Addis Ababa University

School of Public Health and

School of Information Science

Master of Science in Health Informatics



Analyzing the Outbreak Surveillance and Response System in Ethiopia using Data Mining Techniques

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

By:

Yimer Mohammed Hussien

Addis Ababa,

Ethiopia

November, 2012

Addis Ababa University

School of Public Health and

School of Information Science

Master of Science in Health Informatics

Analyzing the Outbreak Surveillance and Response System in Ethiopia using Data Mining Techniques

By

Yimer Mohammed

Name and Signature of Members of the Examining Board

Chairman, Examining Board_____

Ato Ermias Abebe, Advisor_____

Dr. Adamu Addisie, Advisor_____

External Examiner1_____

External Examiner2_____

Dedication

This thesis is dedicated to my adored and beloved Mam and Dad!!

Declaration

This thesis is my original research, and has not been presented for a degree in any other university.

Yimer Mohammed

June, 2012

This thesis has been submitted for examination with our approval as university advisors

---

Ato. Ermias Abebe

Dr. Adamu Addisie

November , 2012

Acknowledgements

My special gratitude goes to my advisors Ato Ermias Abebe and Dr. Adamu Addisie. They are so committed and accountable to push me with constructive assistance, advice and comment to accomplish this research. I would also like to express my deepest gratefulness towards them for their convincing and all rounded help in all aspects of my thesis.

My special thanks also goes towards Ato Abebe and Ato Fikre at PHEM sector for their plausible transparency for the process of business and data understanding in the research.
I would also like to thank Dr. Million at PHEM sector for his full tolerance to the results evaluation and constructive discussion with the findings.

I would also like to thank all Health Informatics department students for their peer evaluation and encouraging comments.

The Last, but not the least I would also like to thank my cousin Alebachew Worku and my parents for their financial and all rounded support.

Finally, without the collaborative efforts of the persons listed above and the others, whose names are not listed here due to many in number for different and varied supports, the process of this thesis were difficult to do and complete with the allocated time and budget. Therefore I am equally thanks for all of them.

Table of Contents

Acknowledgements I

Acronyms V

List of figures VI

List of Tables VII

Abstract VIII

Chapter one 1

1.1. Background /Introduction 1

1.2. *Statements of the problem* 5

1.3. Objectives of the study 8

1.3.1. General objectives 8

1.3.2. Specific objectives 8

1.4. Methodology 9

1.4.1. Study Area/Setting and Period 9

1.4.2. Study Population 9

1.4.3. Study Design 9

1.4.4. Data collection techniques 10

1.4.5. Data analysis techniques and tools 11

1.5. Ethical Considerations 11

1.6. Dissemination of results 11

1.7. Scope and Limitation of the study 12

1.8. Organization of the Thesis 13

1.9. Beneficiaries of the research 14

Chapter Two 15

Literature Review 15

2.1. Introduction 15

2.2. General overview of DM and KD process model 17

2.2.1. KDD overviews 17

2.2.2. Model Development Processes and Techniques 18
2.2.2.1. KD Process model 18
2.2.2.2. Hybrid KD Process Models 19
2.3. Data mining 22
2.3.1. Data mining functionalities 24
2.2.3. Core Ideas in Data mining process 26
2.3. Supervised and Unsupervised Learning in data mining 28
2.4. Classification and Prediction 29
2.4.1. Comparison between Classification and Prediction Methods 30
2.5. Methods of Data preparation for data mining algorithms 30
2.6. Association Rules Mining 31
2.7. Data Clustering Applications 33
2.8. Application of data mining in Health Care 34
2.9. Limitations of Data Mining Applications 38
2.10. Related works 40
2.11. Methods and Techniques used for outbreak surveillance data mining applications 42
2.11.1. Simple K-Means clustering algorithm 42
2.11.2. Association rule mining with Apriori algorithms 44
2.11.2.1. Apriori Algorithm 44
2.12. Classification with Naïve Bayes and Decision Tree classifier 46
2.12.1. Naïve Bayes Classifier 46
2.12.1. Naïve Bayes algorithm 47
2.12.2. Decision Trees and Decision Rules 48
2.12.2.1. Minimizing Model Over-fitting on decision tree 50
2.12.2.2. Classification Metrics for predictive model development 51
Chapter Three 54
3. Public Health Emergency Management 54
3.1. Public health surveillance and response system 54
3.2. Overview of PHEM sector in Ethiopia 56

3.3. Ethiopian Public Health surveillance and Response System 57

Chapter Four 60

4.1. Business understanding and data preprocessing 60

4.1.1. Business understanding 60

4.1.2. Data Understanding and Preprocessing 62

4.1.3. Data Preparation 63

4.1.4. Data Quality Measures 65

4.1.5. Data transformation 66

4.1.6. Data integration, data reduction and feature selection process 67

4.2. Data quality assessment based on timely data reporting activities 69

Chapter Five 72

5.1. Experimental Process for MD 72

5.1.2. Experimentation with Apriori Association-Rule Mining algorithm 73

5.1.2.1. Association rules mining by Apriori Algorithm for PHEM datasets 75

5.1.2.2. Association rules mining by Apriori Algorithm for IDSR datasets 82

5.1.2.3. Association rules mining by Apriori Algorithm for combined Dataset (2004-2012g.C). 89

5.1.2.4. Summary of association rules mining results from Apriori Algorithm 97

5.1.3. Model development for disease clustering with a simple-KMeans Algorithm 100

5.1.4. Prediction MD for Epidemic Typhus disease 108

5.1.4.1. Model preparation For Prediction 109

5.2. General Summary of experimentation results: 119

CHAPTER Six 123

6, CONCLUSIONS AND RECOMMENDATIONS 123

6.1. Conclusions 123

6.2. Recommendation 125

Reference 128

Appendix 1: 133

Appendix 2: 135

Appendix 3: 138

Acronyms

ARFF:	Attribute-Relation File Format
BPR:	Business Process Reengineering
BRHP:	Butajira Rural Health Project
CRISP_DM:	Cross Industry Standard Process for Data Mining
CSV:	Comma Separated Value
DM & KD:	Data Mining and Knowledge Discovery
DSS:	Decision Support System
ED:	Electronic Data
EFDR:	Ethiopian Federal Democratic Republic
EHNRI:	Ethiopian Health and Nutritional Research Institute
ESR:	Environmental Science & Research
FMoH:	Federal Ministry of Health
GPA:	Grade Point Average
HMO:	Health Management Organization
IDSR:	Integrated Disease Surveillance and Response
IHR:	International Health Regulations
KDD:	Knowledge discovery in databases
MD:	Model Development
MIS:	Management information system
MoH:	Ministry of Health
NGO:	Nongovernmental organization
NGOs:	None Governmental Organizations
PHEM:	Public Health Emergency Management
RTA:	Royal Thai Army
RT-PCR:	reverse transcription polymerase chain reaction
SARS:	Severe Acute Respiratory Syndrome
TDIDT:	Top-Down Induction of Decision Trees
Weka:	Waikato Environment for Knowledge Analysis
WHO:	World Health Organization
AAU:	Addis Ababa University

