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

PREDICTING THE OCCURRENCE OF MEASLES OUTBREAK IN ETHIOPIA USING DATA MINING TECHNOLOGY

By

SELAM ASSAMNEW

July, 2011

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

PREDICTING THE OCCURRENCE OF MEASLES OUTBREAK IN ETHIOPIA USING DATA MINING TECHNOLOGY

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

By

SELAM ASSAMNEW

July, 2011

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

PREDICTING THE OCCURRENCE OF MEASLES OUTBREAK IN ETHIOPIA USING DATA MINING TECHNOLOGY

By

SELAM ASSAMNEW

Name and signature of Members of the Examining Board

Name	**Title**	**Signature**	**Date**
_____	Chairperson	_____	_____
_____	Advisor(s),	_____	_____
_____	Examiner,	_____	_____

Declaration

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

_____

Date

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

_____

Advisor

ACKNOWLEDGEMENT

First and for most I'm gratified to the almighty GOD for giving me so much.

I am heartily thankful to my adviser, Dr Million Meshesha, for all I have learned from him and for his continuous help and support in all stages of this thesis, whose encouragement, guidance and support from the initial to the final level enabled me to develop an understanding of the subject.

I would like to express my deep gratitude and respect to Dr Amare Mersha who has inspired me to work on this thesis and his continuous advice and support in arranging opportunities to collect measles surveillance data and to meet with domain experts from WHO for discussion.

My special thanks goes to my brother Fitsum Assamnew whom assisted me morally and financially whenever I needed one qnd the rest of my family who has been there with me all the way.

I am grateful to my sweet husband Muluken Tsehayne (Deliye) whose care, patience, and attention made the whole process of this research work and my pregnancy less difficult than it would be without him.

Lastly, I offer my regards and blessings to all of those who supported me in any respect during the completion of the research.

TABLE OF CONTENT

Contents

LIST OF FIGURES

LIST OF TABLES

ABSTRACT

Measles is a contagious disease caused by measles virus. Measles virus is paramyxovirus of a single serological type. WHO (World Health Organization) expanded program on immunization in 1989 estimates that 1.6 million people die from measles each year in developing countries making it the biggest killer among the six EPI (Expanded Program for Immunizations) target disease. . Furthermore WHO estimates that during 2000–2007, measles deaths declined by 89% in WHO African regions, from approximately 395 000 in 2000 to 45000 in 2007. Although global deaths from measles have decreased markedly in past decades, largely as a result of intensive vaccination efforts, still measles outbreaks continue to occur throughout the regions.

The main objective of this study is to design a predictive model using data mining technology that can help predict the occurrence of measles outbreaks in Ethiopia. This can greatly support the effort to control the outbreak of measles, help efficient use of data and also effective utilization of the already scarce resource of Ethiopia.

Data mining provides automated pattern recognition and attempts to uncover patterns in data that are difficult to detect with traditional statistical methods. The application of data mining in the health care industry has a long and successful history. Data mining has a greater advantage to raise the quality and efficiency of health-related products and services.

The methodology used to achieve the goal of building predictive model using data mining technique for this research was a hybrid six-step Cios KDP. It had six basic steps. These were: problem domain understanding, data understanding, data preparation, data minng, evaluation of the discovered knowledge and use of the discovered knowledge. The required data was collected from WHO measles surveillance database covering the period 2006-2011. Then, data preparation tasks (such as data transformation, deriving of new fields, and handling of missing variables) were undertaken. Naïve bayes and decision tree data mining techniques were employed to build and test the models. Models were built and tested by using a dataset of 15631 records.

The researcher used Naïve bayes and decision tree data mining techniques to build the models. 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 inputting 70% of the records to train a model and then supply the unseen 30% of the record for testing the performance of the model. The next option used by the researcher to improve the performance of the models were to test if a better model could be obtained by excluding one or more of the input variables and training different models. The J48 algorithm has shown better prediction accuracy.

The results from this study were very promising. It proved that applying data mining techniques on measles surveillance data to build a model that predicts the occurrence of measles outbreak in different Ethiopian Regions is possible.

ACRONYMS

CDC	Centers for Disease Control and Prevention
CHIP	Health Improvement Program
CPRS	clinical patient record system
CRISP-DM	Cross-Industry Standard Process for Data Mining
DM	Data mining
EPI	Expanded Programme for Immunizations
HFA	Health for All
IDSR	integrated disease surveillance and response
IRB	Institutional Review Board
IVD	Immunization and Vaccine Development
KDD	Knowledge Discovery from Data
KDP	knowledge discovery process
MoH	Ministry of Health
SDR	standardized death rate
SIA	supplemental immunization activities
SMOTE	Synthetic Minority Over-sampling Technique
TMR	the Medical Record
UNF	United Nations Foundation
UNICEF	United Nations International Children's Emergency Fund
VPDs	vaccine preventable diseases
WHO- AFRO	World Health Organization African Region
WHO	World health organization

CHAPTER ONE

INTRODUCTION

1.1. Background

Measles is a contagious disease caused by measles virus. Measles virus is paramyxovirus of a single serological type. The disease is highly communicable with an incubation period of about 10 days (with a range of 7 to 18 days). The disease is characterized by prodromal fever, conjunctivitis, coryza, cough and presence of koplik spots. A characteristic maculopapular rash appears on the third to seventh day begging on the face and become more general. Man is the only source of the measles virus [1].

WHO (World Health Organization) expanded program on immunization in 1989 estimates that 1.6 million people die from measles each year in developing countries making it the biggest killer among the six EPI (Expanded Programme for Immunizations) target disease. Furthermore WHO reported that during 2000–2007, measles deaths declined by 89% in WHO African regions, from approximately 395 000 in 2000 to 45000 in 2007. Although global deaths from measles have decreased markedly in past decades, largely as a result of intensive vaccination efforts [2], still measles outbreaks continue to occur throughout the regions, highlighting the need to ensure that the regional strategy is fully implemented.

An outbreak consists of an increase in the number of measles cases reported compared with cases reported previously in the same areas during similar time intervals in non outbreak years. Outbreaks occur when the accumulated number of susceptible individuals is greater than the critical number of susceptible individuals, or epidemic threshold, for a given population to sustain transmission. WHO-AFRO defines a suspected measles outbreak as the occurrence of five or more reported suspected measles cases in a health facility or district in one month, with plausible means of transmission An outbreak of

measles is said to have been confirmed when there are 3 or more IgM positive measles cases in a health facility or district in one month [3].

In discussions centered on expanded measles control, elimination, and possible eradication, better prediction of measles outbreak is very important to focus and target vaccination control measures [4]. Measles infection is still prevalent in many developing countries especially in parts of Africa and Asia where more than 20 million measles cases are reported annually [5].

Ethiopia's health status is poor relative to other low-income countries, including those in Sub-Saharan Africa. While the under-five mortality rate is consistently declining, it remains high, with most recent survey estimates placing it at 123 deaths per 1,000 live births [6]. Measles remains a problem in Ethiopia, due mainly to the low measles immunization rate (estimated coverage of 51% in 2001). A total of 3,797 cases and 58 deaths due to measles were reported in 2002-03, while this figure followed a successful measles vaccination campaign in 2002-03.

With the expanded efforts of measles control to help guide policy planning and vaccine strategies reliable prediction of measles outbreak become increasingly important. These can be achieved by applying data mining techniques on measles surveillance data to build a model that predicts the occurrence of measles outbreak in different Ethiopian Regions.

Data mining is the analysis of (often large) observational data sets to find unsuspected relationships and to summarize the data in novel ways that are both understandable and useful to the data owner. The relationships and summaries derived through a data mining exercise are often referred to as models or patterns [7].

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

1.2. Statement of the Problem

Measles is one of the most contagious airborne viral diseases known; it has a high transmission rate, associated with high fever, rashes and vomiting, The virus resides within the nose and throat of infected persons and is easily spread by coughing, sneezing or by close contact. The contagion index reaches nearly 100%, meaning that nearly every unimmunized person coming in close contact with a patient will also become infected with measles [8].

Measles is a leading cause of vaccine-preventable deaths among children. In 2006, an estimated 242,000 people died of measles; about 217,000 of these deaths were among children under 5 years of age. About 1-5% of children with measles die from complications of the disease. In refugee settings, the death rate from measles may be as high as 30% [9]. Global deaths due to measles fell by 48% in 2004 from 871,000 in 1999. The largest reduction occurred in sub Saharan Africa where estimated measles cases and deaths dropped by 60%. These statistics make measles one of the single leading causes of death among children in most developing countries despite the availability of a safe and effective vaccine for more than 40 years [3].

Although national immunization program prevent over 80 million cases of measles and 4.5 million deaths annually, it is estimated that over 30 million cases and 875,000 deaths still occur every year [1]. High transmission of measles despite high coverage of 1-dose measles vaccine has been reported in some developing countries. Some of the reason for vaccine failure in developing countries includes; poor seroconversio, questional potency of vaccines due to problem with cold chain that ultimately affect the quality of vaccine and waning immunity. To prevent outbreak of measles in developing countries, there is a need for second dose of measles vaccine to take care of vaccine failures [1].

Epidemiological investigations of recent outbreaks of vaccine preventable diseases have indicated that incomplete immunization was the major reason for the outbreaks [10]. The study tried to identify the predictors of defaulting from completion of child immunization

among children between ages 9-23 months in Wonago district, South Ethiopia. Methods used in the research were unmatched case control study conducted in eight Kebeles (lowest administrative unit) of Wonago district in south Ethiopia. Census was done to identify all cases and controls. Measles defaulter rate was 76.2%. Knowledge of the mothers about child immunization, monthly family income, postponing child immunization and perceived health institution support were the best predictors of defaulting from completion of child immunization.

Measles Vaccination status by age group among confirmed Measles Cases in Ethiopia, 2007/2008 shows that 57% of the total confirmed cases were <5 Yrs age and only 36% of the confirmed cases have had at least 1 measles dose. A total of 74 outbreaks were reported in 2008 and there were 2,959 confirmed outbreak cases in 56 Woredas. Guji Zone had a huge outbreak of 1,135 confirmed cases of measles [11].

In the above researches [10, 11], about measles outbreaks were conducted by using small proportion of the accumulated data from suspicious site. Besides, in these researches data analysis was conducted by using simple statistical methods. The absence of significant attempt that has been made so far to carry out research in this area using data mining technique rationalizes the importance of this research. Data mining provides automated pattern recognition and attempts to uncover patterns in data that are difficult to detect with traditional statistical methods.

Thus, to alleviate the current problem, this study tries to build a model that predicts the occurrence of measles outbreak in Ethiopia using data mining technique and measles case based surveillance data. This will help to protect the public health from morbidity, mortality and reduce the economic burden from the already scarce resource of the healthcare system in Ethiopia.

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

- Is there any pattern that can be extracted from surveillance data for measles outbreak prediction?

- Which data mining technique is appropriate to predict the occurrence of measles outbreak in Ethiopia?
- What are the factors that contribute to the occurrence of measles outbreak in each region?

1.3. Objective of the Research

1.3.1. General Objective

The general objective of this study is to design a predictive model using data mining technology that can help predict the occurrence of measles outbreaks in Ethiopia. This can greatly support the effort to control the outbreak of measles in Ethiopia.

1.3.2. Specific Objectives

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

- To understand the problem domain by reviewing literatures and documents.
- To clean measles surveillance data by applying preprocessing task like cleaning, transformation and attribute selection
- To build a model using data mining tool on cleaned measles surveillance data, which helps to apply data mining techniques in identifying measles outbreak patterns
- To evaluate the performance of the model with domain experts from WHO and using test datasets
- To report the result and forward recommendations for further studies.

1.4. Research Methodology

1.4.1. Research Design

In this study, a hybrid six-step Cios KDP [12] model was used to achieve the goal of building predictive model using data mining technique. This model was chosen since it exhibits all the advantages of well known and used methodology called CRISP-DM and provides a more general, research-oriented description. It also has more detailed feedback mechanisms that is helpful for achieving the research objective. This methodology is tools independent and combines both aspect of the academic and industrial model.

Based on the hybrid model of the Cios six-step methodology, the required tasks and methods are identified in order to predict the occurrence of measles outbreak in Ethiopia.

1.4.2. Problem Domain Understanding

A model was needed to predict which region and age group would most likely have measles outbreak to take preventive measures before it causes further harm to the society. For understanding the problem domain of measles outbreak in Ethiopia, the researcher used secondary data reference and consultation with domain experts.

To achieve the goal of this research WHO measles surveillance data was used. The Surveillance systems rely on the identification of persons with the clinically recognizable symptoms of measles for detecting and responding to outbreaks, vaccinating susceptible contacts, and assessing the efficacy of vaccines and the impact of vaccination programs. Information about the suspected measles cases were collected through the surveillance using identification of the patient, clinical data, classification and possible source of infection. The research tries to build a model using the collected measles surveillance data to identify which factors have a higher probability of classifying the occurrence of an outbreak.

1.4.3. Data Understanding

The primary source of data for this research is WHO measles surveillance. The database contains measles data collected through surveillance and for this research the surveillance data collected from year 2006 to 2011 is used for building the model. The collected data was in Microsoft Access, it contains a total of 26103 records about patients from all regions of Ethiopia reported from their local health facility. The dataset contains both numeric and nominal values. To get a clear visualization of the data it was later converted to Microsoft Excel. Selection of attribute is made using subjective judgment of the researcher and discussion with domain experts from WHO.

The data contains information about the location, vaccine status, result for measles and Rubella test, age, sex, outcome and date patient were seen at the health facility. And there is more information about the specimen condition, classification and records status of the surveillance data.

1.4.4. Data Preparation

This is one of the crucial steps to produce dataset used for modeling by Weka software. The following steps were undertaken in the preprocessing stage; data cleaning, attribute selection, and data transformation. In order to correct the errors identified trough observation from the preprocessing stage, measures like filling in missing values based on the idea of observing neighboring records, correcting inconsistencies like spelling errors and removing instances with missing values were undertaken. SMOTE was applied to overcome the problem of high imbalance in the class value. Finally, the dataset ready for the data mining process contains only 13 attributes and 15631 records.

1.4.5. Data Mining

To build a predictive model from the cleaned data, Weka data mining software was used. Weka is a tool containing numerous machine learning algorithms that can be applied to to achieve the objective of this research. It supports several standard data mining tasks,

more specifically, data preprocessing, clustering, classification, regression, visualization, and feature selection and also this software is platform independent. The researchers choose Weka software for the data mining purpose, Particularly Decision tree and Naïve Bayes algorithm. These algorithms were proved to be important when applying them in healthcare data from different literatures therefore J48 algorithm and naïve bayes algorithms were applied on the measles surveillance data to come up with the predictive model for predicting the occurrence of measles outbreak.

1.4.6. Evaluation of the Discovered Knowledge

In order to properly interpret knowledge patterns, Performance evaluation methods were used to measure the accuracy of the model created by the Decision tree and Naïve bayes. The methods are true positive rate, false positive rate, accuracy, precision and ROC curve. 10 folds cross validation and 70% split for training and 30% for testing, test option were also used to measure the performance of the model.

1.5. Scope and Limitation of the Research

The scope of this research is to evaluate the potential of data mining technique to predict the occurrence of measles outbreak in Ethiopia. Regional level measles surveillance data for this research was collected from WHO. The data covers the period from 2006-2011, which amounts to 26103 records. Thus, finding of this research can be considered as pertinent to the WHO and FMoH to make precise decision regarding measles outbreak and vaccine strategies.

The following are some of the limitations of this research because of the limited available research time and resource. The research was originally planned to conduct measles outbreak prediction at district level but due to the above reasons, this research was conducted to identify factors that contribute to the outbreak of measles at regional level. Data source of this study was WHO measles surveillance database. Databases available at other related organizations like Federal Ministry of Health, Ethiopian Health and Nutrition Research Institute, UNICEF, National meteorological Agency were not used.

1.6. Significance of the Research

After the completion of this research, it will provide a valuable model to predict the occurrence of measles outbreak in Ethiopia. The finding of this research

- Supports primary health care providers, policy makers, planners to make a better decision.
- Will support the effort of eradicating measles in particular and improving the quality of life in general.
- hepls FMoH and WHO to plan ahead for immunizing the society.
- will be used as a basis to enlighten planning of health programs to channel advocate efforts and researches at the national level,
- Will help health planners to understand the nature and pattern of the occurrence of measles outbreaks in different regions of Ethiopia.
- Can also serve as a baseline data reference and initiative to conduct further studies in the future.

1.7. Dissemination of the result

The result of the research will hopefully be presented at annual students and staff research conference in Addis Ababa University, national conference of Ethiopian public health association, UNICEF and World health organization (WHO).

1.8. Organization of the Paper

This research report is organized into six chapters. The first Chapter briefly discusses background to the problem area, and states the problem, the general and specific objectives of the study, the research methodology, the scope and limitation, significance and dissemination of the research. Chapter two reviews literature on measles and the need for immunization, data mining, data mining technique and methodology respectively. The application of data mining in health care is also briefly discussed in this chapter. Chapter three explains the Data mining techniques applied on measles surveillance data, the

Naïve Bayes and decision trees, used in this research. Moreover Performance Evaluation method used for Predictive Modeling is conversed this chapter. Chapter four is dedicated for the discussion of two basic issues, business understanding and preprocessing. Chapter five presents the experimentation phase of the study at hand. Results of the decision tree and Naïve Bayes experiments were also briefly discussed here. Finally, chapter six provides conclusion of the research, and also presents recommendation for future work.

