_Education: Partial (17.0/5.0)

| | | | Mother_Education_level = Primary: Full (6.0/1.0)

| | | | Mother_Education_level = Secondary: Partial (0.0)

| | | | Mother_Education_level = Higher: Partial (0.0)

| | | Mother_Age = 35-39

| | | | Parity = 3-Feb: Full (6.0/2.0)

| | | | Parity = 7-Jun

| | | | | Listening_To_Radio = Not_At_All: Partial (17.0/3.0)

| | | | | Listening_To_Radio = At_ Least_Once_A_Week: Partial (1.0)

| | | | | Listening_To_Radio = Less_Than_Once_A_Week: Full (2.0)

| | | | Parity = 1.0: Partial (1.0)

| | | | Parity = 9-Aug

| | | | | Listening_To_Radio = Not_At_All: Full (8.0/2.0)

| | | | | Listening_To_Radio = At_ Least_Once_A_Week: Partial (1.0)

| | | | | Listening_To_Radio = Less_Than_Once_A_Week: Partial (2.0)

| | | | Parity = 5-Apr: Partial (14.0/3.0)

| | | | Parity = 11-Oct: Full (1.0)

| | | | Parity = 13-Dec: Partial (0.0)

| | | Mother_Age = 45-49: Partial (5.0/1.0)

| | Father_Education = Higher: Partial (1.0)

| | Father_Education = Primary

| | | Residence = Rural

| | | | Mother_Age = 25-29

| | | | | Listening_To_Radio = Not_At_All: Partial (17.0/6.0)

| | | | | Listening_To_Radio = At_ Least_Once_A_Week: Full (2.0)

| | | | | Listening_To_Radio = Less_Than_Once_A_Week

| | | | | | Wealth_Status = Poor: Full (1.0)

| | | | | | Wealth_Status = Middle: Partial (4.0/1.0)

| | | | | | Wealth_Status = Rich: Full (4.0)

| | | | Mother_Age = 40-44: Partial (4.0/2.0)

| | | | Mother_Age = 15-19: Partial (5.0/2.0)

| | | | Mother_Age = 30-34

| | | | | Wealth_Status = Poor: Full (7.0/3.0)

| | | | | Wealth_Status = Middle

| | | | | | Listening_To_Radio = Not_At_All: Full (4.0/1.0)

| | | | | | Listening_To_Radio = At_ Least_Once_A_Week: Full (2.0)

| | | | | | Listening_To_Radio = Less_Than_Once_A_Week: Partial (2.0)

| | | | | Wealth_Status = Rich: Partial (5.0)

| | | | Mother_Age = 20-24: Full (26.0/9.0)

| | | | Mother_Age = 35-39

| | | | | Wealth_Status = Poor: Partial (2.0)

| | | | | Wealth_Status = Middle: Full (2.0)

| | | | | Wealth_Status = Rich: Partial (3.0)

| | | | Mother_Age = 45-49: Full (0.0)

| | | Residence = Urban

| | | | Listening_To_Radio = Not_At_All

| | | | | Mother_Education_level = No_Education: Partial (6.0/2.0)

| | | | | Mother_Education_level = Primary: Unimmunised (3.0/1.0)

| | | | | Mother_Education_level = Secondary: Partial (0.0)

| | | | | Mother_Education_level = Higher: Partial (0.0)

| | | | Listening_To_Radio = At_ Least_Once_A_Week: Full (4.0)

| | | | Listening_To_Radio = Less_Than_Once_A_Week: Full (2.0)

| | Father_Education = Dont_ know: Partial (2.0)

| | Father_Education = Secondary: Full (7.0/1.0)