HSDP: Health Sector Development Programs

PHC: Primary Health Care

List of figures

Figure 1: The CRISP_DM process model----------19

Figure2: The six-step Hybrid KDP Model----------20

Figure3: Simple K-Means clustering algorithm implementation steps----------43

Figure4: A simple decision tree with the tests on attributes X and Y----------49

List of Tables

Table1. Confusion matrix for two classes----------52

Table2. Epidemic prone reportable disease lists----------61

Table 3: The percentage of reporting efficiency in monthly bases of the PHEM dataset----------70

Table 4: Tabular description of Apriori Algorithm to Association rules generation for PHEM datasets.-----75

Table5: The Run information to the five experimentation of Apriori association rule mining algorithm for IDSR datasets. ----------82

Table6: The Run information to the five experimentation of Apriori association rule mining algorithm for IDSR datasets. ----------89

Table7: Comparison among the three datasets for Apriori association rules mining result---------97

Table8: Selected rules and their respected support and confidence level ----------98

Table9: selecting rule interestingness from the average support and confidence of the three datasets----------99

Table10: Clustering Experimentation analysis for the comparison of within cluster sum of squared errors and Number of iterations performed for each experimentation. ----------102

Table11: cluster result of the three experimentations for cluster instance distribution----------103

Table12: clustering with simple K-means algorithm run information results for better visualization purpose for experiment3 ----------104

Table13: parameter description for attribute selecter or feature generator window in Weka---110

Table14: Gain ratio for best attributes selection process with their ranking lists----------110

Table15: Measurement evaluation for the two decision tree classifiers for comparison --------112

Abstract

The aim of this research work was to show the applicability of data mining techniques for the development of descriptive and predictive model to disease outbreak surveillance datasets in Ethiopia. To do that the three data mining applications such as classification, clustering and association rules mining were undertaken to explore the important applications to the datasets of the PHEM sectors from different perspectives.

A total of 18600 records were collected and assessed from the data store of the surveillance system from the year 2004-2012G.C. After the preprocessing phase of knowledge discovery in databases of data mining application a total of 8796 records were prepared for data mining algorithms. From the total records filtered and prepared for model preparation 4703 were from the IDSR system dataset and the remaining 4093 records were taken from that of the PHEM dataset from the year 2004-2008G.C. and 2009-2012G.C. respectively.

The researcher analyzed two classification algorithms for the prediction of Epidemic typhus disease cases with decision tree J48 classifiers and Naïve Bayes classifiers. Finally the more performing algorithm has been taken for model development. From the results of the experiments done decision tree algorithm had a better performance to classify the disease cases in place and time setting. The accuracy rate of correctly classifying the Epidemic Typhus disease cases by the use of decision tree J48 algorithm was 87.44% whereas with Naïve Bayes classifier was 83.70%. The sensitivity and specificity test was also done for the two classifiers.

The researcher also attempted to analyze the application of association rule mining to find some sort of correlation or patters among disease cases of the surveillance data. The attributes were selected only from the disease cases for the occurrence and nonoccurrence, which were collected in time and place bases. Here, Apriori association rule mining algorithm was run to find interesting patterns among the occurrence and co-occurrence of disease cases which were correlated together. The researcher used 20% for the minimum support and 90% for minimum confidence threshold before the application of the mining algorithm.

The researcher took the combined (integrated) datasets for cluster analysis with the total numbers of 8796 records with 9 attributes. Simple K-Means clustering algorithm was used for the combined datasets since; the algorithm showed the grouping of disease cases with respect to time and place.

In general data mining techniques were important and applicable in the classification, clustering and association rules model development for emerging and reemerging disease cases. But the data

has to have good quality with the inclusion of important attributes of variables for better prediction and description model development The results of the research, apart from its education purpose, were also used for the planning, preparedness, decision making, and disease control and prevention activities to the domain experts.

Chapter one

1.1. Background /Introduction

In the past, when sanitary conditions were poor and diseases were little understood, epidemics occurred periodically and killed thousands of people. One of the largest epidemics ever recorded was the outbreak of bubonic plague that raged throughout Europe, Africa, and Asia from 1347 to 1350. This epidemic, known as the Black Death in Europe, is estimated to have killed one-third of the European population. An outbreak of influenza in 1918 killed more than 20 million people around the world. Such global epidemics are commonly called pandemics **[1]**.

During the past 70 years, there has been a dramatic fall in the incidence of infectious diseases, particularly in developed countries **[2]**. This was due to several factors including: immunization, Anti-microbial chemotherapy, improved nutrition, better sanitation and housing. In less developed countries, however, especially in the tropics, infectious diseases continue to be one of the commonest causes of death, particularly in children **[2]**. In addition to that, Ethiopia, as part of the developing world, has two big health problems. Most of the time infectious diseases (communicable diseases) account about 80% of the health problem in the country which can be prevented by simple sanitary measures and nutritional programs.

As mentioned in **[3],** communicable diseases are considered as major causes of morbidity and mortality, as well as disability in Ethiopia. The high prevalence of communicable diseases in the country was linked to the poorly developed socio-economic and environmental factors that have been inherent for centuries. For this reason, environmental health activities in Ethiopia were maintained and used as a means to control diseases and promote health during periods of “Basic Health Service Provision during 1950-1960’s”, and PHC during the 1970-90’s, and currently in the HSDP.

As mentioned above it is also shown that the majority of the population of Ethiopia, urban and rural alike does not have access to safe and adequate sanitation facilities **[4]**. As well a study carried out in **[4]**, indicates that three-fourth of the health problems of children in Ethiopia are communicable diseases attributed to unsafe and inadequate provision of water and unhygienic waste management, particularly sewage. Additionally, other studies also

revealed that only 11.5 % of the population had access to sanitation facilities while 88.5 % were without proper sanitation facilities. The existing capacity in the wastewater sector in Ethiopia is currently relatively low, in particular in regard to alternative concepts and technologies aiming at more decentralized and environmentally friendly, energy and resource saving solutions.

In 1996, as part of the response to the growing public health problem with communicable diseases, Ethiopia introduced an IDSR strategy focusing on 17 priority diseases. Ethiopia adopted the WHO's infectious disease surveillance and response strategy in 1998, and in October 1999, MoH of Ethiopia and its development partners assessed the country's surveillance system and used the results to adapt a five-year national plan **[5].**

Since Infectious and communicable diseases account for about 60-80% of the health problems in Ethiopia, that Ethiopia has developed a five-year strategic plan and started implementing the IDSR consisting of 23 priority diseases that are epidemic-prone, have public health importance and are targeted for eradication and elimination **[6]**. According to the plan, "Later three diseases were reduced and 20 of them were only followed up in the sector". The systematic evaluation of the first five years of implementation indicated that the country has successfully expanded its surveillance system up to the health centre and hospital level, improving the early detection, reporting, epidemic preparedness and response activities**.**