CHAPTER TWO

LITERATURE REVIEW

Measles is a highly infectious viral disease caused by a Morbillivirus. It only affects humans and rapidly spreads among individuals who have not been vaccinated. Measles vaccination rates are sensitive indicators of functional public health systems. One to two doses of a US$0.14 vaccine could prevent a disease that practically affects 100 percent of the population with an approximate case-fatality rate of 3 to 6 percent in Sub-Saharan Africa [13]. Given that measles vaccine is one of the most cost-effective and low-cost health interventions, low coverage rates of the vaccine in the 1990s is indicative of the poor state of public health infrastructures in various Sub-Saharan Africa populations.

2.1. The Need for Immunization

According to Feachem and Jamison [13] the eradication of smallpox was an outstanding display of concerted global action in a war against microbial invaders. The progress in expanding poliomyelitis and measles vaccination efforts and their elimination from many regions further demonstrates that vaccines are among the most powerful public health tools.

Spotting efficiently and effectively the most likely places or regions to have measles outbreaks at the precise point in time helps the society to be immunized ahead from the occurrence of catastrophic measles morbidity and mortality and a successful eradication of measles from Ethiopia.

It is estimated that immunization saves two million lives per year. Immunization rates for the six major vaccine-preventable diseases—pertussis (whooping cough), tuberculosis, tetanus, polio, measles, and diphtheria—have risen from less than 10 percent in the 1970s to nearly 80 percent today. However, coverage has leveled off more recently. Worldwide,

nearly 30 million children are still not reached each year with routine immunization. Rates in some African countries have dropped to below 50 percent [9].

Immunization from measles is effective, and has resulted in significant reductions in case burden in many parts of the world. Unfortunately, a large percentage of children in the African region never receive their first measles vaccine dose in time for immunity to take hold. The cost of protecting a child against measles is less than USD 1.00, and when correctly administered at 9 months of age, the measles vaccine offers life-long protection to approximately 85% of those vaccinated.

The Role of the Immunization and Vaccine Development (IVD) Programme [9]

i. Strengthen routine immunization against measles.
ii. Establish disease surveillance, laboratory confirmation and data management systems.
iii. Plan and implement supplemental immunization activities (SIAs).
iv. Monitor and evaluate the program implementation.
v. Mobilize resources.
vi. Advocate with local and international partners.

2.2. Data Mining

It is estimated that the amount of information in the world doubles every 20 months [14]; Data mining is becoming more widespread every day, because it empowers companies to uncover profitable patterns and trends from their existing databases. Companies and institutions have spent millions of dollars to collect megabytes and terabytes of data but are not taking advantage of the valuable and actionable information hidden deep within their data repositories. Data mining is becoming more widespread every day, because it empowers companies to uncover profitable patterns and trends from their existing databases. However, as the practice of data mining becomes more widespread, companies that do not apply these techniques are in danger of falling behind and losing market share, because their competitors are using data mining and are thereby gaining the

competitive edge in Discovering Knowledge in Data, the step-by-step hands-on solutions of real-world business problems [15].

Data mining is about solving problems by analyzing data already present in databases. There is a wide availability of huge amounts of data and the imminent need for turning such data into useful information and knowledge. Data mining tools perform data analysis and may uncover important data patterns, contributing greatly to business strategies, knowledge bases, and scientific and medical research [16].

Data mining is the extraction of hidden predictive information from large databases; it is a powerful technology with great potential to help organizations focus on the most important information in their data warehouses [16]. Data mining tools predict future trends and behaviors, helps organizations to make proactive knowledge-driven decisions.

Susan p. [17] defines KDD process as a process that employs methods from various fields such as machine learning, artificial intelligence, pattern recognition, database management and design, statistics, expert systems, and data visualization. It is said to employ a broader model view than statistics and strives to automate the process of data analysis, including the art of hypothesis generation.

Rea [18] further wrote, Data mining is the search for relationships and global patterns that exist in large databases but are hidden within the vast amount of data, such as a relationship between patient data and their medical diagnosis. These relationships represent valuable knowledge about the objects in the database and if the database is a faithful mirror of the real world registered by the database. Indeed, data mining technology has become a new paradigm for decision making and it has attracted a great deal of attention in the information industry and in society as a whole in recent years, due to the wide availability of huge amounts of data and the imminent need for turning such data into useful information and knowledge. The information and knowledge gained can be used for applications ranging from market analysis, fraud detection, credit scoring,

warranty management, and customer retention, to production control and science exploration [18, 16].

2.3. Data Mining Techniques

Data mining techniques provide people with new power to research and to manipulate the existing large volume of data. There are many achievements of application from data mining techniques to various areas such as engineering, marketing, financial, medical and so on. Data mining activities can be categorized in to two, namely prediction and description modeling.

2.3.1. Prediction Modeling

Prediction modeling is a technique that involves using some variables or fields in the data set to predict unknown or future values of other variables of interest. It is used to develop a model to relate a dependent variable with a set of independent variables. There are two types of prediction modeling, namely classification, for categorical dependent variables, and value prediction, for continuous dependent variables.

Classification is appropriate if the goal is to predict group membership of new records based on their characteristics (independent variables). Using classification, the most influential variable is identified and used to split the data into groups. This is then repeated with the next most influential variable until the data are fully characterized. Some examples of classification technique includes decision tree classifier, rule based classifier, neural networks, support vector machines and Naïve Bayes Classifiers. Each technique employs a learning algorithm to identify a model that best fits the relationship between the attribute set and the class level of the input data. The model generated by the learning algorithm should both fit the input data and the class and correctly predict the class label of record it has never seen before. For example, it may be possible to determine a classification criterion or rule that discriminates between different groups of patients with and without side-effects based on age, sex or socio-economic class.

Value prediction uses classification and regression to predict the future outcome of a patient based on, for example, their demographic or socio-economic characteristics. However, we need to use caution as, in any data analysis of continuous outcomes, the results of value prediction can be influenced by the presence of outliers in the data.

2.3.2. Description Modeling

The descriptive model identifies the patterns or relationships in data and explores the properties of the data examined. It focuses on finding patterns describing the data that can be interpreted by humans E.g. Clustering, Summarization, Association rule, Sequence discovery etc.

Clustering uses an algorithm that segregates a database by evaluating the dissimilarity between records. Pairs of records are compared by the values of the individual fields within them, and clustering into groups provides fast and effective ordering in large datasets. Segmentation could be used to group patients with similar symptoms or diagnoses to determine whether there is a drug association. Thus, clustering is a technique of choice if the goal is to reduce a large sample of records to a smaller set of specific homogeneous subgroups (clusters) without losing much information about the whole sample. Because of the heterogeneity between clusters, this analysis can also be helpful in hypothesis development about the nature of the variation between subgroups [19]. For example, if a database contained details of different cardiac pathologies (e.g. valvular heart disease) and medication (e.g. fenfluramine-phentermine), clustering analysis may have segregated patients according to heart disease and identified fenfluramine-phentermine as one of the main factors in this group. Clustering is concerned with grouping together objects that are similar to each other and dissimilar to the objects belonging to other clusters [19].

There are many algorithms for clustering but the most widely used methods are k-means clustering and Agglomerative Hierarchical Clustering in which clusters are formed by

measuring distance between objects and objects with similar distance will be grouped in to one cluster.

Association rules are widely used in data mining to find patterns in data. The patterns reveal combinations of events that occur at the same time. Once identified, these combinations, which are also known as "group associations" can be used to improve decision-making in a wide variety of applications. Example of most widely used algorithm for association rule includes Apriori algorithm which is also known as Market Basket Analysis.

2.4. Data Mining Methodology

The Data mining process model helps organizations to better understand the KDP and provides a roadmap to follow while planning and executing the data mining project. This in turn results in cost and time savings, better understanding, and acceptance of the project results. Such processes are nontrivial and involve multiple steps, reviews of partial results, possibly several iterations, and interactions with the data owners [12]. The knowledge discovery process models usually emphasize independence from specific applications and tools; they can be broadly divided into two.

- Industrial model: those that take into account industrial issues like six-step CRISP model and the five-step model by Cabena et al
- Academic model: those that do not take into account industrial issues. like KDD model, the nine-step model by Fayyad et al. and the eight-step model by Anand and Buchner [12].

2.4.1. KDD Process

Knowledge discovery as a process that involves a series of steps to preprocess the data prior to mining and post processing steps to evaluate and interpret the modeling results [16], as shown in figure 2.1. The first step is Data Selection where data relevant to the analysis tasks are retrieved from the database. This step is followed by Data Cleaning in which noise and inconsistent data are removed from the dataset during this step there may

be a need for Data Integration where multiple data sources may be combined. The third step is Data Transformation where data are transformed or consolidated into forms appropriate for mining by performing summary or aggregation operations once the data is ready for mining, Data mining techniques are applied an essential process where intelligent methods are applied in order to extract data patterns. Then Pattern evaluation step identifies truly interesting patterns representing knowledge using various performance measures. Finally, there is a need of Knowledge presentation for visualization and knowledge representation techniques to present the mined knowledge to the user

Data selection, cleaning and transformation are different forms of data preprocessing, where the data are prepared for mining. The data mining step may interact with the user or a knowledge base. The interesting patterns are presented to the user and may be stored as new knowledge in the knowledge base. The KDD Process is a highly iterative, user involved, multistep process, as can be seen in figure 2.1.

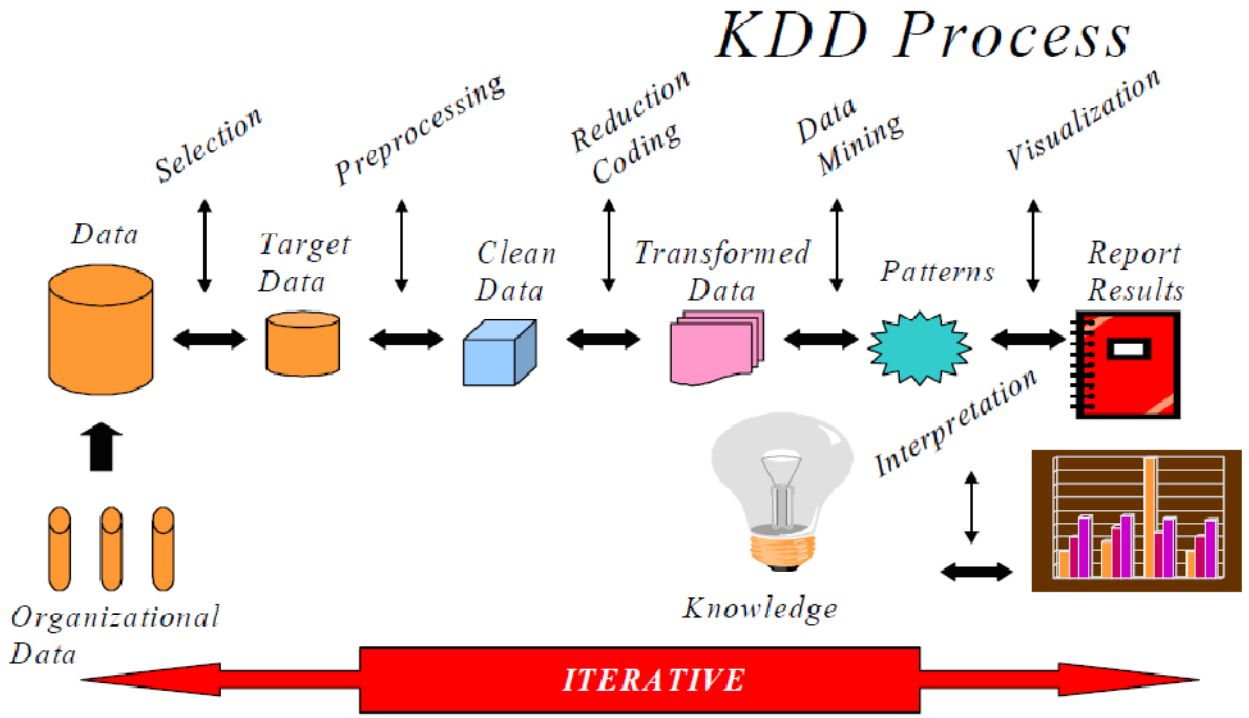


Figure 2.1: The KDD process

Brachman and Anand [20] argued that the KDD process is an interactive and iterative, involving numerous steps with many decisions being made by the user. Fayyad also defined the KDD process as it is preceded by the development of an understanding of the application domain, the relevant prior knowledge and the goals of the end-user. It must be continued by the knowledge consolidation, incorporating this knowledge into the system. KDD has been more formally defined as "the non-trivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data"[13].

2.4.2. CRISP-DM Model

The CRISP-DM KDP model is one of the most widely used data mining methodology for knowledge discovery that consists of six steps as shown in figure 2.2 which are summarized below:

Step 1: **Business understanding.** This step focuses on the understanding of objectives and requirements from a business perspective. It also converts these into a DM problem definition, and designs a preliminary project plan to achieve the objectives. It is further broken into several sub steps, namely, determination of business objectives, assessment of the situation, determination of DM goals, and Generation of a project plan.

Step 2: **Data understanding.** This step starts with initial data collection and familiarization with the data. Specific aims include identification of data quality problems, initial insights into the data, and detection of interesting data subsets. Data understanding is further broken down into collection of initial data, description of data, exploration of data, and Verification of data quality.

Step 3: **Data preparation.** This step covers all activities needed to construct the final dataset, which constitutes the data that will be fed into DM tool(s) in the next step. It includes Table, record, and attribute selection; data cleaning; construction of new attributes; and transformation of data. It is divided into selection of data, cleansing of data, construction of data, integration of data, and Formatting of data sub steps.

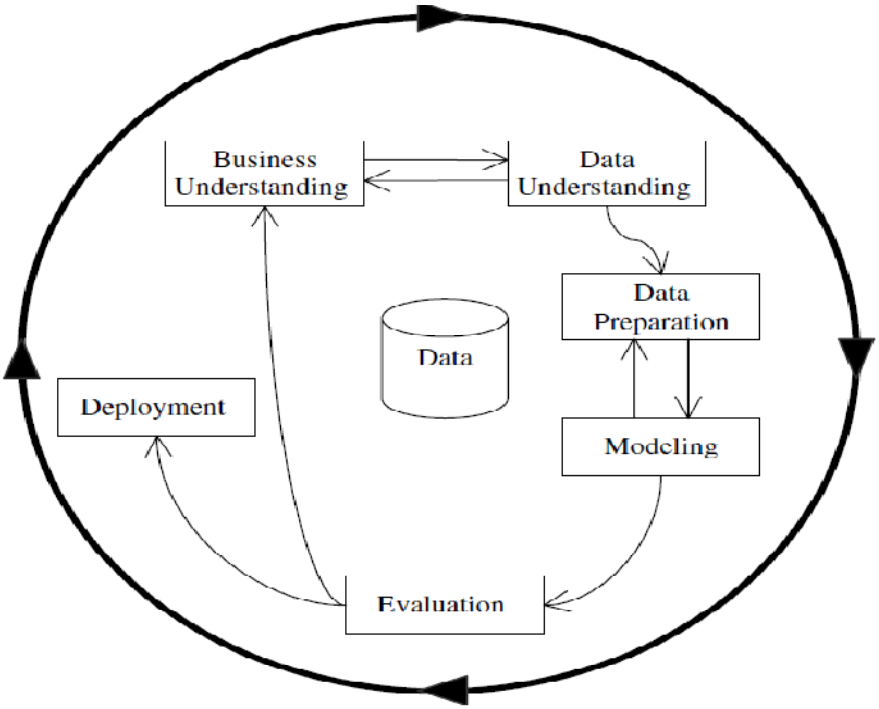


Figure 2.2. The CRISP-DM KD process model

Step 4: **Modeling**. At this point, various modeling techniques are selected and applied. Modeling usually involves the use of several methods for the same DM problem type and the calibration of their parameters to optimal values. Since some methods may require a specific format for input data, often reiteration into the previous step is necessary. This step is subdivided into selection of modeling technique(s), generation of test design, creation of models, and Assessment of generated models.

Step 5: ***Evaluation*.** After one or more models have been built that have high quality from a data analysis perspective, the model is evaluated from a business objective perspective. A review of the steps executed to construct the model is also performed. A key objective is to determine whether any important business issues have not been sufficiently considered. At the end of this phase, a decision about the use of the DM results should be reached. The key sub steps in this step include evaluation of the results, process review, and Determination of the next step.

Step 6: ***Deployment*.** Now the discovered knowledge must be organized and presented in a way that the customer can use. Depending on the requirements, this step can be as simple as generating a report or as complex as implementing a repeatable KDP. This step

is further divided into plan deployment, plan monitoring and maintenance, generation of final report, and Review of the process sub steps.