Place_Of_Delivery = Government_Health_Facility

| Residence = Rural

| | Region = Amhara

| | | Parity = 3-Feb

| | | | Distance_to_health_facility = Big_Problem: Partial (4.0/1.0)

| | | | Distance_to_health_facility = Not_A_Big_Problem

| | | | | Child_Sex = Male: Full (3.0)

| | | | | Child_Sex = Female: Partial (3.0/1.0)

| | | Parity = 7-Jun: Partial (6.0/1.0)

| | | Parity = 1.0: Partial (21.0/5.0)

| | | Parity = 9-Aug: Unimmunised (1.0)

| | | Parity = 5-Apr

| | | | Mother_Age = 25-29: Full (2.0)

| | | | Mother_Age = 40-44: Partial (0.0)

| | | | Mother_Age = 15-19: Partial (0.0)

| | | | Mother_Age = 30-34: Partial (5.0)

| | | | Mother_Age = 20-24: Unimmunised (1.0)

| | | | Mother_Age = 35-39: Partial (0.0)

| | | | Mother_Age = 45-49: Partial (0.0)

| | | Parity = 11-Oct: Full (2.0)

| | | Parity = 13-Dec: Partial (0.0)

| | Region = SNNP

| | | Father_Education = No_Education: Full (7.0/1.0)

| | | Father_Education = Higher: Partial (3.0/1.0)

| | | Father_Education = Primary

| | | | Listening_To_Radio = Not_At_All

| | | | | Mother_Education_level = No_Education: Partial (2.0)

| | | | | Mother_Education_level = Primary: Full (2.0)

| | | | | Mother_Education_level = Secondary: Partial (0.0)

| | | | | Mother_Education_level = Higher: Partial (0.0)

| | | | Listening_To_Radio = At_ Least_Once_A_Week: Full (1.0)

| | | | Listening_To_Radio = Less_Than_Once_A_Week: Partial (5.0)

| | | Father_Education = Dont_ know: Partial (1.0)

| | | Father_Education = Secondary: Partial (1.0)

| | Region = Harari: Partial (22.0/6.0)

| | Region = Oromiya: Partial (36.0/14.0)

| | Region = Benishangul_Gumuz: Partial (20.0/5.0)

| | Region = Tigray

| | | Listening_To_Radio = Not_At_All

| | | | Child_Sex = Male: Full (5.0)

| | | | Child_Sex = Female

| | | | | Father_Education = No_Education: Full (3.0)

| | | | | Father_Education = Higher: Partial (2.0)

| | | | | Father_Education = Primary: Partial (3.0)

| | | | | Father_Education = Dont_ know: Partial (0.0)

| | | | | Father_Education = Secondary: Partial (0.0)

| | | Listening_To_Radio = At_ Least_Once_A_Week: Full (12.0)

| | | Listening_To_Radio = Less_Than_Once_A_Week

| | | | Father_Education = No_Education

| | | | | Mother_Education_level = No_Education: Unimmunised (2.0/1.0)

| | | | | Mother_Education_level = Primary: Partial (2.0)

| | | | | Mother_Education_level = Secondary: Partial (0.0)

| | | | | Mother_Education_level = Higher: Partial (0.0)

| | | | Father_Education = Higher: Partial (0.0)

| | | | Father_Education = Primary: Full (2.0)

| | | | Father_Education = Dont_ know: Partial (0.0)

| | | | Father_Education = Secondary: Partial (1.0)

| | Region = Affar: Partial (9.0/2.0)

| | Region = Gambela: Partial (38.0/10.0)

| | Region = Addis_Ababa: Partial (0.0)

| | Region = Somali

| | | Wealth_Status = Poor

| | | | Marital_Status = Married: Unimmunised (4.0/1.0)

| | | | Marital_Status = Widowed: Partial (2.0)

| | | | Marital_Status = Single: Unimmunised (0.0)

| | | | Marital_Status = Divorced: Unimmunised (0.0)

| | | Wealth_Status = Middle: Partial (0.0)

| | | Wealth_Status = Rich: Partial (7.0)

| | Region = Dire_Dawa: Full (8.0/2.0)

| Residence = Urban

| | Region = Amhara: Partial (26.0/7.0)