As stated in **[7], t**he morbidity and mortality associated with infectious disease outbreaks, which are directly or indirectly linked to ecologic or climate events and trends, pose a growing problem for global public health. Many factors are associated with this growth: disease vector-habitat expansion due to environmental degradation and climate variability; changes in animal and human population dynamics that increase the risk of human exposure to infective pathogens; and the insufficiency of public health infrastructure in resource-limited settings to support and sustain routine infectious disease surveillance, prevention and control activities**.**

Surveillance, risk assessment and outbreak response capacity is a prerequisite for effective management of emerging disease outbreaks and other acute public health events. Effective national surveillance systems generate reliable information for timely risk assessment that informs rapid public health actions **[8].**

According to **[9]**, sub-Saharan Africa will not reach the malnutrition and child mortality targets, as well as some areas in communicable diseases at current pace of implementation. It was also pointed out; there was a need to accelerate improvements in the country's healthcare system if Ethiopia expects to reach its health targets and the MDGs for health. The FMoH introduced the PHEM core process giving the responsibilities of emergencies and disease outbreak condition with the implementation of BPR. Its aim was to improve performance through the introduction of business-process measures and indicators in the government system for communicable disease control and prevention. This might help improve technical and allocative efficiency and thereby increase standards and quality.

As depicted in **[7],** the success of predictive surveillance will depend on the recognition that it is founded on a meta-system of different surveillance activities, linked by communications, coordination and collaboration between multiple, diverse disciplines, and producing advisories and alerts, which have progressively more certainty over the course of an outbreak's emergence**.**

Epidemiological surveillance is a major public health strategy in prevention and control of disease **[7]**. Additionally, Surveillance not only gives us accurate data on epidemiology and burden of disease but also guides us to monitor and control diseases under surveillance. Therefore, surveillance data needs to be capable of identifying clusters; tracking trends; assessing the effectiveness of interventions; and providing timely results. Development of surveillance systems to rapidly identify clusters and outbreaks, and patients at increased risk of acquiring infection will ensure that control measures are best directed to prevent further cases occurring.

According to **[10],** once the cause of the outbreak has been identified, investigators should work to implement longer term control measures to end the current outbreak and prevent future outbreaks. Consequently, well-planned communication can enhance communicable disease outbreak preparation and can hasten containment of an outbreak as well as help to mitigate an outbreak's social and economic impact. At last the need of communicable disease outbreak surveillance and response system objectives are usually take care of patients, prevent further cases, end the outbreak quickly and prevent its recurrence and if possible to eradicate totally.

Advancements in clean water and refrigeration and the development of safe, effective vaccines have greatly decreased such threats; however, common diseases still cause outbreaks and new communicable diseases are emerging **[11].**

As discussed in **[12],** there are also several important factors which have to be addresses since they are influencing the incidence of infectious or communicable diseases like: climate such as temperature, rainfall, humidity, latitude, precipitation etc; socio-demographic influences such as human migration and transportation etc; and drug resistance and nutrition; way of transmission like water, air, food etc; as well as environmental influences such as deforestation; agricultural development; water projects; and urbanization. Therefore the public health functionalities and responsibilities will be challenged and constrained if the problems are not well managed effectively and efficiently in time at each level of administrative functional units.

1.2. *Statements of the problem*

As mentioned in the introduction part of this work, most disease outbreaks are infectious in nature. If they are not early detected and responded with appropriate technology, public health intervention and control, good performing surveillance systems, they can kill many people within a short period of time. More importantly, communicable disease prevention and control is the cornerstone of public health.

One of the most exciting and challenging tasks facing an epidemiologist therefore is working in a public health department to investigate a disease outbreak **[13]**. Usually, the cause and source of the outbreak are unknown. Sometimes large numbers of people are affected. Often, the people in the community are concerned because they fear more people, including themselves, may be stricken unless the cause is found soon. There may be hostilities and defensiveness if an individual, product, or company has been accused of being the cause. Consequently, this pressure-packed situation affects the epidemiologist wherever they come from. Therefore, the epidemiologist must remain calm, professional, and scientifically objective.

For better implementation and improvements of the country's healthcare system, Ethiopia is expected to reach its health targets and the MDGs for health with better surveillance and response system for communicable disease occurrences. Therefore the country has implemented BPR for health and analyzed the critical area of health in the country to improve health care system with skilled human resource. But the results of the new surveillance and response system for disease outbreak situation was not yet analyzed and seen with appropriate techniques and tools.

The PHEM sector under study stores data about every communicable disease outbreaks cases using MS-excel data sheets while it was reported from all respected sites of the country routinely. However, as mentioned in **[14],** data collection activities are the most important step in disease outbreak situation for better data analysis and model development for decision making activities. It was also added that collection of high quality outbreak data nationwide, proper data reporting mechanism, and comprehensive documentation of all recognized outbreaks are essential for any disease surveillance system. They help: to facilitate recognition of relationships or association between disease events occurring in different areas of the country; to convince health professionals and the public of the need for preventive

measures, and to evaluate and improve prevention strategies and if possible for eradication purposes.

The PHEM sector data collection activities were performed in daily and weekly bases for planning and implementation. As mentioned in **[14]**, appropriate data collection process facilitates systematic recording of all the required details of the disease outbreak, thereby providing information for decision making activities on whether and how to proceed further. The information contributes for the recognition of linked multi-district outbreaks and early control measures. Additionally, completed outbreak report forms are used in the production of local and national statistics on outbreak occurrence, including causal agents, modes of transmission, risk factors, affected population, time of onset, more affected areas and soon.

The PHEM sector used the surveillance system outbreak data reports for the process of investigating, controlling, preventing and planning for early warning and preparedness activities for communicable disease outbreak cases throughout the country. Once the data was collected, one can begin to characterize an outbreak by time, place, and person **[13]**. In fact, the PHEM sectors winded up performing this step several times during the course of an outbreak. But, characterizing an outbreak by these variables is descriptive in epidemiological perspectives, because one can describe what has occurred in the population under study. Therefore; the problem in the sector was lack of having analytical machine learning techniques and tools to the surveillance and response system in the country to develop a better descriptive as well as predictive model for efficient and effective decision making processes.

Even though, the process of compiling, describing and analyzing data by time, person and place are desirable for several reasons like: the investigator becomes intimately familiar with the data and with the extent of the public health problem being investigated, providing a detailed description of the health of a population with easy communication, and finally identifying the populations that are at greatest risk of acquiring a particular disease, the sector also lacked appropriate techniques, tools and skilled professionals to mitigate the impact and possibly stop the occurrence and reoccurrence of the next disease outbreaks.

Due to the problem stated above and the discussion with the domain experts, the data collection, analysis, interpretation and dissemination techniques were also done with simple statistical tools like MS-Excel, epi-info, and SPSS. The results were also described in

percentage, mean, median, mode and some other statistic methods like ANOVA, normal distribution, and standard deviation with simple visualization techniques like tables, graphs, charts, normal curve and soon. Therefore, decision was made according to full of traditional statistical techniques for descriptive data analysis purpose with the collected data**.**
Another problem was that large amount of data collected through a long period of time (i.e. many years of data) were not well organized and also were not analyzed using analytical machine learning tools like data mining applications to improve the decision making process.