2.4.3. Hybrid Models

The development of academic and industrial models has led to the development of hybrid models that combine aspects of both. One such model is a six-step KDP model developed by Cios et al [12]. It was developed based on the CRISP-DM model by adopting it to academic research. The main differences and extensions include providing more general, research-oriented description of the steps and introducing a data mining step instead of the modeling step [12],

Further description of the six steps of the model follows,

1. **Problem domain understanding.** This initial step involves working closely with domain experts to define the problem and determine the project goals, identifying key people, and learning about current solutions to the problem. It also involves learning domain-specific terminology. A description of the problem, including its restrictions, is prepared. Finally, project goals are translated into DM goals, and the initial selection of DM tools to be used later in the process is performed.

2. **Data understanding.** 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. **Data preparation.** This step concerns deciding which data will be used as input for DM methods in the subsequent step. It involves sampling, running correlation and significance tests, and data cleaning, which includes checking the completeness of data records, removing or correcting for noise and missing values, etc. The cleaned data may be further processed by feature selection and extraction algorithms (to reduce dimensionality), by derivation of new attributes (say, by discretization), and by summarization of data (data granularization). The end results are data that meet the specific input requirements for the DM tools selected in Step 1.

4. **Data mining**. Here the data miner uses various DM method such as classification, clustering and association rule discovery to derive knowledge from preprocessed data as per the objective of the study.

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

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. A plan to monitor the implementation of the discovered knowledge is created and the entire project documented. Finally, the discovered knowledge is deployed.

2.5. Related Works

Prather [21] conducted a data mining project at Duck University Medical Center using an extensive clinical database of obstetrical patients to identify factors that contribute to prenatal outcomes. The goal of this knowledge discovery effort was to identify factors that will improve the quality and cost effectiveness of prenatal care. The production system database identified for mining was the computer-based patient record system known as “The Medical Record”, or TMR. TMR is a comprehensive longitudinal clinical patient record system (CPRS) developed at Duke University over the last 25 years [21]. For their work, the specific database selected for the data mining project was the prenatal database used by the Department of Obstetrics and Gynecology at Duke University Medical Center. For the purpose of the initial study, the researchers created a sample two-year dataset (1993-1994) from the data warehouse. The researchers confirmed that a large clinical database could be successfully warehoused and mined to identify clinical factors associated with preterm birth. Finally, the authors concluded that data warehousing and

mining technology are applicable to health care, and that the preliminary mining of a clinical data warehouse has produced promising results.

Larvac [22] combined GIS and data mining using among others, Weka with J48 (free, open source, Java-based data mining tools), to analyze similarities between community health centers in Slovenia. Using data mining, they were able to discover patterns among health centers that led to policy recommendations to their Institute of Public Health. They concluded that "data mining and decision support methods, including novel visualization methods, can lead to better performance in decision-making." Medical informatics may use the technologies developed in the new interdisciplinary field of knowledge discovery in databases (KDD) [17].

Health experts have also begun to look at how to apply data mining for early detection and management of pandemics. Kellogg [23] outlined techniques combining spatial modeling, simulation and spatial data mining to find interesting characteristics of disease outbreak. The analysis that resulted from data mining in the simulated environment could then be used towards more informed policy-making to detect and manage disease outbreaks. Wong [24] introduced WSARE, an algorithm to detect outbreaks in their early stages. WSARE, which is short for "What's Strange About Recent Events" is based on association rules and Bayesian networks. Applying WSARE on simulation models have been claimed to result to relatively accurate predictions of simulated disease outbreaks. Of course, these sorts of claims always come with warnings to take precaution when applying these models in real life.

Lloyed-Williams [25] had also analyzed datasets extracted from the World Health Organization's Health for All (HFA) Database using a data mining approach. During the selection process, mortality data based on the following conditions was extracted by the researchers from the HFA database: life expectancy at birth; probability of dying before five years of age; infant mortality; post-neonatal mortality; standardized death rate (SDR) for circulatory diseases; SDR for malignant neoplasm; SDR for external causes of injury and poisoning; SDR for suicide and self-inflicted injury. Data was extracted for 39

European Countries, and then converted into a format acceptable to the software used for that particular project.

Shegaw [26] applied data mining technique to investigate the potential applicability of data mining technology to predict the risk of child mortality based up on community-based epidemiological datasets gathered by the BRHP epidemiological study. The researcher used neural network and decision tree data mining techniques to build and test the models. Using the neural network approach, the best model was identified for the training made by using the default parameters (i. e. training tolerance of 0.1, learning rate of 1.0, and smoothing factor of 0.9) and the following 9 input variables: "ENVIRN", "AGE", "OUTMIG", "HHRELIG", "HHETHNIC", "HHLITERAC", "HHHEALHTH", "HHWATER", AND "WINDOWS". This model had an accuracy rate of 93% at a testing tolerance of 0.4 and was tested with accuracy of 88 % at testing tolerances of 0.2 and 0.1.the researcher constructed by using several classifiers by using See5 decision tree software. This classifier resulted with an accuracy of 95% on training cases and it achieved 95% accuracy on test cases. the researcher concluded that this research work have proved the potential applicability of data mining technology to predict child mortality patterns based solely on demographic, parental, environmental, and epidemiological factors. The encouraging results obtained from both neural networks and decision trees indicate that data mining is really a technology that should be considered to support child health care prevention and control activities at the district of Butajira.

Helen [27] applied data mining techniques to official data such as the 2001 child labor survey to identifying relationships between attributes within the 2001 child labor survey database that can be used to clearly understand the nature of child labor problem in Ethiopia. The researcher used expectation maximization clustering algorithm implemented in knowledge studio version 3.0. For final data preparation and the selected dataset was categorized into five classes. The apriori algorithm was used to generate association rules from the clustered as well as non-clustered selected dataset. The researcher initially, used all of the records with 86 attributes to apriori, and the first 10 best rules were generated. These rules have a minimum coverage or support of 90% and

minimum accuracy of confidence of 95%. In the second round the number of selected attributes was reduced to 63 to generate the first 10 best rules. At this time the minimum support level was 80% and minimum confidence level was 90%. The researcher applied clustering algorithm on the dataset in order to further refine the result. The clustering algorithm, expectation maximization, was run using the 63 attributes used in experiment 2. A total of four clustering models were built by varying the number of clusters from 2 up to 5.the clustering segmented the records into five clusters. Among these five clusters, the third cluster which contains 42.5% of the selected dataset was chosen and given to the apriori algorithm of Weka. Cluster number 3 was selected. The association rule algorithm, apriori, generated its 10 best association rules with minimum coverage of 95% and minimum accuracy of 90%.

Hemalatha and Megala [28] briefly examine the potential use of classification based data mining techniques such as decision tree, Artificial Neural Network to massive volume of Immunization data. In their study data analysis of children with Immunization details (such as BCG, DPG, Polio and Measles) has been made. 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 data mining algorithms it was possible to verify the improvement of quality.

Up to the knowledge of the researcher, no previous researches have been done to predict the prevalence of immunization, including measles outbreak by applying data mining techniques in Ethiopia. Hence this research has a great contribution to generate patterns that help in planning a better strategy for measles outbreak control using data mining technique.

CHAPTER THREE

METHODS FOR MINING MEASLES SURVEILLANCE DATA

There are different data mining techniques presented with their appropriateness to be applied in different health care areas. The data mining techniques used in this research to predict the occurrence of measles outbreak in Ethiopia on measles case based surveillance data were decision tree and Naïve Bayes (NB).

3.1. Decision Tree

Decision trees are an approach of representing a sequence of rules that lead to a set or value. As a result, they are used for directed data mining, mainly classification. One of the main rewards of decision trees is that the model is quite reasonable since it takes the form of explicit rules.

This allows the evaluation of results and the identification of key attributes in the process [29]. It consisting of nodes and branches organized in the form of a tree such that, every interior non-leaf node is labeled with ideals of the attributes. The branches coming out from an inner node are labeled with ideals of the attributes in that node. Each node is labeled with a rank (a worth of the goal characteristic). Tree based models which include classification and regression trees, are the common implementation of induction modeling [30].

The decision tree algorithm used in this research is J48 algorithm, which is an implementation of the C4.5 decision tree learner. This implementation produces decision tree models. The algorithm uses the greedy technique to induce decision trees for classification. A decision-tree model is built by analyzing training data and the model is used to classify unseen data. J48 generates decision trees, the nodes of which evaluate the existence or significance of individual features, C4.5 algorithm uses the concept of information gain or entropy reduction to select the optimal split. Suppose that we have a

variable X whose k possible values have probabilities p_1, p_2, . . . , p_k . 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

$$H(X) = -\sum_j Pj \log_2(Pj)$$

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(pj)'s, with weights equal to the outcome probabilities, resulting in the formula

$$H(X) = -\sum_j Pj \log_2(Pj)$$

C4.5 uses this concept of entropy as follows. Suppose that we have a candidate split S, which partitions the training data set T into several subsets, T_1, T_2, . . . , T_k . The mean information requirement can then be calculated as the weighted sum of the entropies for the individual subsets, as follows

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

Where Pi represents the proportion of records in subset i. We may then define our information gain to be gain(S) = H(T) − H_S(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.

Decision tree can be built from the very small training set as shown in Table 3.1. In this table each row corresponds to an enduring record. The data set contains three predictor attributes, namely Age, Gender, symptoms and one goal attribute, namely disease whose values to be predicted from symptoms indicates whether the corresponding enduring have a certain disease or not.

Age	Gender	Symptoms	Disease
5	Female	Medium	Yes
3	Male	High	Yes
2	Female	Medium	Yes
4	Female	High	Yes
10	Female	Low	No
9	Male	Low	No
11	Female	Low	No

Table 3.1 Data set used to build decision tree of Figure 3.1

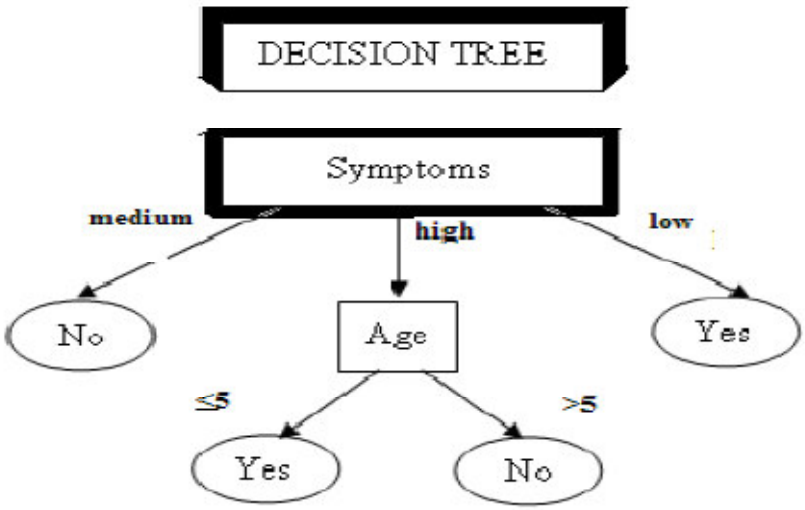


Figure 3.1 A decision tree built from the data in Table 3.1

3.2. Naïve Bayes Classifiers

Naïve Bayes (NB) approach is a very popular classification method that does not use rules, a decision tree or any other explicit representation of the classifier. Rather, it uses the branch of Mathematics known as probability theory to find the most likely of the possible classifications [20]. It relies on all attributes being categorical

The Naïve Bayes Classification algorithm

The Naïve Bayes (NB) algorithm gives us a way of combining the prior probability and conditional probabilities in a single formula, which is used to calculate the probability of each of the possible classifications, then choose the classification with the largest value.

Given a set of k mutually exclusive and exhaustive classifications c_1, c_2, . . . , c_k, which have prior probabilities $P(c_1)$, $P(c_2)$, . . . , $P(c_k)$, respectively, and n attributes a_1, a_2, . . . , a_n which for a given instance have values v_1, v_2, . . . , v_n respectively, the posterior probability of class c_i occurring for the specified instance can be shown to be proportional to

$$P(c_i) \times P(a_1 = v_1 \text{ and } a_2 = v_2 \ldots \text{and } a_n = v_n \mid c_i)$$

Making the assumption that the attributes are independent, the value of this expression can be calculated using the product

$$P(c_i) \times P(a_1 = v_1 \mid c_i) \times P(a_2 = v_2 \mid c_i) \times \ldots \times P(a_n = v_n \mid c_i)$$

Then the Naïve Bayes classifier calculates this product for each value of i from 1 to k and the classification with the highest value is chosen for predication

The formula in the algorithm combines the prior probability of c_i with the values of n possible conditional probabilities involving a test on the value of a single attribute.

$$P(c_i) \times \prod_{j=1}^{n} P(a_j = v_j \mid class = c_i)$$

When using the Naïve Bayes method to classify a series of unseen instances the most efficient way to start is by calculating all the prior probabilities and also all the conditional probabilities involving one attribute, though not all of them may be required for classifying any particular instance.

3.3. Performance Evaluation for Predictive Modeling

Once a predictive model is developed using the measles surveillance data, the model should be checked as to how it will perform for the future data that it has not seen during the model building process. The researcher has used three 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 [30]. A confusion matrix is a table of size m by m. An entry, CMi,j in the first m rows and m columns indicates the number of tuples of class i that were labeled by the classifier as class j. For a classifier to have good accuracy, ideally most of the tuples would be represented along the diagonal of the confusion matrix, from entry $CM_{1,1}$ to entry $CM_{m,m}$, with the rest of the entries being close to zero.

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

Figure 3.2: Simple confusion matrix

As shown in figure 3.2 a confusion matrix table of size two by two, the following measures can be calculated to measure the accuracy of the model, true positive rate, false positive rate, accuracy, Precision, recall, F – measure and ROC curve

The **true positive rate** of a classifier is estimated by dividing the correctly classified positives by the total positive count.

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

The **false positive rate** of the classifier is estimated by dividing the incorrectly classified negatives by the total negatives.

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

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

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

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

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

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

$$F - measure = \frac{2}{\frac{1}{precision} + \frac{1}{Recall}}$$

ROC (Receiver Operating Characteristics Analysis) curve

ROC analysis is performed by drawing curves in two dimensional spaces, with axes defined by the TP rate and FP rate [31]. The TP Rate and FP Rate values of different classifiers on the same test set are often represented diagrammatically by a ROC Graph.. On a ROC Graph, as shown in Figure 4.4, 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. The performance of different types of classifier with different parameters can be compared by inspecting their ROC curves.

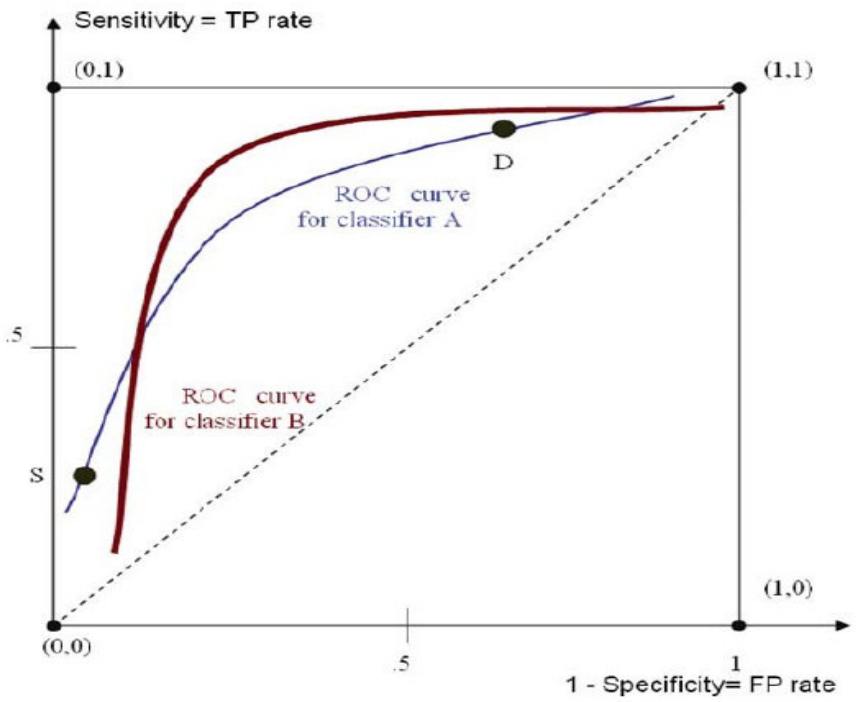


Figure 3.3. Example of ROC Curve for two classifiers

In order to decide which of the two classifiers in figure 3.3 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 3.3 for two classifiers, in this case the popular method called Area Under Curve (AUC) is used. 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 is. All the above measures were taken to measure the performance of models resulted from the three classifiers mentioned in section 3.1.

CHAPTER FOUR

BUSINESS UNDERSTANDING AND PREPROCESSING

Measles remains a leading cause of death among young children in the world, despite the availability of a safe and effective vaccine for the past 40 years. In 1999–2000, more than 873,000 measles deaths occurred annually. In response, the Measles Initiative was founded in 2001 as a long term partnership that aims to support the goal of reducing global measles deaths by 90 per cent by 2010 compared to 2000 estimates. Leading the effort is the American Red Cross, United Nations Foundation (UNF), Centers for Disease Control and Prevention (CDC), UNICEF, and WHO.