| | Region = SNNP: Full (25.0/9.0)

| | Region = Harari

| | | Distance_to_health_facility = Big_Problem

| | | | Marital_Status = Married

| | | | | Father_Education = No_Education: Partial (3.0)

| | | | | Father_Education = Higher: Full (2.0)

| | | | | Father_Education = Primary: Full (7.0/3.0)

| | | | | Father_Education = Dont_ know: Partial (0.0)

| | | | | Father_Education = Secondary: Partial (3.0)

| | | | Marital_Status = Widowed: Partial (0.0)

| | | | Marital_Status = Single: Partial (0.0)

| | | | Marital_Status = Divorced: Full (2.0)

| | | Distance_to_health_facility = Not_A_Big_Problem: Full (81.0/22.0)

| | Region = Oromiya

| | | Marital_Status = Married

| | | | Mother_Age = 25-29

| | | | | Parity = 3-Feb: Partial (7.0)

| | | | | Parity = 7-Jun: Partial (0.0)

| | | | | Parity = 1.0: Full (6.0/2.0)

| | | | | Parity = 9-Aug: Partial (0.0)

| | | | | Parity = 5-Apr: Full (2.0)

| | | | | Parity = 11-Oct: Partial (0.0)

| | | | | Parity = 13-Dec: Partial (0.0)

| | | | Mother_Age = 40-44: Partial (1.0)

| | | | Mother_Age = 15-19: Full (2.0)

| | | | Mother_Age = 30-34: Unimmunised (5.0/3.0)

| | | | Mother_Age = 20-24

| | | | | Parity = 3-Feb: Full (4.0/1.0)

| | | | | Parity = 7-Jun: Partial (0.0)

| | | | | Parity = 1.0: Partial (9.0/2.0)

| | | | | Parity = 9-Aug: Partial (0.0)

| | | | | Parity = 5-Apr: Partial (0.0)

| | | | | Parity = 11-Oct: Partial (0.0)

| | | | | Parity = 13-Dec: Partial (0.0)

| | | | Mother_Age = 35-39: Full (4.0)

| | | | Mother_Age = 45-49: Partial (0.0)

| | | Marital_Status = Widowed: Partial (0.0)

| | | Marital_Status = Single: Full (1.0)

| | | Marital_Status = Divorced: Partial (5.0)

| | Region = Benishangul_Gumuz: Partial (19.0/6.0)

| | Region = Tigray: Full (47.0/12.0)

| | Region = Affar: Partial (24.0/8.0)

| | Region = Gambela

| | | Marital_Status = Married

| | | | Child_Sex = Male: Partial (11.0/4.0)

| | | | Child_Sex = Female: Full (19.0/8.0)

| | | Marital_Status = Widowed: Partial (2.0)

| | | Marital_Status = Single: Partial (0.0)

| | | Marital_Status = Divorced: Full (1.0)

| | Region = Addis_Ababa: Full (178.0/30.0)

| | Region = Somali

| | | Father_Education = No_Education: Partial (6.0)

| | | Father_Education = Higher: Full (3.0)

| | | Father_Education = Primary

| | | | Distance_to_health_facility = Big_Problem: Unimmunised (2.0/1.0)

| | | | Distance_to_health_facility = Not_A_Big_Problem: Partial (5.0)

| | | Father_Education = Dont_ know: Partial (0.0)

| | | Father_Education = Secondary

| | | | Distance_to_health_facility = Big_Problem: Partial (3.0)

| | | | Distance_to_health_facility = Not_A_Big_Problem: Full (9.0/4.0)

| | Region = Dire_Dawa

| | | Mother_Education_level = No_Education

| | | | Parity = 3-Feb

| | | | | Distance_to_health_facility = Big_Problem: Partial (10.0/3.0)

| | | | | Distance_to_health_facility = Not_A_Big_Problem: Full (16.0/4.0)

| | | | Parity = 7-Jun: Partial (5.0/2.0)