The primary goal of this research work was therefore, to analyze the Outbreak surveillance and response system of Ethiopia using data mining application for better decision making activities. The datasets in the PHEM core process were so huge as mentioned before, and the simple statistical tools would not support to provide significant information and knowledge from these large collections of data as done by the sector professional for the purpose of early warning, planning, preparedness and decision making activities. Furthermore, the researcher planned to perform an effective analysis of the dataset with data mining application to model descriptive, predictive and association rule mining for effective and efficient surveillance and response system for better implementation and decision making process.

Research questions

1. Are the qualities of data in the PHEM sector data store appropriate for data mining application?
2. Are there any associations among emerging and reemerging disease outbreak cases in the PHEM dataset?
3. Could it be possible to develop a predictive model for disease outbreak cases from the existing PHEM surveillance and response dataset?

1.3. Objectives of the study

1.3.1. General objectives

To analyze the disease outbreak surveillance and response system of Ethiopia to point out the possibilities of applying and showing better results with data mining applications for predictive and descriptive model preparation for emerging and re-emerging disease outbreak cases, and show the co-occurrence of disease cases in the country.

1.3.2. Specific objectives

The specific objectives of this research work are to:-

- ✓ Carry out a thorough review of a sort of related literature in data mining techniques and communicable disease outbreak surveillance and response system for better analysis.
- ✓ See if data mining techniques can be applied for better results to the surveillance and response system or not.
- ✓ Prepare the existing datasets for better use of data mining algorithms and methods
- ✓ Show the most prevailing disease cases to different locations of the country.
- ✓ Evaluate the functionality and the performance of the model under study.
- ✓ Develop a new model and make recommendation.

1.4. Methodology

1.4.1. Study Area/Setting and Period

The study was conducted at the EHNRI specifically in PHEM Sector from February, 2012 to March, 2012 for data collection activities.

1.4.2. Study Population

All disease outbreak cases which were reported weekly as an emerging and re-emerging disease outbreak to the PHEM sector have been considered in the study. The occurrences and association (correlation) of cases which do have a great contribution in public health treats can also be investigated with data mining applications. The target populations were all the people of the country who are affected and infected by a certain communicable disease outbreaks. Out of the 20 epidemic-prone priority diseases in the surveillance system, only the weekly reportable outbreak cases such as Malaria, Meningitis, Relapsing Fever, Typhoid Fever and Epidemic Typhus have been incorporated and analyzed in the study. Samples has been taken from the existing outbreak reports of the country surveillance and response system datasets between 2004-2012G.C. The *regional, zonal, Woreda, Health Facilities* and other *Governmental and nongovernmental organizations* reports of all disease outbreak cases (primary or index cases and secondary cases) which are resulted from communicable disease outbreaks to the population has been added and analyzed by data mining applications.

1.4.3. Study Design

The KDD process model specially the hybrid KDD process model would be applied on the data which were obtained from different disease outbreak cases reports of the country to the PHEM Sector. Data mining tools, techniques and algorithm were applied to discover unseen patters from the surveillance datasets. That's why, these techniques and algorithms are an efficient way of showing clusters, categories and associations between the different regions and times of the country with respect to diseases occurrence. Some rules like association rules mining would be used to determine the patterns of disease outbreak cases of the country

to show the correlation between different types of outbreaks within a certain time and place. Apriori association rule mining algorithm would be used for the co-occurrence of two or more diseases cases for pattern discovery. Classification algorithms were also used to predict the future occurrence of a certain communicable disease outbreaks in time and place. For this purpose the researcher used decision tree and Naïve Bayes classifier for Epidemic Typhus disease prediction. Disease segmentation would be done by the use of Simple K-Means clustering algorithm data mining techniques to show the variation disease prevalence in time and place setting..

1.4.4. Data collection techniques

The process of data collection was carried out much more quickly and at much less expense from the secondary data collected in the PHEM sector from weekly reportable routine surveillance datasets. The outbreak reports of different regions and different time periods in the country had to be collected and stored in Excel workbooks by the Sector. The use of secondary data that were already available in the sector would be involved in the study, this data collection strategy and techniques were appropriate for description and a little bit for prediction activities for the occurrences of outbreak cases and to show the patterns of association among them.

In the national PHEM surveillance system the data collection tools were that of Microsoft excel, Microsoft access, Microsoft word processor, internet (web browser for email communication) and telephone conversation, letter writing. As well the data collection techniques were divided in to two categories. The first one was daily reportable disease data collection mechanisms such as line listing, daily epidemic report, and case based reporting system. The second one was weekly disease outbreak reporting (Routine surveillance data collection) mechanisms. These reporting divisions were formulated due to the fatality, morbidity and natural history of the disease outbreak cases and prevention or control mechanisms set for action.

1.4.5. Data analysis techniques and tools

Before the data analysis techniques, data had to be collected as mentioned before. The data that were already collected were transformed in the format that data mining tools can process or run the datasets after some preprocessing activities. The data mining tools such as WEKA 3.6 or WEKA 3.7 were used for the analysis and discovery of the hidden patterns in the dataset of disease outbreak reports. Since the surveillance data was stored in Microsoft excel worksheet, it had to be converted and transformed to comma Separated Value (CSV) text format or Attribute-Relation File Format (ARFF) file format.

Finally, Multidimensional and multilevel data could be analyzed easily and accurately with the application of data mining techniques and algorithms to discover the hidden knowledge in the dataset. It was because of data mining algorithms which were efficient techniques to extract the hidden knowledge from the huge datasets for decision making activities.

1.5. Ethical Considerations

AAU, School of Public Health provided ethical clearance and approval for the research. To conduct the study, Permission has also been obtained from PHEM Sector located in EHNRI. Informed consents of participants were obtained for the study before conducting the interview. The general as well as the specific objectives of the study was explained to the participants or domain experts briefly. The respondents were informed about that they have the right to involve into or leave from the study while the studying period.

1.6. Dissemination of results

Findings were presented in graphical, tabular and textual formats. This study was used to identify the affected or vulnerable communities that were exposed to a certain disease outbreak cases and classify them in times and places. The final findings were presented to AAU school of Informatics, AAU School of Public Health, EHNRI and FMoH.

1.7. Scope and Limitation of the study

The scope of this research is to analyze the current surveillance and response system of Ethiopia with the potential applicability of data mining techniques in supporting the outbreak disease surveillance and response system of the country.

The research was done only for five communicable disease outbreak cases. It was because of; some of the diseases in the system did have a national concern data for some disease outbreaks. After the application of data mining techniques of this research work could comparatively be considered as significant in assessing the potential applicability of data mining technology to the whole disease outbreak cases. The use of the current descriptive study undertaken in the sector was strictly limited to analyze the best use of surveillance datasets with appropriate and potential applicability of data mining technology to support the health system as a whole.

Developing a model with different data mining algorithm was constrained with a serious of time problem that was given to undertake this research work. There was also a limitation of time to see alternative data mining tools to run for different applications than the researcher's chosen data mining tool (i.e. WEKA) for predictive and descriptive model development and association rule generation. Therefore the application of data mining techniques would be limited to some data mining algorithms and WEKA data mining tool.