Infant and under five mortality rates in Ethiopia are among the highest in the world. Diarrheal diseases, vaccine preventable diseases (VPDs) and malnutrition are responsible for a majority of childhood deaths in Ethiopia. Measles is one of vaccine preventable diseases that cause catastrophic deaths. Therefore to expanded measles control, elimination, and possible eradication the Expanded Program on Immunization (EPI) started in Ethiopia in 1980 with the intention of increasing the Immunization coverage by 10% annually and reach 100% coverage in 1990 [31].

4.1. Business Understanding

The EPI program includes six vaccine preventable diseases including measles, diphtheria, pertussis, tetanus, polio and tuberculosis. The program is run by the Ministry of Health (MoH) in close cooperation with WHO, UNICEF and other partners and implemented in each region by the Regional Health Bureaus. WHO provides technical assistance to the Ministry of Health in resource planning, as well as social mobilization.

Main rationale of this research is applying data mining techniques on Measles surveillance data to detect continuing measles transmission and evaluate vaccination strategies. This research result helps improve measles control in identifying and

managing outbreaks. It also assists to predict outbreaks by identifying geographic areas and age groups at high risk.

From the organization founded the measles initiatives, the business understanding for WHO of Ethiopia is discussed in this section.

4.1.1. World Health Organization

WHO is the directing and coordinating authority for health within the United Nations system. It is responsible for providing leadership on global health matters, shaping the health research agenda, setting norms and standards, articulating evidence-based policy options, providing technical support to countries and monitoring and assessing health trends.

In the 21st century, health is a shared responsibility, involving equitable access to essential care and collective defense against transnational threats. WHO has a leading role in strategy development, consensus building, and programme monitoring. WHO provides the overall technical leadership and strategic planning for the management, coordination and monitoring of global measles control activities. WHO is also responsible for ensuring that all components of the measles mortality reduction strategy are technically sound and successfully implemented. Ethiopia is one of the WHO Africa region countries for measles mortality reduction program.

According to World Health Organization African Region (WHO AFRO), a measles outbreak is defined as five or more reported suspected cases of measles in a health facility or local government area in one month with a plausible means of transmission [4]. This defines the threshold that should elicit appropriate control measures like case finding, line listing of additional cases, improved case management and concentrated immunization activities in that local government area.

4.2. Measles Case Based Surveillance Data

Measles case-based surveillance is part of the national integrated disease surveillance and response (IDSR) system and a key component of the measles control program. Case-based measles surveillance was initiated in 2003 and since 2004, case-based laboratory supported surveillance has been implemented nationwide. Data on reported suspected measles cases are entered into a database, which is analyzed, on a weekly basis. In 2005, the reporting format was standardized so that all surveillance indicators could be monitored on a weekly basis [4]. It is a system whereby every suspected measles case should be detected and undergo laboratory investigation or the first five cases in the situation of outbreaks in WHO accredited laboratories.

Measles control strategies are based on the assumption that measles virus transmission occurs in chains of transmission of clinically recognizable measles cases. Surveillance systems rely on the identification of persons with the clinically recognizable symptoms of measles for detecting and responding to outbreaks, vaccinating susceptible contacts, and assessing the efficacy of vaccines and the impact of vaccination programs [32]. As illustrated in figure 4.1, the community should report any person with a fever and rash to the nearest health facility. Whereas, health care providers should suspect measles in any person with generalized rash and fever plus one of the following: cough or coryza (runny nose) or conjunctivitis (red eyes). Cases should also be reported in any person in whom a clinician strongly suspects measles. Health workers are expected to investigate and report all such cases and their possible contacts to the next higher level. Cases should be reported immediately through the IDSR reporting mechanism. In addition, blood (serum sample) should be collected from every isolated suspected case using a case-based investigation form or up to 5 cases per 30 days from each Woreda. The blood should be sent to the EHNRI (Laboratory Ethiopian health and nutrition research institute) for analysis and serologic testing. Six copies of the IDSR case-based reporting form should be completely filled. One copy should go with the specimen while the remaining copies should be filed at the reporting health facility, the Woreda health office, zonal health desk, regional health bureau, and Federal Ministry of Health focal person offices.

Patient-specific data such as age, sex, vaccination status and date of rash onset should be obtained from the patient and parents and provided in the case investigation form. During suspected outbreaks (if more than five suspected cases occur at a health facility or in a district within 30 days) serum specimens should be collected only from the first five. The remaining suspected measles cases should be line-listed and sent to the next higher level including to Federal Ministry of Health on the appropriate format. The main aim of this research is to build a model that predicts measles outbreak in Ethiopia using the measles surveillance data.

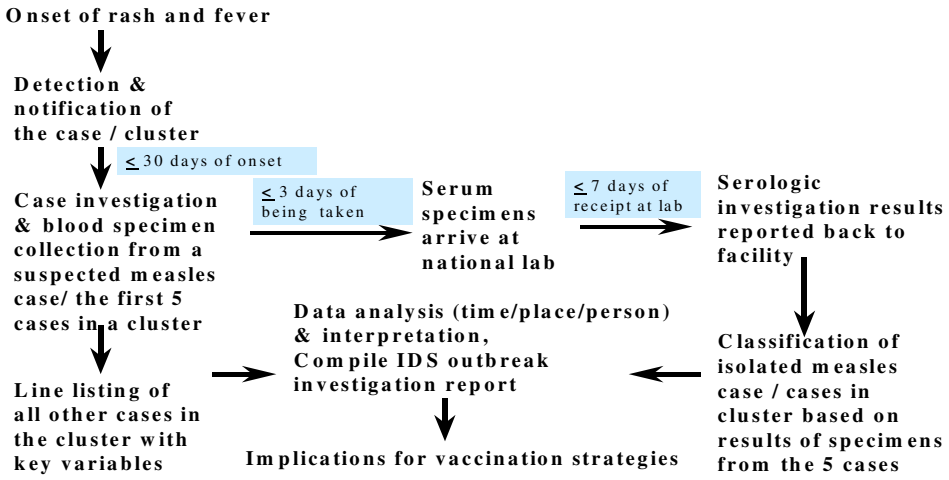


Figure 4.1 the process measles surveillance

Now that the measles campaigns have been completed in all areas of Ethiopia the specificity of clinically diagnosed measles is low. Therefore, suspected cases and suspected outbreaks should be serologically confirmed. To confirm measles diagnosis, 5 ml. of blood serum should be collected from all sporadic cases, and from the first five cases in suspected outbreaks occurring in a given location within 30 days.

From January 1 to June 30, 2005, 710 suspected measles cases were reported. For 597 (98%) of suspected cases (excluding epi-linked cases), adequate blood specimens were collected. In all, 73 (86%) of 85 Zones or special Woredas have reported at least one suspected measles case with a blood specimen. However, the annualized rate of suspected cases with blood specimens per 100,000 persons is currently at 1.63, which falls short of

the 2.0/100,000 population target. (Note that the performance target is increased from 1.0/100,000 to 2.0/100,000 in 2005.)[32].

To investigate measles outbreak the surveillance data collected on each reported case includes the age, date of onset of rash, possible source of infection, basic clinical information and the outcome (alive or dead). A blood sample is also taken for laboratory analysis from the initial cases in the outbreak to confirm that the outbreak is indeed due to measles. To facilitate data analysis, a line listing of cases is created. Where outbreaks were detected, further examination of the case based information and attempted to define the extent and cause of each outbreak is done [32].

4.3. Preprocessing

Data preprocessing is a stage where the data set is prepared to be useful for data mining purposes. The following steps are undertaken in the preprocessing stage; data cleaning, attribute and feature selection, and data transformation. The overriding objective is to minimize GIGO: to minimize the "garbage" that gets into our model so that we can minimize the amount of garbage that our models give out.

The initial measles surveillance data from the WHO database contains the following attributes listed in table 4.1.

No	Attribute name	Data type	Description
1.	ReportingHealthfacility	Text	Name of the Reporting health facility
2.	IdNumber	Text	Id number
3.	DateRecformdistric	Date/Time	Date district sent record
4.	DateReceivedNatlev	Date/Time	Date record received at national level
5.	NamesOfPatient	Text	Name of the patient
6.	DateOfBirth	Date/Time	Date of birth of the patient
7.	AgeInyears	Number	Age in years of the patient

No	Attribute name	Data type	Description
8.	AgeInmonths	Number	Age in month of the patient
9.	Sex	Text	sex
10.	PatientsResidence	Text	Patient's residence
11.	ReportingDistrict	Text	Name of Reporting district
12.	Towncity	Text	Town/ city of the patient
13.	Urbanrural	Text	Urban or rural
14.	DateSeenHealthFaci	Date/Time	The Date patient were seen at the health facility
15.	NumberOfVaccinedos	Number	Number of vaccine doses taken to the patient
16.	DateHealthfacility	Date/Time	Date health facility notified
17.	DateOfLastvaccinat	Date/Time	Date of last vaccination
18.	DateOfonset	Date/Time	date of onset of rash
19.	Inoutpatient	Text	Inpatient or outpatient
20.	Outcome	Text	Outcome (whether the patient died from measles or not)
21.	ProvinceOfResidenc	Text	Province of residence of the patient
22.	DateSentFormtodist	Date/Time	Date form sent to district
23.	DateSpecimencollec	Date/Time	Date of specimen collection
24.	FinalClassificatio	Text	Final classification whether result is 1-Confirmed by lab, 2-Confirmed by epi linkage, 3-Compatible/clin. no spec, 4-Discarded - IgM negative, 5-Pending lab results
25.	DateSpecimenSentto	Date/Time	Date specimen sent to laboratory
26.	DateLabReceivedSpe	Date/Time	Date laboratory received specimen
27.	SpecimenCondition	Text	Specimen condition whether specimen is 1-Adequate, 2-Not adequate
28.	MeaslesIgm	Text	Measlesigm whether measle 1-Positive, 2-Negative, 3-Indeterminate, 4-Not Done, 5-

No	Attribute name	Data type	Description
			Pending, 9-Unknown
29.	RubellaIgm	Text	Rubellaigm whether Rubella 1-Positive, 2-Negative, 3-Indeterminate, 4-Not Done, 5-Pending, 9-Unknown
30.	OtherLabResults	Text	Other lab results
31.	DateLabSentResultt	Date/Time	Date laboratory sent the result
32.	CountryCode	Text	Country code
33.	DistrictofResidenc	Test	District of residence
34.	Age	Number	Age
35.	DateDistrictRecLab	Date/Time	Date District recorded laboratory result
36.	Specimensource	Text	Specimen source
37.	Outbreak	Text	Outbreak whether there is 1-Yes, 2-No
38.	DataType	Text	Data type whether it is Case-based or line-list
39.	RecStatus	Number	Record status
40.	UniqueKey	Number	Unique key
41.	number	Number	Number
42.	vx_status	Text	Vaccine status whether the patient is vaccinated, unvaccinated or unknowm
43.	age_group	Text	Age group
44.	DT	Text	Data type whether it is 1-Case-based or 2-line-list

Table: 4.1 List of Variables in the Initial Dataset

4.3.1. Data Field Selection:

Discussing the importance of selecting relevant features (attributes) in any data mining task, Liu and Motoda [33] wrote that the abundance of potential features constitutes a serious obstacle to the efficiency of most learning algorithms. Popular methods such as k-nearest neighbor, C4.5, and back propagation are slowed down by the presence of many features, especially if most of these features are redundant and irrelevant to the learning

task." The authors further stated that some algorithms may be confused by irrelevant or nosily attributes and construct poor classifiers. Therefore, eliminating some attributes, which are assumed to be irrelevant to build the model can increase the accuracy of the classifier, save the computational time, and simplify results obtained.

Some of the data or attributes in the initial dataset was not pertinent to the data mining goal and were ignored. Of the variables given in table 4.1 *IdNumber* , *ReportingHealthfacility, DateRecformdistric, DateReceivedNatlev, NameOfPatient, ReportingDistrict, OtherlabResults, DateSentFormtodist, DateSpecimencollec, DateSpecimenSentto, DateHealthFacility, DateOfOnset, DateLabReceivedSpe, SpecimenCondition, DateLabSentResultt, CountryCode, DateDistrictRecLab, Specimensource, RecStatus, UniqueKey* and *number*, were ignored as having no data mining value based on the discussion with domain experts from WHO.

4.3.2. Data Cleaning

All raw data sets initially prepared for data mining are often large; many are related to human beings and have the potential for being messy. One should expect to find missing values, distortions, misreporting, inadequate sampling, and so on in these initial data sets. Thus Data cleaning routines attempt to fill in missing values smooth out noise while identifying outliers, and correct inconsistencies in the data.

After ignoring attributes that have no data mining value, the remaining attributes were checked for missing values, inconsistencies and other interpretable observations. The data collected had a huge number of variables with missing values. Table 4.2 summarizes variables and percentage (%) of missing values associated with each attribute.

No	**Attribute name**	**Percentage of missing values**
1.	AgeInyears	0.13%
2.	AgeInmonths	42.96%
3.	Sex	0.18%
4.	Urbanrural	26.46%

No	Attribute name	Percentage of missing values
5.	DateSeenHealthFacility	26.53%
6.	NumberOfVaccinedoses	0%
7.	MeaslesIgm	35.15%
8.	PatientsResidence	0.015%
9.	DistrictofResidence	0%
10.	ProvinceOfResidence	0%
11.	DateOfBirth	90.86%
12.	Towncity	34.35%
13.	Dateoflastvaccination	97.93%
14.	Dateofonset	0%
15.	FinalClassification	2.73%
16.	Vx_status	0%
17.	Age_group	20.8%
18.	InOutpatient	2.4%
19.	OutCome	0.076%
20.	RubellaIgm	35.15%
21.	DataType	0%
22.	DT	0%
23.	Outbreak	41.28%

Table: 4.2 List of Variables with their Missing Values

After identifying percentage of missing instances, attributes with a higher percentage of missing values have been removed from the dataset due to the fact that they may compromise the research goal. The removed attributes are namely DateOfBirtth (90.86%), Towncity (34.35%) and Dateoflastvaccination (97.93%). Attribute DT is also removed since it is a duplicate of the attribute datatype.

In the process of dealing with errors identified from the dataset and correcting those errors, filling in missing values, smoothing noisy data, identifying or removing outliers, and resolving inconsistencies was done for data cleaning. In exclusive to this research, the main goal being the prediction of measles outbreak in Ethiopia, Missing values are filled based on the idea of observing neighboring record values for resolving inconsistencies. Missing values for numeric attribute age is filled using the difference between date of birth and date seen at health facility. The nominal attribute sex has the modal value 'M' in the dataset therefore the 23 missing values are filled using this value 'M'. Under the dependent variables, Patient residence has 4 missing values; the universal constant value unknown was used to fill in the missing instances for the attribute. Another approach to work with missing value is removing instances with missing values, the independent variable outbreak has 41.28% missing values, 50% from the 41.28% 0f the missing instances were excluded from the dataset not to compromise the result of the data mining goal.

4.3.3. Data Transformation

Data transformation is necessary for two purposes to fix problems with the data such as missing values and categorical variables that take on too many values, and to bring information to the surface by creating new variables to represent trends and other ratios and combinations.

Attributes *Ageinyears* and *ageinmonth* contains the patients age which are of the same value, therefore attribute age was derived from the two variables by using ageinyear + (ageinmonth/12) formula. Furthermore reducing the cardinality of the attribute to a manageable size makes the result easy to interpretable and to design classifiers with better generalization capabilities. Therefore, attribute *age-cat* was derived from attribute age containing values (>1, 1-4, 5-9, 10-14, <15) these categories are based on the surveillance data and outbreak investigations to strengthen measles immunization programmes guide line.

Another attribute *Season* with values (summer, spring, winter, autumn) and *year* were created by considering the month values and the year value from *Date seen at health facility* attribute respectively. Finally missing values of *vx_status* were filled by considering *number of vaccine dozes* attributes and *vx_status* was taken to build the predictive model since it has less dimensionality than *number of vaccine dozes*. Table 4.3 provides summary of the original attributes and derived attributes with their values.

No	Original attributes	Derived attributes	Values
1	Age	Age_cat	>1, 1-4, 5-9, 10-14, <15
2	Date seen at health facility	year	2006, 2007, 2008, 2009, 2010, 2011
3	Date seen at health facility	Season	summer, spring, winter, autumn
4	Number of vaccine dozes	Vx_status	Vaccinated, unvaccinated, unknown

Table 4.3 Summary of Derived Attributed with Their Values

The final selected dataset with their description are summarized in Table 4.4.