| | | | Parity = 1.0

| | | | | Father_Education = No_Education: Partial (4.0/1.0)

| | | | | Father_Education = Higher: Partial (0.0)

| | | | | Father_Education = Primary: Full (5.0/1.0)

| | | | | Father_Education = Dont_ know: Partial (0.0)

| | | | | Father_Education = Secondary: Partial (6.0/1.0)

| | | | Parity = 9-Aug: Partial (2.0)

| | | | Parity = 5-Apr

| | | | | Distance_to_health_facility = Big_Problem: Full (8.0)

| | | | | Distance_to_health_facility = Not_A_Big_Problem

| | | | | | Child_Sex = Male: Full (5.0/1.0)

| | | | | | Child_Sex = Female: Partial (2.0)

| | | | Parity = 11-Oct: Partial (1.0)

| | | | Parity = 13-Dec: Full (0.0)

| | | Mother_Education_level = Primary: Full (47.0/8.0)

| | | Mother_Education_level = Secondary

| | | | Parity = 3-Feb: Full (8.0)

| | | | Parity = 7-Jun: Full (0.0)

| | | | Parity = 1.0: Partial (6.0/2.0)

| | | | Parity = 9-Aug: Full (0.0)

| | | | Parity = 5-Apr: Full (1.0)

| | | | Parity = 11-Oct: Full (0.0)

| | | | Parity = 13-Dec: Full (0.0)

| | | Mother_Education_level = Higher: Full (7.0/2.0)

Place_Of_Delivery = Other

| Mother_Age = 25-29

| | Father_Education = No_Education: Unimmunised (8.0/3.0)

| | Father_Education = Higher: Full (1.0)

| | Father_Education = Primary: Partial (11.0/5.0)

| | Father_Education = Dont_ know: Unimmunised (2.0)

| | Father_Education = Secondary: Unimmunised (0.0)

| Mother_Age = 40-44

| | Region = Amhara: Unimmunised (0.0)

| | Region = SNNP: Unimmunised (1.0)

| | Region = Harari: Unimmunised (0.0)

| | Region = Oromiya: Unimmunised (0.0)

| | Region = Benishangul_Gumuz: Partial (3.0)

| | Region = Tigray: Unimmunised (0.0)

| | Region = Affar: Unimmunised (2.0)

| | Region = Gambela: Unimmunised (0.0)

| | Region = Addis_Ababa: Unimmunised (0.0)

| | Region = Somali: Unimmunised (0.0)

| | Region = Dire_Dawa: Unimmunised (0.0)

| Mother_Age = 15-19: Unimmunised (5.0/1.0)

| Mother_Age = 30-34

| | Listening_To_Radio = Not_At_All: Unimmunised (12.0/2.0)

| | Listening_To_Radio = At_ Least_Once_A_Week: Partial (1.0)

| | Listening_To_Radio = Less_Than_Once_A_Week: Partial (5.0/2.0)

| Mother_Age = 20-24: Partial (19.0/4.0)

| Mother_Age = 35-39: Partial (18.0/6.0)

| Mother_Age = 45-49: Unimmunised (1.0)

Place_Of_Delivery = Private_Health_Facility

| Residence = Rural: Partial (24.0/6.0)

| Residence = Urban

| | Region = Amhara: Full (2.0)

| | Region = SNNP: Partial (3.0/1.0)

| | Region = Harari: Full (19.0/5.0)

| | Region = Oromiya

| | | Mother_Age = 25-29: Partial (3.0)

| | | Mother_Age = 40-44: Partial (0.0)

| | | Mother_Age = 15-19: Partial (0.0)

| | | Mother_Age = 30-34: Full (3.0/1.0)

| | | Mother_Age = 20-24: Partial (1.0)

| | | Mother_Age = 35-39: Partial (0.0)

| | | Mother_Age = 45-49: Partial (0.0)

| | Region = Benishangul_Gumuz: Full (3.0/1.0)

| | Region = Tigray: Partial (4.0/2.0)