Another limitation of this study was lack of sufficient related works or literatures particularly in relation to data mining applications on communicable disease outbreaks. There was also lack of prior experience to data mining applications and skilled professional in the sector to provide proper and timely feedback to the model as well as to the stated problem in general. Therefore the researcher has tried to analyze the surveillance system and develop a model only from the datasets found in the sector by the use of data mining applications for the first time. Additionally, application of data mining technology in epidemiological datasets especially in disease outbreak surveillance data was a very time consuming and energy intensive process due to

the large amount of data collected and reported to the sector from different locations of the country with timely and weekly bases.

Finally, the data didn't have enough information (low quality) as per the objectives of the research with data mining techniques. For instance the prediction or classification of disease outbreak cases were not simply done on the datasets which consists of small numbers of attributes and with missing the determinant factors of disease outbreak cases to occur.

1.8. Organization of the Thesis

This thesis was divided into six chapters. The first chapter was an introduction part, which contained background to the research work, statement of the problem addressed, objective of the research, and methodologies adopted for the study, Ethical consideration, dissemination of results, scopes and limitation of the study, and organization of the thesis work.

The second chapter dealt about data mining techniques, methods/algorithms used, and its application in the health care sector, limitations of data mining applications and related literature.

The third chapter was devoted to give further understanding about the outbreak surveillance and response system in general and the existing outbreak surveillance and response activities in PHEM in particular.

The fourth chapter provided discussions about business understanding and data preprocessing. In this chapter the main issues are business understanding, data understanding and preprocessing, data preparation, data quality measures, the data collection and analysis tools and techniques, data transformation, and data integration, data reduction, and feature selection processes.

The fifth chapter provides experimentation about the different data mining algorithms and steps that were undertaken in the research work. The issues considered here are model development and testing results for association rules generation, disease data clustering, and

epidemic typhus prediction by using weka data mining software tool. Finally, the sixth chapter was aimed to the final parts of this thesis work. The issues are the concluding remarks and recommendations forwarded based on the research findings.

1.9. Beneficiaries of the research

Even thought the research was done for educational purpose, its findings would be applicable for emerging diseases surveillance and response system of the country for better decision support activities and the data analysis could show the defects or pitfalls of data collection and analysis system of the PHEM sector under the FMoH of Ethiopia. It also provided an input for a better data collection, collation, dissemination, interpretation and presentation purpose to take measurable action for disease outbreak mitigation and if possible eradication for the future occurrences of disease outbreak cases in the country. It would also be a direction to a research in the PHEM sector and for AAU research center and provide a comprehensive documentation for other researchers in the field.

Finally, the researcher provided rewarding recommendation with varied alternatives to the FMoH and the PHEM sector for data collection, collation, dissemination, interpretation, presentation, analysis techniques and tools to the surveillance and response system. As a result, quality data which can enhance the effectiveness, efficiency and total quality management of the system for planning, managing, preparedness, early warning and decision-making processes will be obtained. Therefore, this quality data would be used for a better model development with analytical machine learning tools like data mining.

Chapter Two

Literature Review

2.1. Introduction

With the a rapid growth in size and number of available databases, mining for knowledge, regularities or high-level information from data became essential to support decision-making and predict future behavior **[15].** As stated in **[16]** nowadays the modern science and engineering are based on using first-principle models to describe physical, biological, and social systems**.** As a result of data mining techniques, which are used for achieving the above goals, can be classified into the following categories: classification, clustering, association rule mining, sequential pattern analysis, prediction, data visualization etc.

It has been estimated that the amount of data stored in the world's databases doubles every 20 months, and although it would surely be difficult to justify this figure in any quantitative sense, we can all relate to the pace of growth qualitatively **[17]**. In addition to that, as the flood of data swells and machines that can undertake the searching become common place, the opportunities for data mining increase.

According to **[16],** the above approach starts with a basic scientific model, such as Newton's laws of motion or Maxwell's equations in electromagnetism, and then builds upon them various applications in mechanical engineering or electrical engineering. In this approach, experimental data are used to verify the underlying first-principle models and to estimate some of the parameters that are difficult or sometimes impossible to measure directly. As mentioned in **[18],** it is possible to say that we are overwhelmed with data as mentioned above. It is because of the amount of data in the world, in our lives, continues to increase and there's no end in sight. However, in many domains the underlying first principles are unknown, or the systems under study are too complex to be mathematically formalized. It is also stated that with the growing use of computers, there is a great amount of data being generated by such systems. In the absence of first-principle models, such readily available data can be used to derive models by estimating Thus there is currently a paradigm shift from classical modeling and analyses based on first principles for developing models and the corresponding analyses directly from data **[16].**

These days the Omni-present personal computers make it too easy and provides an inexpensive multi-gigabyte disk to postpone decisions about what to do with all this stuff we simply buy another disk and keep it all**[18]**. In addition to this, **[17]** added the ubiquitous electronics record to our decisions, our choices in the supermarket, our financial habits, are comings and goings. It is also described that as technology grows faster and faster, People have been seeking patterns in data since human life began. It is because of hunters seek patterns in animal migration behavior, farmers seek patterns in crop growth, politicians seek patterns in voter opinion, and lovers seek patterns in their partners' responses and physicians seek patterns in disease outcomes and soon.

As mentioned in **[16], d**ue to the emerging data storage technologies and different electronic instruments, we have grown accustomed gradually to the fact that there are tremendous volumes of data filling our computers, networks, and lives. Therefore; Government agencies, scientific institutions, and business areas have all dedicated enormous resources to collecting and storing data for better data storing techniques and for better use of the information and knowledge from the data**.**

The World Wide Web overwhelms us with information; meanwhile, every choice we make is recorded and all these are just personal choices **[18]**. As a result we would all testify to the growing gap between the *generation* of data and our *understanding* of it. It is also addressed that as the volume of data increases, inexorably, the proportion of it that people understand decreases, alarmingly. Therefore; lying hidden in all this data is information, and those information also can provide potentially useful information or knowledge, which is rarely made explicit or taken advantages of it.

In reality, only a small amount of data will ever be used because, in many cases, the volumes are simply too large to manage, or the data structures themselves are too complicated to be analyzed effectively **[16]**. How could this happen? The primary reason is that the original effort to create a data set is often focused on issues such as storage efficiency; it does not include a plan for how the data will eventually be used and analyzed. Due to the above problems, the need to understand large, complex, information-rich data sets is common. Therefore; the ability to extract useful knowledge hidden in these data and to act on that knowledge is becoming increasingly important in today's competitive world.

2.2. General overview of DM and KD process model

2.2.1. KDD overviews

The enormous expansion of the volume of collected data from different data sources and fields yields to an increase of databases of all sizes and designs **[19]**. However, it is obviously clear that the manual analysis of such amounts of data is impossible when the data sets are increasing and huge. For that reason, it is important to develop tools that can access and analyze such data to extract interesting knowledge that can be helpful in the decision making process. Therefore Knowledge discovery is a technique that can discover interesting information from such large databases.