No	Attribute name	Description
1.	Age-cat	Age category
2.	Sex	sex
3.	Vx_status	Whether the patient is vaccinated or not or not known
4.	MeaslesIgm	Measles igm test
5.	Season	Season where the patient seen to the health facility
6.	RubellaIgm	Rubella igm test
7.	Urbanrural	Patient resides in urban or rural area
8.	ProvinceOfResidence	Province of residence of the patient
9.	Inoutpatient	Whether patient is in patient or outpatient
10.	Datatype	Data type whether it is 1-Case-based or 2-line-list
11.	Finalclassification	Final classification whether result is 1-Confirmed by lab, 2-Confirmed by epi linkage, 3-Compatible/clin. no spec, 4-Discarded - IgM negative, 5-Pending lab results

No	Attribute name	Description
12.	year	Year seen at health facility
13.	outbreak	Whether there is an outbreak of measles or not

Table: 4.4 Final Selected Variables with Their Description

In conclusion, the final dataset ready to be used for the data mining purpose contains the 13 attributes and 15631 records, as depicted in table 4.4.

CHAPTER FIVE

EXPERIMENTATION

In this study an attempt was made to design a model that enables to predict the outbreak of measles in Ethiopia. To this end, J48 decision tree and Naïve Bayes classifiers were experimented on. WHO measles surveillance database is consulted to extract the dataset required for training and evaluating the models created by the classifiers. For creating prediction model a total size of 15631 datasets are used for training and testing. The validation is done using 10-fold cross validation and 70% split test option.

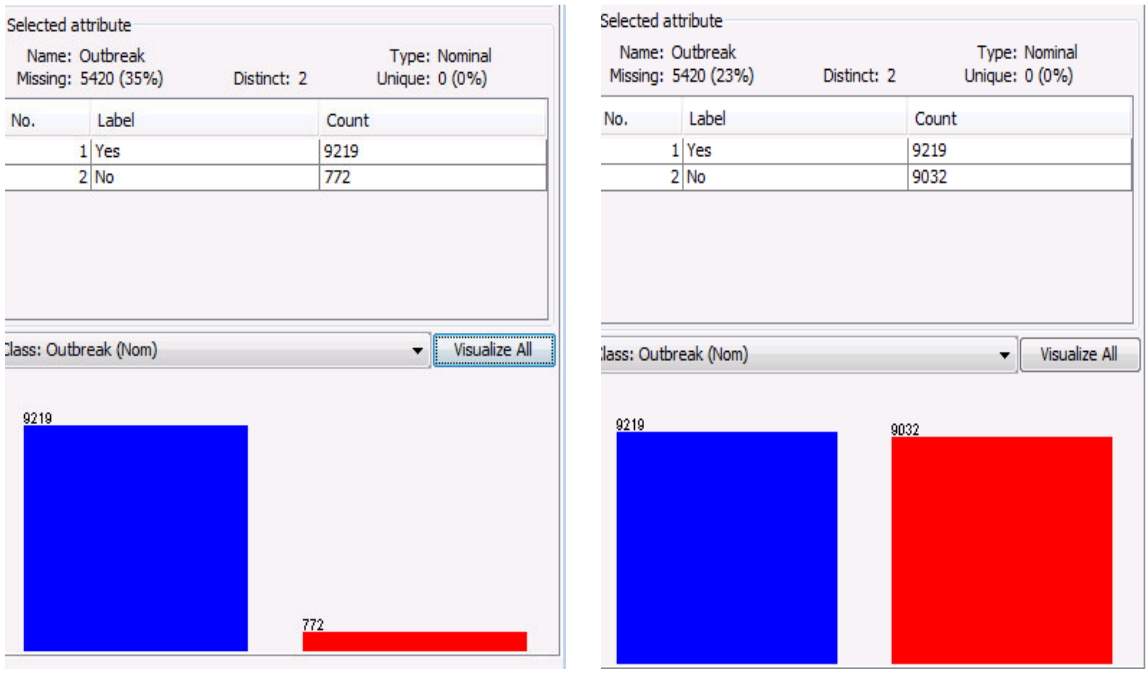
5.1. Dataset Preparation

The data collected for this research from WHO measles surveillance database was in Microsoft Access format. The dataset initially had 44 attributes and 26103 records but after the preprocessing stage, it was reduced to 13 attributes and 15631 records for building the predictive model. The data was extracted to Microsoft Excel for preprocessing purpose. It was then later converted to arff format which is compatible with Weka software.

After data is loaded in Weka, Automatic operations by filters like NumericToNominal are applied on three numeric attributes so that all the 13 attributes have nominal value. These attributes are namely *Inoutpatient, Finalclafication* and *year*.

A dataset is imbalanced if the classification categories are not approximately equally represented [34]. Performance of machine learning algorithms is typically evaluated using predictive accuracy. However, this is not appropriate when the data is imbalanced and/or the cost difference of error is large. In the case of measles surveillance data the class variable outbreak has a higher imbalance with ratio 1:12. Therefore, the researcher used SMOTE Automatic operations by filter where minority classes are oversampled by

generating synthetic examples of minority class and adding them to the dataset. This way, the class distribution in the dataset changes and probability of correctly classifying minority class increases.



(a) (b)

Figure 5.1 Side by side review of the class variable outbreak using SMOTE: (a) Original data; (b) balanced data using SMOTE

Figure 5.1 shows a side by side review of the class variable *outbreak* after SMOTE operation applied to the minority class. Originally there were 9219 records in the majority class and only 772 records in the minority class but after applying SMOTE the difference between the classes were reduced only to 197 records.

5.2. Model Building

In the initial experiment the researcher took 13 attributes for building predictive model. The selection of attribute is made using subjective judgment of the researcher and discussion with domain experts from WHO. To build the predictive model, the arff format of the selected dataset was given to Weka, J48 and naïve Bayes algorithm.

5.2.1. Predictive Model Building Using J48 Decision Tree

The type of classification selected for the first experimentation was J48 decision tree. The classifier has operation that are completely interactive and they benefit from powerful visualization features. The experiment on the decision tree predictive model building was based on the measles surveillance data from WHO, that has been preprocessed and introduced to the Weka software.

In the experiments, variable "*Outbreak*" was set as dependent variable and the remaining other attributes were set as independent variables. The classification tree was built using all the default parameters suggested by the Weka software. Six experiments were done with the J48 algorithm using different combination and numbers of attributes and inputting all the records with a 10-fold cross-validation mode, and inputting 70% of the records to train a model and then supply the unseen 30% of the record for testing the performance of the model. Furthermore the result obtained from these experiments is summarized in table 5.2 with their respective accuracies, true positive rate and true negative rate.

Experiment no	**No of attributes**	**True positive rate**		**True negative rate**		**Accuracy of the model**	
		10 fold	**70% split**	**10 fold**	**70% split**	**10 fold**	**70% split**
1	13	94.9%	94.8%	98.9%	99.3%	96.87%	97.06%
2	9	94.9%	94.8%	98.9%	99.3%	96.87%	97.06%
3	8	89.7%	87.5%	91.1%	92%	90.43%	89.76%

Table 5.1 Summary of the Three Decision Tree Experiment Results

The result of the decision tree model in experiment one shows that though each record in the dataset included 13 attributes including dependent variable, the decision tree selected 9 variables. These variables used in the model are *Datatype, Finalclassification, Provinceofresidence, Rubellaigm Inoutpatient, Sex, Vx_Status, Season, Urbanrural, Outcome, Age_Cat, Measlesigm* and *Outbreak*. It excluded *RubellaIgm, OutCome, InOutPatient* and *FinalClassification* variable as having no statistical significance to classify records to the predefined class.

The model has accuracy of 96.87% using 10 fold cross validation and 97.06% accuracy using 70% split test options. Moreover the model has a true positive rate of 94.9%and true negative rate of 98.9%for 10 fold cross validation and 94.8%true positive rate and 99.3% true negative rate for 70% split test. In the mean time the second experiment used 9 attribute that experiment one identified to be statistically significant to construct the decision tree. It exhibited the same accuracy, true positive rate and true negative rate as to that of experiment one. The third experiment used 8 attributes namely *Age-Cat, Sex, Vx_Status, Season, Provinceofresidence, Year, Urbanrural* and *Outbreak*. The 10 fold cross validation experiment scored 90.43% accuracy, 89.7% true positive rate and 91.1% true negative rate. Whereas the 70% split test scored 89.7% accuracy, 87.5% true positive rate and 92%true negative rate.

The best decision tree model produced was from experiment 2. The model shows a better performance evaluation than other models. The 70% split test model also scored a better performance than 10 fold cross validation. Therefore, the 9 attributes, used to build the decision tree for experiment 2 with 70% split test option, are taken to be statistically significant in splitting the decision tree. Furthermore, opinions gathered from the domain experts from WHO indicated that these attributes have a great role in the prediction task.

Figure 5.2 shows the tree view of the predictive model built using J48 algorithm using 9 attributes. For clear understanding of the tree, the run information for the model is annexed at annex III.

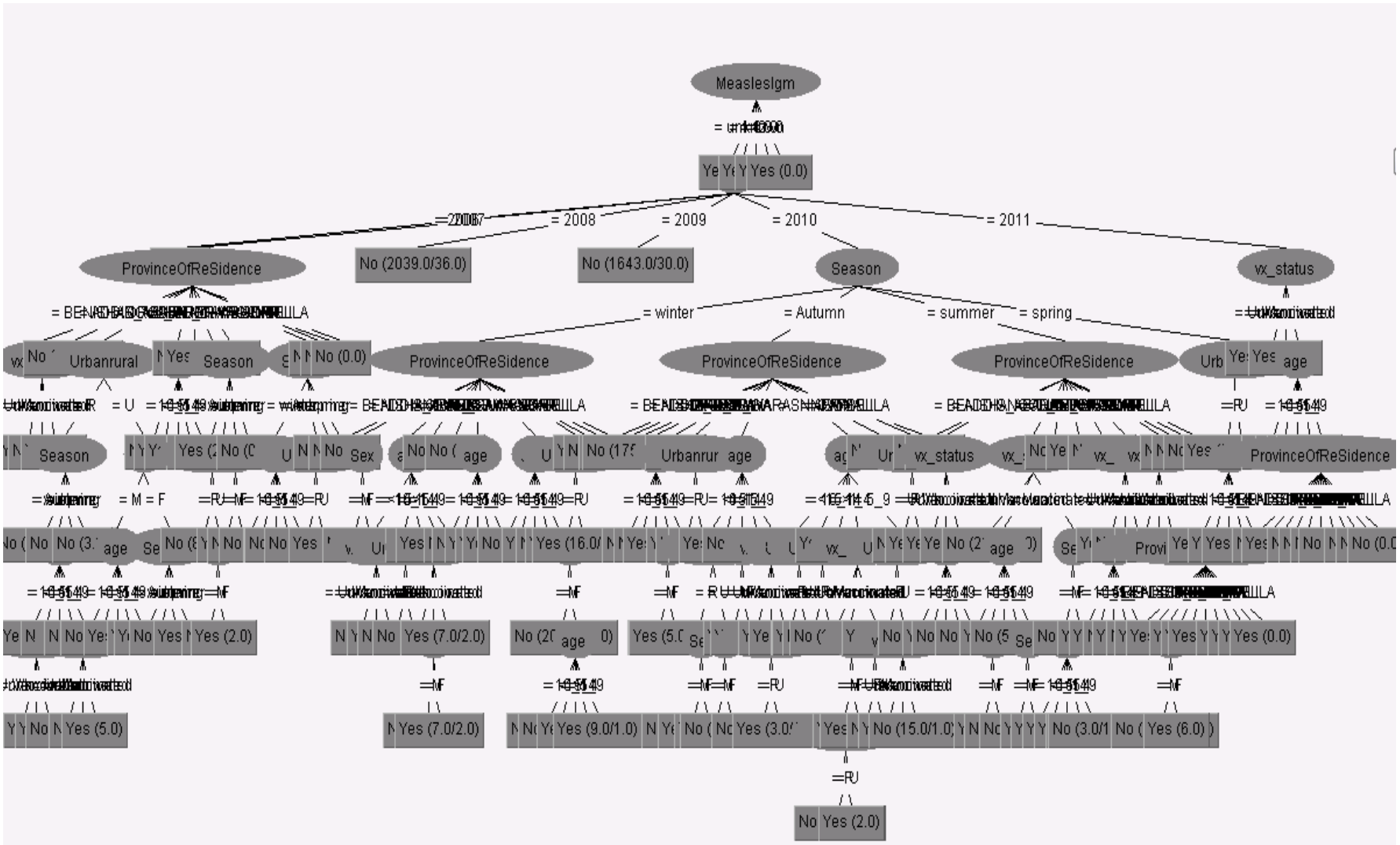


Figure 5.2 Tree View of the Predictive Model Using J48 Algorithm

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 the decision tree shown in table 5.2 illustrate that out of the total 5444 records 2556 records were correctly classified as "yes" and 2728 records were correctly classified as "no". The classifier incorrectly classified 140 records as "no" and 20 records as "yes" while in fact they belong to "yes" and "no" class respectively. In general the classifier correctly classified 5284 records and incorrectly classified 160 records out of a possible 5444. It has ignored 1671 missing Instances while building the model. The accuracy of the classifier to correctly predict the class value as "yes" and no is 97.061%.

=== Confusion Matrix ===		
a	b	← classified as
2556	140	a = Yes
20	2728	b = No

Table 5.2 Confusion Matrix for J48 Decision Tree Model

In order to see how well the classifier can recognize "yes" tuples (the positive tuples) and how well it can recognize "no" tuples (the negative tuples), the sensitivity and specificity measures can be used. Sensitivity is also referred to as the true positive (recognition) rate, while specificity is the true negative rate. Furthermore the classifier has 94.8% sensitivity and 99.3% specificity which discloses that decision tree classifier has an acceptable capability of recognizing the true class values.

ROC Analysis for J48 Decision Tree Model

ROC analysis provides tools to select possibly optimal models and to discard suboptimal ones. ROC analysis is related in a direct and natural way to cost/benefit analysis of

diagnostic decision making. Figure 5.3 shows the area under ROC for the Outbreak Instances. Class value yes gives the ROC accuracy of 91.44%.

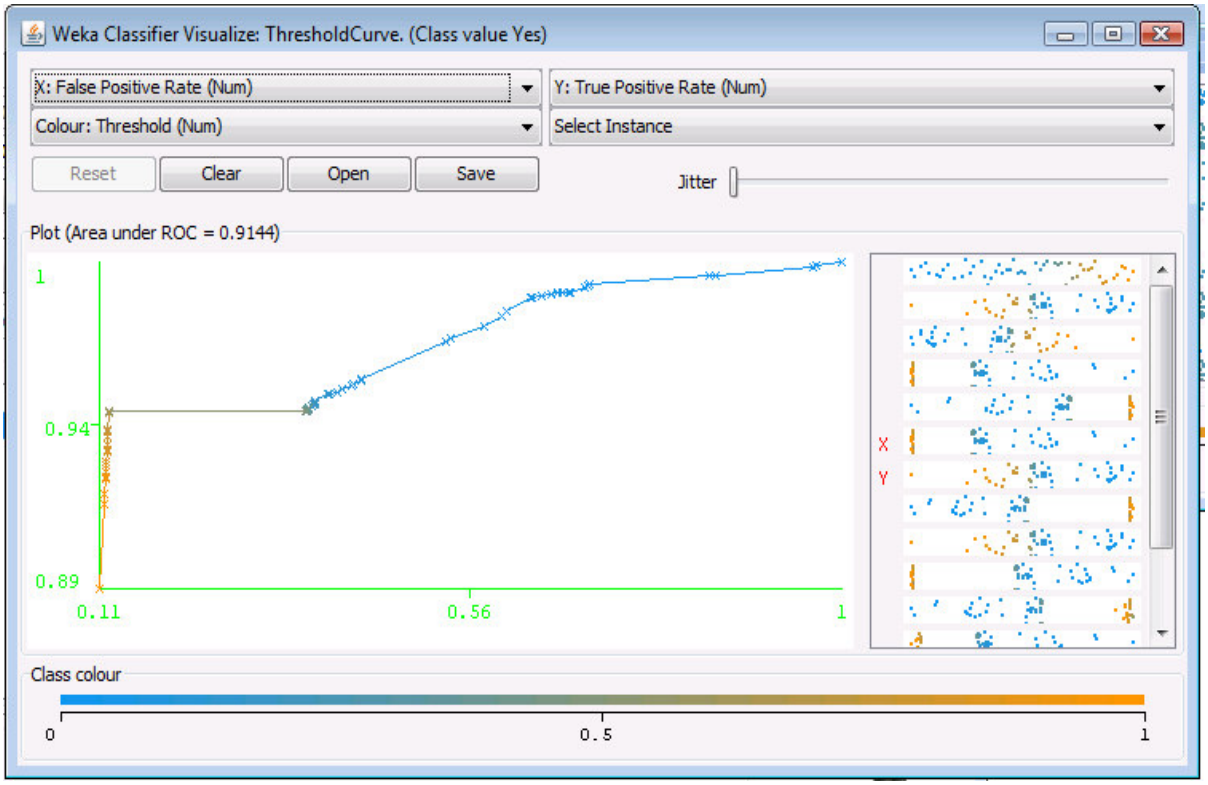


Figure 5.3 ROC curve of the decision tree model

In the above figure the vertical axis of ROC curve represents the true positive rate. The horizontal axis represents the false-positive rate. As shown in figure 5.3, initially the curves moves steeply up from zero showing that there are more true positives than false positive. Later, as we start to encounter fewer and fewer true positives, and more and more false positives, the curve cased off and became more horizontal. The area under the curve for the model is 0.9144 which is closer to 1. Consequently the corresponding model has an outstanding accuracy in classifying the data.