| | Region = Affar: Unimmunised (2.0/1.0)

| | Region = Gambela

| | | Father_Education = No_Education: Partial (3.0/1.0)

| | | Father_Education = Higher: Partial (1.0)

| | | Father_Education = Primary: Partial (3.0/1.0)

| | | Father_Education = Dont_ know: Unimmunised (0.0)

| | | Father_Education = Secondary: Unimmunised (4.0)

| | Region = Addis_Ababa: Full (61.0/10.0)

| | Region = Somali

| | | Distance_to_health_facility = Big_Problem: Full (3.0/1.0)

| | | Distance_to_health_facility = Not_A_Big_Problem: Unimmunised (2.0/1.0)

| | Region = Dire_Dawa

| | | Mother_Education_level = No_Education: Full (4.0/1.0)

| | | Mother_Education_level = Primary: Partial (10.0/2.0)

| | | Mother_Education_level = Secondary: Full (7.0/2.0)

| | | Mother_Education_level = Higher: Full (3.0/1.0)

Number of Leaves : 529

Size of the tree : 690

Time taken to build model: 0.4 seconds

=== Stratified cross-validation ===

=== Summary ===

Correctly Classified Instances 5131 62.497 %

Incorrectly Classified Instances 3079 37.503 %

Kappa statistic 0.3068

Mean absolute error 0.3271

Root mean squared error 0.4202

Relative absolute error 83.3303 %

Root relative squared error 94.8534 %

Total Number of Instances 8210

=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
	0.365	0.056	0.567	0.365	0.444	0.78	Unimmunised
	0.806	0.576	0.633	0.806	0.709	0.613	Partial
	0.424	0.097	0.629	0.424	0.506	0.738	Full
Weighted Avg.	0.625	0.355	0.621	0.625	0.608	0.676	

=== Confusion Matrix ===

a	b	c	<-- classified as
505	842	36	a = Unimmunised
339	3653	539	b = Partial
46	1277	973	c = Full