As stated in **[20],** across a wide variety of fields, data are being collected and accumulated at a dramatic pace. There is an urgent need for a new generation of computational theories and tools to assist humans in extracting useful information (knowledge) from the rapidly growing volumes of digital data. As a result these theories and tools are becoming the subject of the emerging field of KDD.

"KDD is the non-trivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data**"[19].** It was also mentioned that the process is nontrivial means that it is not straightforward and some search methodology need to be followed; the extracted patterns should be novel or in other words unknown previously; they also have to be useful, this usefulness implies that the patterns can play a positive role in the decision making process; finally, they must be understandable at least for the data analyst and it run with a number of iterative steps**.**

The unifying goal of KDD is extracting high-level knowledge from low-level data in the context of large data sets including how the data are stored and accessed, how algorithms can be scaled to massive data sets. Therefore; KDD and DM terms are relatively new, but their applications had been developed much earlier than the inception of these definitions **[21]**. Additionally KDD is generally defined as the process of extracting implicit information and knowledge from data. KDD also helps humans make sense of huge amounts of data by mining patterns.

2.2.2. Model Development Processes and Techniques

2.2.2.1. KD Process model

Historically, the notion of finding useful patterns in data has been given a variety of names, including data mining, knowledge extraction, information discovery, information harvesting, data archaeology, and data pattern processing **[20]**. Additionally, KDD has evolved, and continues to evolve, from the intersection of research fields such as machine learning, pattern recognition, databases, statistics, AI, knowledge acquisition for expert systems, data visualization, and high-performance computing. In general the driving force behind KDD is the database field (the second D in KDD).

Academic Research Models for KD process

The efforts to establish a KDP model were initiated in academia. In the mid-1990s, when the DM field was being shaped, researchers started defining multistep procedures to guide users of DM tools in the complex knowledge discovery world. The main emphasis was to provide a sequence of activities that would help to execute a KDP in an arbitrary domain. All the activities are set as a continuation of the previous of former iterations and the next iteration is the products of the preceding steps.

Industrial Models for KD process

Industrial models quickly followed academic efforts. Several different approaches were under-taken, ranging from models proposed by individuals with extensive industrial experience to models proposed by large industrial consortiums. Two representative industrial models are the five-step model by Cabena et al., with support from IBM and the industrial six-step CRISP-DM model, developed by a large consortium of European companies. The latter has become the leading industrial model. The CRISP-DM (CRoss-Industry Standard Process for Data Mining) was first established in the late 1990s. The development of this process model enjoys strong industrial support **[22].**

The goal of the project was to define and validate an industry- and tool-neutral data mining process model that which would make the development of large as well as small data mining projects faster, cheaper, more reliable and more manageable **[23].** The following picture shows the CRISP-DM with six consecutive steps. As shown in the picture some of the processes are iterative and the others are straight forward directly to the next process.

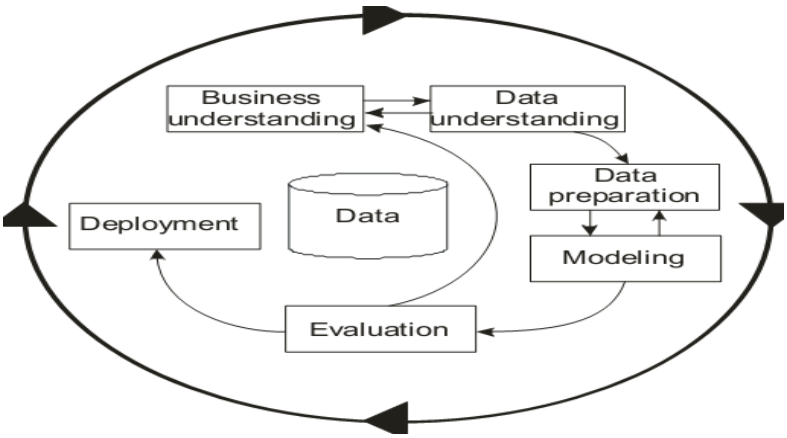


Figure 1: The CRISP_DM process model

2.2.2.2. Hybrid KD Process Models

As explained in **[22],** since the 1990s, several different KDPs have been developed. The initial efforts were led by academic research but were quickly followed by industry. The first basic structure of the model was proposed by Fayyad et al. and later improved/modified by others. The process consists of multiple steps, which are executed in a sequence. Each subsequent step is initiated upon successful completion of the previous step, and requires the result generated by the previous step as its input. Another common feature of the proposed models is the range of activities covered, which stretches from the task of understanding the project domain and data, through data preparation and analysis, to evaluation, understanding, and application of the generated results. All the proposed models also emphasize the iterative nature of the model, in terms of many feedback loops that are triggered by a revision process. As mentioned above it is also the combination of the educational and industrial knowledge process models.

Although the models usually emphasize independence from specific applications and tools, they can be broadly divided into those that take into account industrial issues and those that do not. However, the academic models, which usually are not concerned with industrial issues, can be made applicable relatively easily in the industrial setting and vice versa **[22]**

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

The main differences and extensions of the hybrid and CRISP-DM model include:

- providing more general, research-oriented description of the steps,
- introducing a data mining step instead of the modeling step,
- introducing several new explicit feedback mechanisms, (the CRISP-DM model has only three major feedback sources as shown in the figure above, while the hybrid model has more detailed feedback mechanisms as shown below in the figure) and
- Modification of the last step, since in the hybrid model, the knowledge discovered for a particular domain may be applied in other domains.