5.2.2. Predictive Model Building Using Naïve Bayes Classifiers

The type of classification model selected for the second experiment to build the predictive model was Naïve Bayes. Six experiments were done with the naïve bayes algorithm using different combination and numbers of attributes and inputting all the records with a 10-fold cross-validation mode, and inputting 70% of the records to train a model and then supply the unseen 30% of the record for testing the performance of the model. Table 5.2 summarizes the result obtained from these experiments with their respective accuracies, true positive rate and true negative rate.

Experiment no	**No of attributes**	**True positive rate**		**True negative rate**		**Accuracy of the model**	
		10 fold	**70% split**	**10 fold**	**70% split**	**10 fold**	**70% split**
1	13	86.1%	86.4%	100%	100%	93%	93.25%
2	9	86.6%	86.6%	99.7%	99.7%	93.1%	93.31%
3	8	69.3%	70.5%	69.9%	69.2%	69.5%	69.85%

Table 5.3 Summary of Naïve Bayes Experiment Results

From table 5.3 we can see that experiment one used 13 attributes and the model scored 93% accuracy, 86.1% true positive rate and 100% true negative rate with 10 fold cross validation and 93.25% accuracy, 86.4% true positive rate and 100% true negative rate with 70% split. The second experiment used 9 attributes and scored the best accuracy of 93.31% with 70% split test option from the other experiments. Experiment three scored the list accuracy with both test option. Even though experiment one has a higher true negative rate with both test options; experiment two has a higher true positive rate and accuracy than experiment one and three. Hence experiment two with 9 attributes and 70%

split test option were chosen to have a better performance evaluation for naïve bayes classifier.

Confusion Matrix for Naïve Bayes Classifier

Table 5.4 shows the confusion matrix for model built using Naïve Bayes in experiment two. The confusion matrix depicts that the amount of records that is incorrectly classified as "yes" outbreak by the model is only 8 with percentage close to 0. Whereas the amount of records that is incorrectly classified as "no" there is no outbreak is 356 with percentage of 11.5%.

=== Confusion Matrix ===		
a	b	←classified as
2340	356	a = Yes
8	2740	b = No

Table 5.4 confusion matrix for Naïve Bayes model

In order to access how well the classifier can recognize "yes" tuples (the positive tuples) and how well it can recognize "no" tuples (the negative tuples), the sensitivity and specificity measures were used. The classifier has 86.8% sensitivity which shows that the model has acceptable capability of recognizing the true positive value of the class "yes" and 99.7% specificity which reveal that naive bayes classifier has an immense capability of recognizing the true class value "no".

ROC Analysis for Naïve Bayes Classifier

The area under ROC for the Outbreak Instances produced from the Naïve Bayes classifier is shown in Figure 5.4. The vertical axis of ROC curve represents the true positive rate. The horizontal axis represents the false-positive rate. Class value yes gives the ROC accuracy of 95.63%.

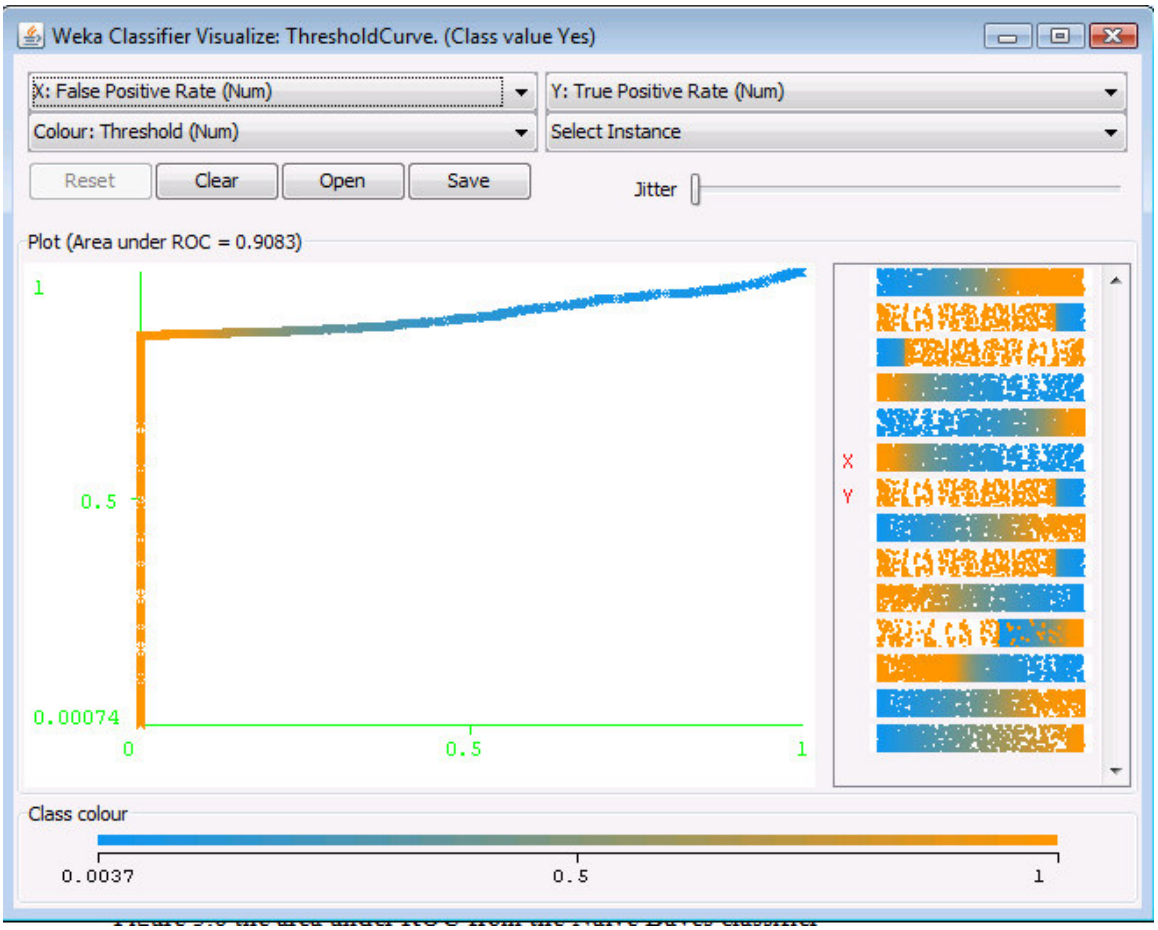


Figure 5.4 the area under ROC from the Naïve Bayes classifier

In figure 5.4, at first the curve moves sharply up from zero showing that there are more true positives than false positive. Then the curve starts to become more horizontal as it encounters less true positives, and more false positives. The area under the curve for the model is 0.9083 which is closer to 1. Consequently the corresponding model has a good accuracy in classifying the data.

5.3. Discussion

In this study, experiments have been carried out with two classification algorithms, i.e. J48 algorithm and Naïve Bayes classifier to build a model that predicts measles outbreak in Ethiopia. From the experiments 9 attributes were identified to make sound rule and

better accuracy. The results of the experiment performance for J48 and Naïve Bayes are summarized in Table 5.5, including their accuracy, sensitivity, specificity and area under the ROC. Comparison of the performance evaluation results in table 5.5 between J48 and Naïve Bayes are illustrated in Figures 5.5.

Testing criteria	J48	Naïve Bayes
Accuracy (%)	97.06	93.31
Sensitivity (%)	94.8	86.8
Specificity (%)	99.3	99.7
Area under the ROC (%)	91.44	90.83

Table 5.5. Performance Summary of J48 and Naïve Bayes Classifiers

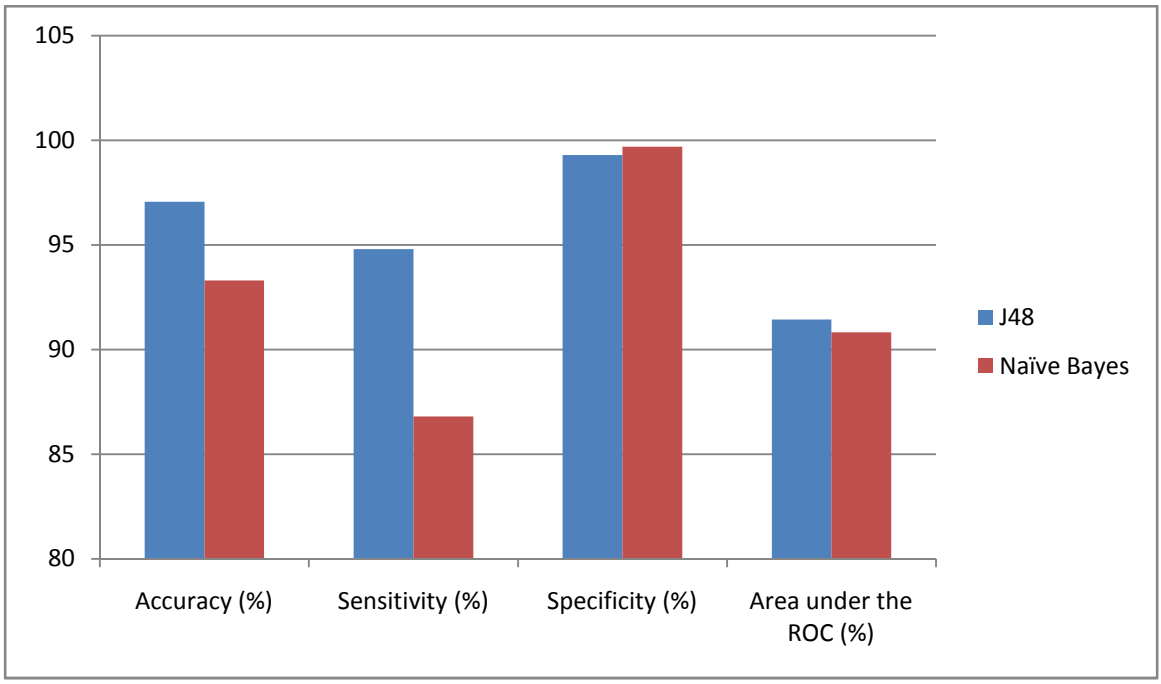


Figure 5.5 Experimental Result Summaries of J48 and Naïve Bayes Classifiers

The scores of the Naïve Bayes classifiers for correctly classifying the true Negative (No-outbreak) is slightly higher than J48. However, the overall score of the Naïve Bayes model is lower than that of the decision tree model. The Naïve Bayes fails to recognize

true positive (Yes-outbreak) as good as that of No-outbreak. These could be due to the reason, that there are missing values in the class variable. The J48 decision tree model has shown adequate capability of recognizing true positive and true negative of the class value.

Naïve Bayes works well when tested on actual datasets, particularly when combined with attribute selection procedures that eliminate attributes redundancy. Dependencies between attributes inevitably reduce the power of Naïve Bayes to differentiate what is going on. They can, however, be eliminated by using a subset of the attributes in the decision procedure, making a careful selection of which ones to use.

Although there is an agreement between the decision tree and naïve bayes models, they disagree on the classification of several outbreaks. Some outbreaks are misclassified by one of the models though not by the other as discussed in the next section. However for this research, as illustrated in figure 5.5, J48 classifier has a higher accuracy, sensitivity and area under the ROC than Naïve Bayes classifiers which makes it a better classifier for the given measles surveillance data.

5.4. Classifier error

There is a lot to be learned from closely examining the errors made by a classification model. These errors represent the difference between what the model predicts and what the actual outcome turns out to be in the real world. Whenever a model turns out to be worth considering for application, the next step is to examine why classification errors occur in the test dataset. Sometimes, the predicted and actual value may differ in predicting a record to a certain class label. This shows that the record that is labeled by an expert to one class may be labeled by the classifier to other class. This kind of phenomenon often reduces the performance of the system. The classifier mostly predicts the records in to a certain class as there are similar attributes that lie in the same class boundary. But the attribute that determines the class boundary of the given record is suppressed because of the data-driven (attribute similarity) trend applied by the classifier

The model built using J48 classifier with 70% split test option used a total 5444 records for testing the model performance. 2556 records and 2728 records were correctly classified as “yes” and “no” respectively. The classifier incorrectly classified 140 records as “no” and 20 as “yes”. Table 5.6 demonstrates sample of records that are incorrectly classified with those that are correctly classified by the model.

Sample no.	**age**	**Sex**	**Season**	**Year**	**Measles Igm**	**vx_status**	**ProvinceOf ReSidence**	**Urban rural**	**Outbreak**	
									predicted	**expert**
1.	1_4	M	Autumn	2010	1.0	UnVaccinated	SNNPR	R	No	No
	1_4	M	Autumn	2010	1.0	UnVaccinated	OROMIA	R	No	Yes
2.	<15	M	Autumn	2010	1.0	unknown	AMHARA	U	Yes	No
	<15	M	spring	2010	1.0	unknown	OROMIA	U	Yes	Yes
3.	<15	M	winter	2008	1.0	unknown	OROMIA	U	No	Yes
	<15	M	winter	2007	1.0	unknown	OROMIA	U	No	No
4.	1-4	M	Autumn	2010	1.0	unknown	OROMIA	R	No	No
	1-4	M	Autumn	2010	1.0	UnVaccinated	OROMIA	R	No	Yes

Table 5.6 Sample of Record That Shows Classifier and Expert Judgments Variation

As shown in table 5.6 in sample 1, the expert classified the existence and non existence of measles outbreak based on province of residence of the patient while the remaining records have the same value. Then the Classifier wrongly classified the existence of measles outbreak as no based on the similar values without considering the province of residence of the patient which has a different value. Here what determined the class label is the province of residence of the patient. The classifier took the similarity of all the attributes disregarding the difference in province of residence and classified them as no measles outbreak while in fact the expert classified them as yes there is measles outbreak. Similarly in sample 3, the classifier classified existence of an outbreak as no without considering the year patient went to health facility when in fact the expert solely depends on it to classify the model as yes.

In sample 2, the classifier classified the existence of measles outbreak as yes without considering the season and province of residence of the patient values while the expert used these factors to classify the non existence of measles outbreak. In wrapping up our discussion, the misclassification of the model were due to the underestimating of a single attributes value and also taking the similarity of the other attributes as the predominant predictive values.

5.4. Generating Rules from Decision Tree

After consecutive experiments in building the best decision tree model, the next step is to generate, rules by tracing through the branches up to leafs. A rule is a correlation found between the main variable (dependent) and the others (independent).The corresponding rules extracted form decision trees is listed below and Some of the rules believe to be interesting are randomly selected and presented as follows.

Rule 1:

IF MeaslesIgm = 1.0 year = 2007 and ProvinceOfReSidence = BENISHANGUL_G and vx_status = unknown: THEN Yes (9.0)

Rule 2:

IF MeaslesIgm = 1.0 year = 2007 and ProvinceOfReSidence = OROMIA and Urbanrural = R and Season = Autumn and age = 10_14: THEN Yes (11.0/3.0).