Appendix 4: Visual Basic code

```
private void predictBtn_Click(object sender, EventArgs e)
        {


            if (PodCmb.SelectedIndex == 0 && RegCmb.SelectedIndex == 2)
                if (FlrCmb.SelectedIndex == 0)
                {
                    Imstatus.Text = "PARTIAL";

                }

            if (PodCmb.SelectedIndex == 0 && RegCmb.SelectedIndex == 2)
                if (FlrCmb.SelectedIndex == 1)
                  Imstatus.Text = "PARTIAL";

           if(PodCmb.SelectedIndex==0 && RegCmb.SelectedIndex==10)
               if(ageCmb.SelectedIndex==3 && FlrCmb.SelectedIndex==0)
                   if(MesCmb.SelectedIndex==0 && PCmb.SelectedIndex==2)
                       if(MarsCmb.SelectedIndex==1 && MesCmb.SelectedIndex==0)
                       Imstatus.Text = "PARTIAL";

           if (PodCmb.SelectedIndex == 0 && RegCmb.SelectedIndex == 1)
               if (MesCmb.SelectedIndex == 0 && WsCmb.SelectedIndex == 0)
                   if (FlrCmb.SelectedIndex == 0 && ageCmb.SelectedIndex == 2)
                       if(PCmb.SelectedIndex==3)
                       Imstatus.Text = "UNIMMUNISED";
           if (PodCmb.SelectedIndex == 0 && RegCmb.SelectedIndex == 1)
               if (MesCmb.SelectedIndex == 0 && WsCmb.SelectedIndex == 0)
                   if (FlrCmb.SelectedIndex == 0 && ageCmb.SelectedIndex == 2)
                       if(PCmb.SelectedIndex==2 && DhfCmb.SelectedIndex==1)
                           Imstatus.Text = "UNIMMUNISED";

             if(PodCmb.SelectedIndex==0 && RegCmb.SelectedIndex==1)
                 if(MesCmb.SelectedIndex==0 && ageCmb.SelectedIndex==2)
                     if(WsCmb.SelectedIndex==2 && HesCmb.SelectedIndex==0)
                         if(FlrCmb.SelectedIndex==0)
                     Imstatus.Text = "UNIMMUNIZED";

            if(PodCmb.SelectedIndex==0 && RegCmb.SelectedIndex==8)
                  if(WsCmb.SelectedIndex==0 && MesCmb.SelectedIndex==0)
                      if (FlrCmb.SelectedIndex == 0 && HesCmb.SelectedIndex == 0)
                          Imstatus.Text = "UNIMMUNIZED";
              if (PodCmb.SelectedIndex == 1 && ResCmb.SelectedIndex == 1)
                  if (RegCmb.SelectedIndex == 2 && PCmb.SelectedIndex == 1)
                      if (DhfCmb.SelectedIndex == 1 )
                          Imstatus.Text = "PARTIAL";
              )

               if(PodCmb.SelectedIndex==1 && ResCmb.SelectedIndex==1)
                   if(RegCmb.SelectedIndex==8 && WsCmb.SelectedIndex==0)
                       if(MarsCmb.SelectedIndex==3)
                           Imstatus.Text = "PARTIAL";


            if(PodCmb.SelectedIndex==1 && ResCmb.SelectedIndex==1)
                if(RegCmb.SelectedIndex==9 && HesCmb.SelectedIndex==1)
```

```
            if(FlrCmb.SelectedIndex==0 && MesCmb.SelectedIndex==0)
                Imstatus.Text = "PARTIAL";

    if(PodCmb.SelectedIndex==2 && ResCmb.SelectedIndex==0)
        if(RegCmb.SelectedIndex==9)
            Imstatus.Text = "PARTIAL";

    if (Imstatus.Text == "")
        MessageBox.Show("There is NO Rule as you Enter");

}

private void Form1_Load(object sender, EventArgs e)
{

}

private void RestBtn_Click(object sender, EventArgs e)
{
    ageCmb.SelectedIndex = -1;
    MarsCmb.SelectedIndex = -1;
    PodCmb.SelectedIndex = -1;
    DhfCmb.SelectedIndex = -1;
    Imstatus.Text = "";  // FOR IMMUNIZATION STATUS!
    MesCmb.SelectedIndex = -1;
    HesCmb.SelectedIndex = -1;
    SexCmb.SelectedIndex = -1;
    PCmb.SelectedIndex = -1;
    RegCmb.SelectedIndex = -1;
    ResCmb.SelectedIndex = -1;
    WsCmb.SelectedIndex = -1;
    FlrCmb.SelectedIndex = -1;

}
```

Appendix 5

Addis Ababa University

School of Graduate Studies

School of Information Science and School of Public Health

The Prototype Validity Questionnaire

Direction:

Dear respondents,

This questionnaire is designed to check the validity of the prototype which I designed as part of my MA study. The truthfulness of your responses will contribute much to the validity of the prototype. You are, therefore, cordially requested to be honest to provide accurate information. I would like to let you know that any information you provide in this questionnaire will be kept strictly confidential and will only be used for this study. There is no need to write your name on any part of this questionnaire.

Thank you in advance

Gender: 1. Male 2. Female

	Poor	Fair	Good	Very Good	Excellent
The features of interface in terms of ***easy to use***					
The features of interface in terms of ***easy to understand***					
The features of interface in ***facility decision making***					
The features of interface in terms of ***time taking***					
The features of interface in ***formulating Prediction***					

DECLARATION

I, the undersigned, declare that this thesis is my original work, has not been presented for a degree in this or any other university, and that all the material sources used in this thesis have been fully acknowledged.

_____

Hiwot Abebe

June, 2014

This thesis has been submitted for examination with our approval as university advisors.

_____

Million Meshesha (PhD)

June, 2014

_____

Wubegzier Mekonnen (PhD)

June, 2014