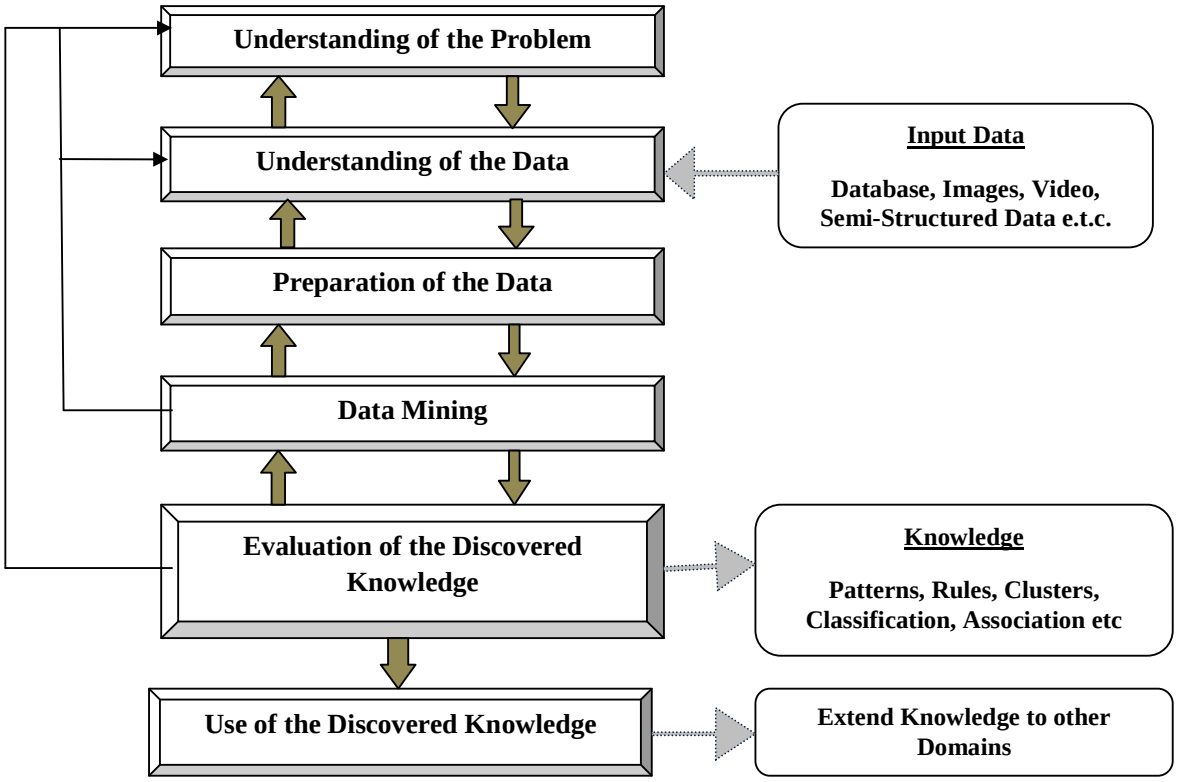


Figure2: The six-step Hybrid KDP Model

1. **Understanding of the problem domain**. This initial step involves working closely with domain experts to define the problem and determine the project goals, identifying key people, and learning about current solutions to the problem. It also involves learning domain-specific terminology. 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. **Understanding of the data**. This step includes collecting sample data and deciding which data, including format and size, will be needed. Background knowledge can be used to guide these efforts. Data are checked for completeness, redundancy, missing values, plausibility of attribute values, etc. Finally, the step includes verification of the usefulness of the data with respect to the DM goals.

3. **Preparation of the data**. This step concerns deciding which data will be used as input for DM methods in the subsequent step. It involves sampling, running correlation and significance tests, and data cleaning, which includes checking the completeness of data records, removing or correcting for noise and missing values, etc. The 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 methods and techniques to derive knowledge from preprocessed data. Appropriate data mining Algorithm will be selected and tested to develop better discovered knowledge.

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.

Finally the major application of the hybrid model has been used in medicine and software development areas. Example applications include development of computerized diagnostic systems for cardiac SPECT images and a grid data mining framework called GridMiner-Core. It has also been applied to analysis of data concerning intensive care, cystic fibrosis, and image-based classification of cells and outbreak detection and recognition.

Understanding the very important aspect of the KDP is the relative time spent in completing each of the steps. Evaluation of this effort enables precise scheduling. Several estimates have been proposed by researchers and practitioners alike. The specific estimated values depend on many factors, such as existing knowledge about the considered project domain, the skill level of human resources, and the complexity of the problem at hand, to name just a few. The common theme of all estimates is an acknowledgment that the data preparation step is by far the most time-consuming part of the KDP.

2.3. Data mining

According to **[24],** the history of DM & KD is not much different. In the early 1990s, when the KDD processing term was first coined, there was a rush to develop DM algorithms that were capable of solving all the problems of searching for knowledge in data. Apart from developing algorithms, tools were also developed to simplify the application of DM algorithms. From the viewpoint of DM & KD process models, the year 2000 marked the most important milestone. The number of applications which are applied in the data mining and knowledge discovery (DM & KD) projects has increased enormously over the past few years. Additionally in [24], it is also described that as DM & KD development projects became more complex, a number of problems emerged: continuous project planning delays, low productivity and failure to meet user expectations.
THIS

As described in **[25],** Data Mining is the process of extracting information and knowledge from large data sets through the use of algorithms and techniques drawn from the field of: Statistics, Machine Learning and Data Base Management Systems as stated above. In **[17],** Data mining is defined as the process of discovering patterns in data. The process must be automatic or (more usually) semiautomatic. Data mining is also defined in **[16]**, as an iterative process within which progress by discovery, through either automatic or manual methods. Therefore; the term data mining has mostly been used by statisticians, data analysts, and the management information systems (MIS) communities and gained popularity in the database field. It is also shown that Knowledge is the end product of a data-driven discovery and has been popularized in the AI and machine-learning fields.

Traditional data analysis methods often involve manual work and interpretation of data that is slow, expensive and highly subjective. The process of Data Mining, popularly called as knowledge discovery in large data, enables firms and organizations to make calculated decisions by assembling, accumulating, analyzing and accessing very large corporate data **[25].** As a result**,** it uses variety of tools like query and reporting tools, analytical processing tools, and Decision Support System (DSS) tools for the analysis and the process of data mining or knowledge activities.

With a use of data mining approach stated in **[17];** the patterns discovered must be meaningful in that they lead to some advantage, usually an economic advantage. The data is invariably present in substantial quantities and useful patterns allow us to make nontrivial predictions on new data. Due to the large amounts of data at the present time described in **[16],** Data mining is most useful in an exploratory analysis scenario in which there are no predetermined notions about what will constitute an "interesting" outcome. In general data mining, as mentioned in the KDD above, is the process to the search for new, valuable, and nontrivial information in large volumes of data and a cooperative effort of humans and computers. Here the best results from structural patterns are achieved by balancing t8he knowledge of human experts in describing problems and goals with the search capabilities of computers.

According to **[17],** Even if the data is stored electronically and searched by computers, economists, statisticians, forecasters, and communication engineers have long worked with the idea that patterns in data can be sought automatically, identified, validated, and used for prediction purposes. What is new is the staggering increase in opportunities for finding patterns in data is very inspiring. The uncontrolled growth of databases in recent years, databases on such everyday activities as customer choices, brings data mining to the forefront of new business technologies.

As the world grows in complexity, overwhelming us with the data it generates, data mining becomes our only hope for elucidating the patterns that underlie it. Intelligently analyzed data is a valuable resource and it can lead to new insights into commercial settings, new investigations, new knowledge, and new inventions in different field of studies for competitive advantages **[17].**

Data mining is one of the fastest growing fields in the computer industry **[16]**. Once a small interest area within computer science and statistics, it has quickly expanded into a field of its own. One of the greatest strengths of data mining is reflected in its wide range of methodologies and techniques that can be applied to a host of problem sets. Since data mining is the entire data warehousing, data-mart, and decision-support community, encompassing professionals from such industries as retail, manufacturing, telecommunications, healthcare, insurance, transportation, and business process reengineering.

There are two primary goals of data mining. These are in general prediction and description activities. Prediction involves using some variables or fields in the data set to predict unknown or future values of other variables of interest. Description, on the other hand, focuses on finding patterns describing the data that can be interpreted by humans **[16].** Therefore, it is possible to put data-mining activities into one of the two categories:

1. Predictive data mining, which produces the model of the system described by the given data, or
2. Descriptive data mining, which produces new, nontrivial information based on the available data

On the predictive end of the spectrum, the goal of data mining is to produce a model, expressed as an executable code, which can be used to perform classification, prediction, estimation, or other similar tasks. On the other, descriptive, end of the spectrum, the goal is to gain an understanding of the analyzed system by uncovering patterns and relationships in large data sets. The relative importance of prediction and description for particular data-mining applications can vary considerably.