Rule 3:

IF MeaslesIgm = 1.0 year = 2007 and ProvinceOfReSidence = OROMIA and Urbanrural = U and Sex = F and Season = Autumn: THEN Yes (13.0/1.0)

Rule 4:

IF MeaslesIgm = 1.0 year = 2007 and ProvinceOfReSidence = AMHARA and Season = Autumn: THEN Yes (24.0/3.0)

Rule 5 :

IF MeaslesIgm = 1.0 and year = 2010 and Season = winter and ProvinceOfReSidence = OROMIA and age = <15 and vx_status = unknown: THEN Yes (8.0)

Rule 6:

IF MeaslesIgm = 1.0 and year = 2010 and Season = winter and ProvinceOfReSidence = OROMIA and age = 10_14 and Urbanrural = R: THEN Yes (10.43/1.0)

Rule 7:

IF MeaslesIgm = 1.0 and year = 2007 and ProvinceOfReSidence = OROMIA and Urbanrural = R and Season = winter : THEN No (333.0/11.0)

Rule 8:

IF MeaslesIgm = 1.0 and year = 2010 and Season = winter and ProvinceOfReSidence = OROMIA and age = 5_9: THEN Yes (24.0/5.0)

Rule 9:

IF MeaslesIgm = 1.0 and year = 2010 and Season = winter and ProvinceOfReSidence = SNNPR and Urbanrural = R and Sex = F and age = 1_4: THEN Yes (10.0/2.0)

Rule 10

IF MeaslesIgm = 1.0 and year = 2010 and Season = winter and ProvinceOfReSidence = SNNPR and Urbanrural = U: THEN Yes (16.0/1.0)

Rule 11

IF MeaslesIgm = 1.0 and year = 2010 and Season = Autumn and ProvinceOfReSidence = AMHARA and age = <15 and Urbanrural = R and Sex = F: THEN Yes (21.0/2.0)

Rule 12

IF MeaslesIgm = 1.0 and year = 2010 and Season = Autumn and ProvinceOfReSidence = AMHARA and age = 10_14: THEN Yes (32.0/6.0)

Rule 13

IF MeaslesIgm = 1.0 and year = 2010 and Season = Autumn and ProvinceOfReSidence = AMHARA and age = 5_9 and vx_status = unknown: THEN Yes (10.0/3.0)

Rule 14

IF MeaslesIgm = 1.0 and year = 2010 and Season = Autumn and ProvinceOfReSidence = SNNPR and age = 1_4 and vx_status = Vaccinated and Urbanrural = R: THEN Yes (10.0/3.0)

Rule 15

IF MeaslesIgm = 1.0 and year = 2010 and Season = summer and ProvinceOfReSidence = AMHARA and vx_status = UnVaccinated: THEN Yes (12.0/2.0)

Rule 16

IF MeaslesIgm = 1.0 and year = 2010 and Season = spring and Urbanrural = R and age = <15 and ProvinceOfReSidence = AMHARA: THEN Yes (22.0/1.0)

Rule 17

IF MeaslesIgm = 1.0 and year = 2010 and Season = spring and Urbanrural = R and age = 10_14: THEN Yes (17.0/1.0)

Rule 18

MeaslesIgm = 1.0 and year = 2010 and Season = spring and Urbanrural = R and age = 1_4: Yes (60.79/3.0)

Rule 19

MeaslesIgm = 1.0 and year = 2010 and Season = spring and Urbanrural = R and age = 5_9: THEN Yes (29.0/1.0)

Rule 20

MeaslesIgm = 1.0 and year = 2010 and Season = spring and Urbanrural = U: Yes (262.82/1.0)

Rule 21

MeaslesIgm = 1.0 and year = 2011 and vx_status = unknown: Yes (38.0/1.0)

Rule 22

MeaslesIgm = 1.0 and year =2011 and vx_status = UnVaccinated: Yes (31.0)

Rule 23

MeaslesIgm = unknown: Yes (7872.0)

Rule 24

MeaslesIgm = 4.0: Yes (67.0)

The first rule shows that if the patient is positive for MeaslesIgm test and has unknown vaccine status and the year patient went to the health facility is 2007 and Region is Benishagul Gumuz and then there is a higher probability that measles outbreak will occur in those conditions. The 7th rule indicated that if MeaslesIgm test is positive and year is 2007 and region is Oromia and setting is in Rural and Season is winter then there is a higher probability that outbreak will not occur on those circumstances.

It can be revealed from the above rules that most important variables for building model to classify the occurrence of measles outbreak in Ethiopia were MeaslesIgm, year, Season, ProvinceOfReSidence, vaccine status and age. Therefore these attributes play a significant role in classifying records at the higher level of the tree which indicates their statistical significance than other variables like sex and Urbanrural.

5.5. Discussions of Results on occurrence of measles outbreak

The rules generated from the decision tree in section 5.4 predict the occurrence of measles outbreak in Ethiopia. The rule considers difference conditions of the attributes age, Sex, Season, year, Urbanrural, MeaslesIgm, vx_status and ProvinceOfReSidence to predict for the class value outbreak. The attributes MeaslesIgm, year, Season, ProvinceOfReSidence, vaccine status and age were identified as having a higher statistical significance in classifying the predicted value for outbreak.

Amhara, Oromia, SNNPR and Benishangul Gumuz regions were identified to be the most susceptible for the occurrence of measles outbreak. This could be due to the fact that these regions occupy the highest population density in the country comprising 19249514, 30905622, 16704782 and 755046 population size respectively.

In summer the incidence of measles outbreak is very rare according to the prediction of the model. And there seem to be a pick outbreak occurrence in autumn, winter and spring. This could be to due to the reason that during these season children are engaged in school which result contact with different children from different background that raises the contagiousness of the airborne viral diseases.

The rule shows that the majority of outbreaks occur when measles Igm test is unknown. This rule is justified by domain experts as when measles outbreak is suspected serum specimens is collected only from the first five cases. The remaining suspected measles cases is line-listed and conformed for measles by epidemiological link without further testing their measles igm status.

It is indicated in the rules that age groups between 1-4, 5-9, 10-15, and <15 were found to be the most vulnerable group for measles outbreak. Member of this age group are more interactive in the community which increase their contagious risk and they may not be vaccinated in earlier age and therefore they are susceptible for measles outbreak.

CHAPTER SIX

CONCLUSION AND RECOMMENDATIONS

6.1. Conclusion and Summary

6.1.1. Summary

Data mining is extracting meaningful patterns and rules from large quantities of data. It is clearly useful in any field where there are large quantities of data and something worth learning. In this respect, widespread use of medical information systems and explosive growth of medical databases require traditional manual data analysis to be coupled with methods for efficient computer-assisted analysis. Extensive amounts of data gathered in health care databases require specialized tools for storing and accessing data, for data analysis, and for effective use of data. Medical informatics may use the technologies developed in the new interdisciplinary field of knowledge discovery in databases (KDD), and particularly data mining. Today, numerous health care organizations are using data mining tools and techniques in order to raise the quality and efficiency of health-related products and services.

This research tried to investigate the potential applicability of data mining technology in developing a model to predict the occurrence of measles outbreaks in Ethiopia, so that it can support the effort to control measles outbreak in Ethiopia.

This investigation was conducted according to the Hybrid process model. The data was collected from WHO measles surveillance database organized from 2006 to 2011 for the research purpose. Analyzing the large volume of measles surveillance data and extracting useful information and knowledge for decision making about measles outbreak prediction was done. First the data was preprocessed for data cleaning, attribute and feature selection, and data transformation.

This experimental research, which engaged a hybrid methodological approach, made use of two predictive modeling techniques, J48 decision tree and Naïve Bayes, to address the problem. The experiment result shows that J48 decision tree outperformed naïve bayes classifier. Hence J48 decision tree with 97.06% accuracy prediction model building was selected to extract interesting rules to cite.

6.1.2. Conclusion

In conclusion, results from the study have shown that the problem of predicting measles outbreak in Ethiopia could be supported by the use of data mining, in particular with decision trees technique. Moreover, further extensive experiments at district level and using various data sources available with those organizations working in related public health will enhance the result obtained in this study.

6.2. Recommendation

This research work was carried out for academic purpose and it has revealed the potential applicability of data mining technology to predict measles outbreak in Ethiopia using measles surveillance data. Accordingly, based on the findings of this research work, the researcher forwards the following recommendations for future work particularly in relation to the possible application of data mining technology in supporting the effort to measles in Ethiopia.

- Results of this research could be improved through extensive tests and use of other prediction techniques such as neural networks, support vector machine or a combination of them. So further experiments need to be done for better classification performance.

- The measles surveillance data has a large quantity of missing values and there is inconsistency in filling out the required information in the dataset. Therefore there is a need to investigate the possibility of designing a classifier that works better given incomplete surveillance data.

- Although in this study encouraging results were obtained, the experiment is made only using data taken from WHO measles surveillance database. Since there are numerous other organization that has direct or indirect effect on the measles control program like Federal Ministry of Health, Ethiopian Health and Nutrition Research Institute, UNICEF, National meteorological Agency, further investigation should be done by integrating the various measles data sources.

- In this study the prediction of measles outbreak was done at regional level. The researcher proposed to continue the research at district level in order to get a clear alertness about the locations of the measles outbreak.

- There is a need to design knowledge base system in order to organize the knowledge extracted using data mining from measles surveillance database for providing the necessary advisory service for experts, researchers and organizations working in measles control program in Ethiopia.

- There is a need to make a data warehouse for the measles surveillance data in WHO and other related organizations that works with measles control program so that further researches using data mining technology can easily be achieved in the future.

REFERENCE

1. T. M. Akande. (2007.) A review of measles vaccine failure in developing countries. Nigerian Medical Practitioner. 52 No 5-6, Available at URL: http://www.ajol.info/index.php/nmp/article/viewFile/28917/5243 (accessed on December 28, 2010).
2. Aaby, P., Clemens, C., (1989). Measles Immunization research: a review. Bulletin of the World Health Organization.
3. Health Security and Environment. World Health Organization. Geneva 27, Switzerland.
4. Federal Ministry of Health and WHO Ethiopia. (2007). National Guidance for Measles Surveillance and outbreak investigation
5. World Health Organization (2009). Measles Fact Sheet No 286. Geneva Switzerland. Available at URL: http://www.who.int/mediacentre/factsheets/fs286/en. (Accessed on: December 20, 2010)
6. GAVI (2010). Support for civil society organizations in Ethiopia. Available at URL: http://www.gavialliance.org/resources/CSO_Ethiopia.pdf (accessed on January 3, 2010)
7. David H., Heikki M. and Padhraic S. (2001). Principles of Data Mining: The MIT Press.
8. Infectious diseases: Measles. Available at URL: http://www.infection-research.de/infectious_diseases/masern/ (accesses on February 9, 2011).
9. WHO. Measles pre-elimination. Available at URL: http://www.afro.who.int/en/clusters-a-programmes/ard/immunization-and-vaccines-development/programme-components/measles-pre-elimination.html (accessed on march 3, 2011)
10. Tadesse H., Deribew A., Woldie M. (2008). Predictors of defaulting from completion of child immunization in south Ethiopia. A case control study. BMC Public Health, Ethiopia Available at URL: http://www.biomedcentral.com/1471-2458/9/150 (accessed on January 15, 2011)

11. Neghist T. (2009). Progress of Measles Control in Ethiopia.
12. Krzysztof J., Witold P., Roman W., Lukasz A. (2007). Data Mining: A Knowledge Discovery Approach. New York: Springer-Verlag New York Inc.
13. R. G. Feachem, D. T. Jamison. (1991). Disease and Mortality in Sub-Saharan Africa. Washington, DC. Available at URL: http://www.dcp2.org/file/66/Disease%20and%20Mortality%20in%20SSA.pdf (accessed on January 6, 2011)
14. W. J. Frawley, G. P. Shapiro, C. J. Matheus. (1991). Knowledge Discovery in Databases: An Overview; Knowledge Discovery in Databases 1-27; AAAI Press,
15. Daniel T. L. (2005). Discovering knowledge in data: An introduction to data mining. Canada: John Wiley & sons, inc.
16. Han, Jiawei and Kamber, Micheline. (2001). Data Mining: concepts and Techniques. San Fransisco: Morgan kufman.
17. Susan p. (2001). Effective Use of the KDD Process and Data Mining For Computer Performance Professionals. Int. CMG Conference
18. Rea, Allan. (2002). Data Mining: An introduction Student Notes. Available URL: http://www.pcc.qub.ac.uk/tec/courses/datamining/stu_notes/dm_book_1.html (Accessed on: February 5, 2011)
19. Max B. (2007). Principles of Data Mining. : Springer.
20. Brachman, R. J. & Anand, T. (1996). The process of knowledge discovery in databases.
21. Prather, Jonathan C. et. al. (2001). Medical Data Mining: Knowledge Discovery in a clinical Data Woehouse. Available at URL: http://www.amia.org/pubs/symposia/D004394.PDF (accessed on February 3, 2011)
22. Larvac, Nada. (1998). Data Mining in Medicine: Selected Techniques and Applications. Available URL.: http://citeseer.nj.nec.com/lavrac98data.html (Accessed on: February 3, 2011)
23. P. J. Kellogg , S. D. Bale , F. S. Mozer , T. S. Horbury ,and H. Reme. Methodologic and approaches to spiral Epidemiology. Available at URL:

http://www.scribd.com/doc/6301518/Methodologic-Issues-and-Approaches-to-Spatial-Epidemiology (accessed on April 25, 2011)

24. Wong, W.K., Moore, A., Cooper, G. and Wagner, M., (2005). What's Strange About Recent Events (WSARE): An Algorithm for the Early Detection of Disease Outbreaks. Journal of Machine Learning Research. 6, 1961-1998.
25. Lloyd - Williams, Michael. (1997). Discovering the Hidden secrets in your Data - the data Mining approach to Information.
    Available at URL: http://informationr.net/ir/3-2/paper36.html (accessed on January 11 , 2011)
26. Shegaw A. (2002). Application of data mining technology to predict child mortality patterns: the case of butajira rural health project (BRHP), M.Sc. Thesis, Addis Ababa University, Ethiopia.
27. Helen T. (2003). Application Of Data Mining Technology To Identify Significant Patterns In Census Or Survey Data: The Case Of 2001 Child Labor Survey In Ethiopia, M.Sc. Thesis, Addis Ababa University, Ethiopia
28. M. Hemalatha, S. Megala, Mining Techniques in Health Care: A Survey Of Immunization, Journal of Theoretical And Applied Information Technology vol 25. No. 2 -- 2011
29. Curet, Jackson and M, Tarar. A Designing and Evaluating a case-based learning and reasoning agent in unstructured decision making, Information Intelligence and Systems.
30. Jiawei H., Micheline K. (2006). Data Mining: Concepts and Techniques. San Francisco: Morgan Kaufmann
31. 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.
32. Ethiopian IDSR/AFP/EPI newsletter. Diseases prevention and control department and family health department, MoH: June 2005, 15.
33. Liu H., Motoda H. (1998). Feature Selection for knowledge Discovery and Data Mining Available at URL:

http://www.databaseheadquarters.com/bookstore/management2/079238198XAMUS141630.shtml (accessed on May 5, 2011).

34. Nitesh V., Kevin W., Lawrence O., W. Philip. (2007). SMOTE: Synthetic Minority Over-sampling Technique. Journal of Artificial Intelligence Research, volume 16, 321 -- 357

ANNEX I

THE IDS GENERIC CASE INVESTIGATION FORM

Reporting Health Facility:				Reporting Woreda/Zone				
IDS Case-based Reporting Format **From Health Facility/Health Worker to Woreda/Zone Health office/department**								
Cholera	Dracunculiasis	Neonatal Tetanus	Measles	Meningitis	Plague	Viral Hemorrhagic Fever	Yellow Fever	Others/specify
Date of form received at the national level: / / (Day/Month/Year)								
Name of Patient:								
Date of Birth (DOB): / / (Day/Month/Year)				**Age (If DOB unknown):**		Year	Month (if <12)	Day (NNT only)
Sex:		M=Male F=Female						
Patient's Address:		Urban		Rural				
Kebele:					House number:			
Woreda:		Zone:			Region:			
Locating Information:	If applicable or If the patient is neonate or child, please write full name of mother and father of the patient							
Date Seen at Health Facility: / /	**Date Health Facility notified Woreda/Zone:** / /		**Date of Onset:** / /					
Number of vaccine doses received:	9=unknown For cases of Measles, NT (TT in mother), Yellow Fever, and Meningitis (For Measles, TT, YF- by card & for Meningitis, by history)							
Date of last vaccination:	/ / (Measles, Neonatal Tetanus (TT in mother), Yellow Fever, and Meningitis only)							
Blank variable #1 of the case:								
Blank variable #2 of the case:								
In/Out Patient	1=Inpatient	2=outpatient						
Outcome	1=Alive	2=Dead	3=Unknown					
Final Classification of case	1=Confirmed	2=probable	3-Discarded	4=Suspect				
Person Completing the form; Name:	**Signature:**							
Date form sent to Woreda/Zone: / / (Day/Month/Year)								

If Lab Specimen Collected

For Health Facility: If lab specimen is collected, complete the following information and send a copy of this form to the lab with the specimen.											
Date of specimen collection: _____/_____/_____											
Type of specimen:		Stool		Blood		CSF		Other/specify			
Date specimen sent to lab: _____/_____/_____											
ID Number:											
For the Lab: Complete this section and return the form to Woreda/Zone health facility team or clinician											
Date lab received specimen: _____/_____/_____											
Specimen Condition:		Adequate		Not adequate							
Disease/Condition:											
Type of Test:											
Result:			+ = Positive		- = Negative		P = pending				
Malaria	P. Faliciparum										
	P. Vivax										
Cholera (culture)											
Cholera direct exam; specify the method used: _____											
Meningitis: N meningitides	Culture										
	Latex										
	Gram stain										
Meningitis: S. pneumoniae	Culture										
	Latex										
	Gram stain										
Meningitis: H. influenzae	Culture										
	Latex										
	Gram stain										
Shigella Dysenteriae	Culture										
	Type		Type 1		Other types		No Shigella				
Typhoid Fever	Widal ("O" > 1:160)										
	Blood culture										
	Stool culture										
Epidemic Typhus: Serum test (OX19)											
Result:			+ = Positive		- = Negative		I= Indeter.		P=Pending		
Serologic Investigation (Viral Detection)	Yellow fever (IgM)										
	Measles (IgM)										
	Rubella (IgM)										
	RVF (IgM)										
	Ebola (IgM)										
Other lab test (specify)	Results:										
Date lab sent results to Woreda/Zone/health facility:					_____ /_____ /_____						
Name of lab sending results:											
Other pending results:											
Name of lab technician sending the results:						Signature:					
Date Woreda/Zone receive lab results: ____/____/_____					Woreda/Zone:						
Date lab results sent to health facility by Woreda/Zone: _____/_____/_____											
Date lab results received at the health facility: _____/_____/_____											

ANNEX II

MEASLES OUTBREAK LINE LIST

Health Facility _____ Woreda/ Zone _____

(Use this form if there are more than 5 cases of suspected measles cases seen in a health facility in a village in 1 Week. This indicates that an Outbreak may be occurring. Include the first 5 cases on this form)

Month _____ Year _____ Region _____

ID Number	Kebele/ Town	Name/sex	Date of Birth or age(Age in year (e.g. 4y) Age in Months if <12 months (e.g. 9m)	Date of onset of rash	No. of valid measles doses (0,1,2,99)	Lab .Specimen		Outcome 1=Alive 2= Dead 9=Unknownn	Comments
						Blood taken Yes/No	Results Pos/Neg		

ANNEX III

RUN INFORMATION FOR DECISION TREE CONSTRUCTED AS AN OUTPUT FOR EXPERIMENT 2

=== Run information ===

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

Relation: sample2-Weka.filters.unsupervised.attribute.Remove-R4,6-7,11,14-15,17-18,20,23-Weka.filters.unsupervised.attribute.Remove-R12-Weka.filters.unsupervised.attribute.NumericToNominal-R4,8,11-Weka.filters.supervised.instance.SMOTE-C2-K5-P700.0-S1-Weka.filters.supervised.instance.SMOTE-C2-K5-P47.0-S1-Weka.filters.unsupervised.attribute.Remove-R6,8,11-12

Instances: 23717

Attributes: 9

age

Sex

Season

year

MeaslesIgm

vx_status

ProvinceOfReSidence

Urbanrural

Outbreak

Test mode: split 70.0% train, remainder test

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

J48 pruned tree

MeaslesIgm = 1.0
| year = 2006: No (0.0)
| year = 2007
| | ProvinceOfReSidence = BENISHANGUL_G
| | | vx_status = unknown: Yes (9.0)
| | | vx_status = UnVaccinated: No (34.0)
| | | vx_status = Vaccinated: Yes (1.0)
| | ProvinceOfReSidence = ADDIS_ABABA: No (185.0/7.0)
| | ProvinceOfReSidence = OROMIA
| | | Urbanrural = R
| | | | Season = winter : No (333.0/11.0)
| | | | Season = Autumn
| | | | | age = <15
| | | | | | vx_status = unknown: No (29.0/2.0)
| | | | | | vx_status = UnVaccinated: Yes (1.0)
| | | | | | vx_status = Vaccinated: Yes (3.0)
| | | | | age = 10_14: Yes (11.0/3.0)
| | | | | age = <1: No (0.0)
| | | | | age = 1_4: No (131.0/9.0)
| | | | | age = 5_9
| | | | | | vx_status = unknown: No (71.0/3.0)
| | | | | | vx_status = UnVaccinated: No (67.0)
| | | | | | vx_status = Vaccinated: Yes (5.0)
| | | | Season = summer: No (14.0/1.0)
| | | | Season = spring : No (3.0)
| | | Urbanrural = U
| | | | Sex = M
| | | | | age = <15: No (114.0/4.0)
| | | | | age = 10_14: No (15.0/1.0)

| | | | | age = <1: Yes (4.0)

| | | | | age = 1_4: Yes (6.0/2.0)

| | | | | age = 5_9: Yes (1.0)

| | | | Sex = F

| | | | | Season = winter : Yes (7.0/3.0)

| | | | | Season = Autumn: Yes (13.0/1.0)

| | | | | Season = summer: No (11.0/1.0)

| | | | | Season = spring : Yes (0.0)

| | ProvinceOfReSidence = HARERI

| | | age = <15: No (0.0)

| | | age = 10_14: Yes (1.0)

| | | age = <1: Yes (1.0)

| | | age = 1_4: Yes (3.0/1.0)

| | | age = 5_9: No (29.0)

| | ProvinceOfReSidence = DIRE_DAWA: No (3.0/1.0)

| | ProvinceOfReSidence = TIGRAY: Yes (10.0/3.0)

| | ProvinceOfReSidence = AMHARA

| | | Season = winter

| | | | Urbanrural = R: No (80.0/5.0)

| | | | Urbanrural = U

| | | | | Sex = M: No (2.0)

| | | | | Sex = F: Yes (2.0)

| | | Season = Autumn: Yes (24.0/3.0)

| | | Season = summer

| | | | Sex = M: Yes (3.0)

| | | | Sex = F: No (24.0)

| | | Season = spring : No (0.0)

| | ProvinceOfReSidence = SNNPR

| | | Season = winter

| | | | age = <15: Yes (2.0)

| | | | age = 10_14: No (12.0/1.0)

| | | | age = <1: Yes (1.0)

| | | | age = 1_4: No (5.0/2.0)

| | | | age = 5_9: Yes (7.0)

| | | Season = Autumn

| | | | Urbanrural = R: No (22.0/5.0)

| | | | Urbanrural = U: Yes (6.0)

| | | Season = summer: No (79.0)

| | | Season = spring : No (0.0)

| | ProvinceOfReSidence = AFAR: No (0.0)

| | ProvinceOfReSidence = SOMALI: No (2.0)

| | ProvinceOfReSidence = GAMBELLA: No (0.0)

| year = 2008: No (2039.0/36.0)

| year = 2009: No (1643.0/30.0)

| year = 2010

| | Season = winter

| | | ProvinceOfReSidence = BENISHANGUL_G: No (0.0)

| | | ProvinceOfReSidence = ADDIS_ABABA

| | | | Sex = M: Yes (5.0)

| | | | Sex = F: No (24.0/1.0)

| | | ProvinceOfReSidence = OROMIA

| | | | age = <15

| | | | | vx_status = unknown: Yes (8.0)

| | | | | vx_status = UnVaccinated: No (51.0/3.0)

| | | | | vx_status = Vaccinated: No (37.0)

| | | | age = 10_14

| | | | | Urbanrural = R: Yes (10.43/1.0)

| | | | | Urbanrural = U: No (13.57/1.57)

| | | | age = <1: No (11.0)

| | | | age = 1_4

| | | | | vx_status = unknown: No (105.0/4.0)

| | | | | vx_status = UnVaccinated

| | | | | | Sex = M: No (97.0/2.0)

| | | | | | Sex = F: Yes (7.0/2.0)

| | | | | vx_status = Vaccinated: Yes (7.0/2.0)

| | | | age = 5_9: Yes (24.0/5.0)

| | | ProvinceOfReSidence = HARERI: No (2.0)

| | | ProvinceOfReSidence = DIRE_DAWA: No (0.0)

| | | ProvinceOfReSidence = TIGRAY

| | | | age = <15: No (0.0)

| | | | age = 10_14: No (12.0)

| | | | age = <1: Yes (1.0)

| | | | age = 1_4: Yes (3.0)

| | | | age = 5_9: No (0.0)

| | | ProvinceOfReSidence = AMHARA

| | | | age = <15: Yes (10.0/2.0)

| | | | age = 10_14: No (15.0/1.0)

| | | | age = <1: Yes (1.0)

| | | | age = 1_4: No (2.0)

| | | | age = 5_9: Yes (1.0)

| | | ProvinceOfReSidence = SNNPR

| | | | Urbanrural = R

| | | | | Sex = M: No (201.0/22.0)

| | | | | Sex = F

| | | | | | age = <15: No (12.0/1.0)

| | | | | | age = 10_14: No (36.0/1.0)

| | | | | | age = <1: No (0.0)

| | | | | | age = 1_4: Yes (10.0/2.0)

| | | | | | age = 5_9: Yes (9.0/1.0)

| | | | Urbanrural = U: Yes (16.0/1.0)

| | | ProvinceOfReSidence = AFAR: Yes (3.0)

| | | ProvinceOfReSidence = SOMALI: No (15.0)

| | | ProvinceOfReSidence = GAMBELLA: No (0.0)

| | Season = Autumn
| | | ProvinceOfReSidence = BENISHANGUL_G: No (1.0)
| | | ProvinceOfReSidence = ADDIS_ABABA: No (262.0/9.0)
| | | ProvinceOfReSidence = OROMIA: No (1758.0/145.0)
| | | ProvinceOfReSidence = HARERI
| | | | age = <15: No (0.0)
| | | | age = 10_14: No (0.0)
| | | | age = <1: Yes (1.0)
| | | | age = 1_4
| | | | | Sex = M: No (40.0)
| | | | | Sex = F: Yes (5.0/2.0)
| | | | age = 5_9: Yes (3.0)
| | | ProvinceOfReSidence = DIRE_DAWA: No (0.0)
| | | ProvinceOfReSidence = TIGRAY
| | | | Urbanrural = R: No (42.0)
| | | | Urbanrural = U: Yes (4.0)
| | | ProvinceOfReSidence = AMHARA
| | | | age = <15
| | | | | Urbanrural = R
| | | | | | Sex = M: No (165.0/14.0)
| | | | | | Sex = F: Yes (21.0/2.0)
| | | | | Urbanrural = U
| | | | | | Sex = M: Yes (7.0/3.0)
| | | | | | Sex = F: No (138.0/2.0)
| | | | age = 10_14: Yes (32.0/6.0)
| | | | age = <1
| | | | | Sex = M: Yes (2.0)
| | | | | Sex = F: No (16.0)
| | | | age = 1_4: No (377.0/21.0)
| | | | age = 5_9
| | | | | vx_status = unknown: Yes (10.0/3.0)

| | | | | vx_status = UnVaccinated

| | | | | | Urbanrural = R: No (109.0/12.0)

| | | | | | Urbanrural = U: Yes (3.0/1.0)

| | | | | vx_status = Vaccinated: Yes (5.0/1.0)

| | | ProvinceOfReSidence = SNNPR

| | | | age = <15

| | | | | Urbanrural = R: Yes (6.0/1.0)

| | | | | Urbanrural = U: No (11.0)

| | | | age = 10_14

| | | | | Urbanrural = R: Yes (7.0/1.0)

| | | | | Urbanrural = U: No (15.0/2.0)

| | | | age = <1: Yes (5.0/1.0)

| | | | age = 1_4

| | | | | vx_status = unknown: No (100.0/6.0)

| | | | | vx_status = UnVaccinated

| | | | | | Sex = M

| | | | | | | Urbanrural = R: No (57.0/8.0)

| | | | | | | Urbanrural = U: Yes (2.0)

| | | | | | Sex = F: Yes (7.0/1.0)

| | | | | vx_status = Vaccinated

| | | | | | Urbanrural = R: Yes (10.0/3.0)

| | | | | | Urbanrural = U: No (49.0)

| | | | age = 5_9

| | | | | Urbanrural = R: Yes (22.0/4.0)

| | | | | Urbanrural = U

| | | | | | vx_status = unknown: Yes (1.0)

| | | | | | vx_status = UnVaccinated: Yes (3.0/1.0)

| | | | | | vx_status = Vaccinated: No (15.0/1.0)

| | | ProvinceOfReSidence = AFAR: No (11.0)

| | | ProvinceOfReSidence = SOMALI

| | | | Urbanrural = R: No (47.0)

| | | | Urbanrural = U: Yes (4.0)

| | | ProvinceOfReSidence = GAMBELLA: No (1.0)

| | Season = summer

| | | ProvinceOfReSidence = BENISHANGUL_G: No (1.0)

| | | ProvinceOfReSidence = ADDIS_ABABA

| | | | vx_status = unknown: Yes (6.0)

| | | | vx_status = UnVaccinated

| | | | | age = <15: No (18.0/4.0)

| | | | | age = 10_14: Yes (1.0)

| | | | | age = <1: Yes (2.0)

| | | | | age = 1_4: No (13.0/2.0)

| | | | | age = 5_9: Yes (1.0)

| | | | vx_status = Vaccinated: Yes (12.0)

| | | ProvinceOfReSidence = OROMIA

| | | | vx_status = unknown: No (299.0/8.0)

| | | | vx_status = UnVaccinated

| | | | | age = <15: No (23.0/2.0)

| | | | | age = 10_14

| | | | | | Sex = M: Yes (3.0)

| | | | | | Sex = F: No (24.0)

| | | | | age = <1: Yes (4.0/1.0)

| | | | | age = 1_4: No (56.0/2.0)

| | | | | age = 5_9

| | | | | | Sex = M: No (3.0/1.0)

| | | | | | Sex = F: Yes (6.0)

| | | | vx_status = Vaccinated

| | | | | Sex = M

| | | | | | age = <15: Yes (0.0)

| | | | | | age = 10_14: Yes (0.0)

| | | | | | age = <1: Yes (0.0)

| | | | | | age = 1_4: Yes (3.0)

| | | | | | age = 5_9: No (3.0/1.0)

| | | | | Sex = F: No (45.0/1.0)

| | | ProvinceOfReSidence = HARERI: No (1.0)

| | | ProvinceOfReSidence = DIRE_DAWA: Yes (3.0)

| | | ProvinceOfReSidence = TIGRAY: No (45.0)

| | | ProvinceOfReSidence = AMHARA

| | | | vx_status = unknown

| | | | | age = <15: Yes (5.0)

| | | | | age = 10_14: Yes (1.0)

| | | | | age = <1: No (1.0)

| | | | | age = 1_4: Yes (0.0)

| | | | | age = 5_9: No (2.0)

| | | | vx_status = UnVaccinated: Yes (12.0/2.0)

| | | | vx_status = Vaccinated: No (13.0/2.0)

| | | ProvinceOfReSidence = SNNPR

| | | | vx_status = unknown: Yes (2.0)

| | | | vx_status = UnVaccinated: Yes (4.0)

| | | | vx_status = Vaccinated: No (25.0)

| | | ProvinceOfReSidence = AFAR: No (0.0)

| | | ProvinceOfReSidence = SOMALI: No (2.0)

| | | ProvinceOfReSidence = GAMBELLA: No (0.0)

| | Season = spring

| | | Urbanrural = R

| | | | age = <15

| | | | | ProvinceOfReSidence = BENISHANGUL_G: Yes (0.0)

| | | | | ProvinceOfReSidence = ADDIS_ABABA: Yes (0.0)

| | | | | ProvinceOfReSidence = OROMIA

| | | | | | Sex = M: No (23.39/4.39)

| | | | | | Sex = F: Yes (6.0)

| | | | | ProvinceOfReSidence = HARERI: Yes (0.0)

| | | | | ProvinceOfReSidence = DIRE_DAWA: Yes (0.0)

| | | | | ProvinceOfReSidence = TIGRAY: Yes (8.0/2.0)

| | | | | ProvinceOfReSidence = AMHARA: Yes (22.0/1.0)

| | | | | ProvinceOfReSidence = SNNPR: Yes (2.0)

| | | | | ProvinceOfReSidence = AFAR: Yes (0.0)

| | | | | ProvinceOfReSidence = SOMALI: Yes (0.0)

| | | | | ProvinceOfReSidence = GAMBELLA: Yes (0.0)

| | | | age = 10_14: Yes (17.0/1.0)

| | | | age = <1: Yes (3.0)

| | | | age = 1_4: Yes (60.79/3.0)

| | | | age = 5_9: Yes (29.0/1.0)

| | | Urbanrural = U: Yes (262.82/1.0)

| year = 2011

| | vx_status = unknown: Yes (38.0/1.0)

| | vx_status = UnVaccinated: Yes (31.0)

| | vx_status = Vaccinated

| | | age = <15: Yes (1.0)

| | | age = 10_14: No (0.0)

| | | age = <1: No (0.0)

| | | age = 1_4: Yes (5.0)

| | | age = 5_9

| | | | ProvinceOfReSidence = BENISHANGUL_G: No (0.0)

| | | | ProvinceOfReSidence = ADDIS_ABABA: No (0.0)

| | | | ProvinceOfReSidence = OROMIA: Yes (5.0/2.0)

| | | | ProvinceOfReSidence = HARERI: No (0.0)

| | | | ProvinceOfReSidence = DIRE_DAWA: No (0.0)

| | | | ProvinceOfReSidence = TIGRAY: No (0.0)

| | | | ProvinceOfReSidence = AMHARA: Yes (1.0)

| | | | ProvinceOfReSidence = SNNPR: No (13.0/1.0)

| | | | ProvinceOfReSidence = AFAR: No (0.0)

| | | | ProvinceOfReSidence = SOMALI: No (0.0)

| | | | ProvinceOfReSidence = GAMBELLA: No (0.0)

MeaslesIgm = unknown: Yes (7872.0)

MeaslesIgm = 4.0: Yes (67.0)

MeaslesIgm = 2.0: Yes (0.0)

MeaslesIgm = 3.0: Yes (0.0)

Number of Leaves : 209

Size of the tree : 282

Time taken to build model: 0.09 seconds

=== Evaluation on test split ===

=== Summary ===

Correctly Classified Instances	5284	97.061 %
Incorrectly Classified Instances	160	2.939 %
Kappa statistic	0.9412	
Mean absolute error	0.0541	
Root mean squared error	0.1671	
Relative absolute error	10.8181 %	
Root relative squared error	33.406 %	
Total Number of Instances	5444	
Ignored Class Unknown Instances	1671	

=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
	0.948	0.007	0.992	0.948	0.97	0.914	Yes
	0.993	0.052	0.951	0.993	0.972	0.971	No
Weighted Avg.	0.971	0.03	0.972	0.971	0.971	0.943	

=== Confusion Matrix ===

```
   a    b   <-- classified as
2556  140 |    a = Yes
  20 2728 |    b = No
```