2.3.1. Data mining functionalities

As stated in **[26]**, there are much functionality in data mining applications which provides a useful results and patterns: such as Characterization, Discrimination, Association, Classification, Clustering analyzes, and Data evolution analysis:

- **Characterization** is a summarization of the general characteristics or features of a target class of data. For example, the characteristics of students can be produced, generating a profile of all the University first year computing science students, which may include such information as a high GPA and large number of courses taken.
- **Discrimination** is a comparison of the general features of target class data objects with the general features of objects from one or a set of contrasting classes. For example, the general features of students with high GPA's may be compared with the general features of students with low GPA's. The resulting description could be a general comparative profile of the students such as 75% of the students with high GPA's are fourth-year computing science students while 65% of the students with low GPA's are not.
- **Association** is the discovery of association rules showing attribute-value conditions that occur frequently together in a given set of data. For example, a data mining system may find association rules like major(X, "computing science"") ⇒ owns(X, "personal computer") [support = 12%, confidence = 98%] where X is a variable representing a student. The rule indicates that of the students under study, 12% (support) major in computing science and own a personal computer. There is a 98% probability (confidence or certainty) that a student in this group owns a personal computer.
- **Classification** differs from prediction in that the former is to construct a set of models (or functions) that describe and distinguish data class or concepts, whereas the latter is to predict some missing or unavailable, and often numerical, data values. Their similarity is that they are both tools for prediction: as mentioned above classification is used for predicting the class label of data objects and prediction is typically used for predicting missing numerical data values.
- **Clustering analyzes** data objects without consulting a known class label. The objects are clustered or grouped based on the principle of maximizing the intra-class similarity and minimizing the inter-class similarity. Each cluster that is formed can be viewed as a class of objects. Clustering can also facilitate taxonomy formation, that is, the organization of observations into a hierarchy of classes that group similar events together.
- **Data evolution analysis** describes and models regularities or trends for objects whose behavior changes over time. Although this may include characterization, discrimination, association, classification, or clustering of time-related data, distinct features of such an analysis include time-series data analysis, sequence or periodicity pattern matching, and similarity-based data analysis.

2.2.3. Core Ideas in Data mining process

In general as stated in **[27],** data mining have the following important and core ideas in processing a large datasets to discover knowledge in the data.

Classification

As mentioned above in classification and prediction part of data mining functionalities, classification is perhaps the most basic form of data analysis. For example the recipient of an offer can respond or not respond; an applicant for a loan can repay on time, repay late or declare bankruptcy; a credit card transaction can be normal or fraudulent; the victim of an illness can be recovered, still ill, or deceased; the occurrence of outbreak is detected early, or late; a disease outbreak occur not occur; and so on. A common task in data mining is to examine data where the classification is unknown or will occur in the future, with the goal of predicting what that classification is or will be. Similar data where the classification is known are used to develop rules, which are then applied to the data with the unknown classification.

Prediction

As mentioned above, Prediction is similar to classification, except we are trying to predict the value of a continuous numerical variable (e.g., amount of purchase, salary of an employee, number of outbreak in a certain region or location), rather than a class (e.g. purchaser or non purchaser, salary of par time employee or full time employee, occurrence of outbreak under expected, expected or above expected level). Of course, in classification we are trying to predict a class, but the term \prediction" in this context is the prediction of the value of a continuous variable. (Sometimes in the data mining literature, the term \estimation" is used to refer to the prediction of the value of a continuous variable, and \prediction" may be used for both continuous and categorical data.)

Association Rules/Affinity analysis

Large databases of customer transactions lend themselves naturally to the analysis of associations among items purchased, or \what goes with what." *Association rules* or *affinity analysis* can then be used in a variety of ways. For example, grocery stores can use such information after a customer's purchases have all been scanned to print discount coupons, where the items being discounted are determined by mapping the customer's purchases onto

the association rules. Another example of association rule mining is that, the occurrence of a disease outbreak due to the occurrence of the other will be visualized here.

Predictive Analytics

Classification, prediction, and to some extent affinity analysis (association rule mining), constitute the analytical methods employed in \predictive analytics."

Data Reduction

As we know natural data collection may be complicated and huge. Therefore; Sensible data analysis often requires distillation of complex data into simpler data. Rather than dealing with thousands of product types, an analyst might wish to group them into a smaller number of groups. This process of consolidating a large number of variables (or cases) into a smaller set is termed as data reduction.

Data Exploration

Unless our data project is very narrowly focused on answering a specific question determined in advance (in which case it has drifted more into the realm of statistical analysis than of data mining), an essential part of the job is to review and examine the data to see what messages they hold, much as a detective might survey a crime scene. Here, full understanding of the data may require a reduction in its scale or dimension to allow us to see the forest without getting lost in the trees. Since Similar variables (i.e. variables that supply similar information) might be aggregated into a single variable incorporating all the similar variables. In other way round or analogously, records might be aggregated into groups of similar records.

Data Visualization

Another technique for exploring data to see what information they hold is through graphical analysis. This process includes looking at each variable separately as well as looking at relationships between variables. For numeric variables we use histograms and box-plots to learn about the distribution of their values, to detect outliers (**extreme observations**), and to find other information that is relevant to the analysis task. Similarly, for categorical variables we use bar charts and pie charts. We can also look at scatter plots of pairs of numeric

variables to learn about possible relationships, the type of relationship, and again, to detect outliers **[38].**

2.3. Supervised and Unsupervised Learning in data mining

According to [27], a fundamental distinction among data mining techniques is between supervised and unsupervised learning methods. Supervised learning" algorithms are those used in classification and prediction. We must have data available in which the value of the outcome of interest is known. These training data are the data from which the classification or prediction algorithm learns, or is trained, about the relationship between predictor variables and the outcome variable. Once the algorithm has learned from the training data, it is then applied to another sample of data (the validation data) where the outcome is known, to see how well it does in comparison to other models. If many different models are being tried out, it is prudent to save a third sample of known outcomes (the test data) to use with the final, selected model to predict how well it will do. Finally, in **[27] t**he model can then be used to classify or predict the outcome of interest in new cases where the outcome is unknown**.**

Supervised learning is used to estimate an unknown dependency from known input-output samples and the so called Classification and regression are common tasks supported by this type of learning **[16]**. Additionally, the term "supervised" denotes that, the output values for training samples will be classified with respect to the known classes. Different types of techniques support this kind of learning, and some of them such as prediction like logistic regression, and classification like decision rules and decision trees. Under the unsupervised learning scheme, only samples with input values are given to a learning system, and there is no notion of the output during the learning process. The goal of unsupervised learning is to discover "natural" structure in the input data. The best example is description data modeling techniques like clustering and association rule mining.

Finally, Simple linear regression analysis is an example of supervised learning (though rarely called that in the introductory statistics course where you likely first encountered it). The *Y* variable is the (known) outcome variable and the *X* variable is some predictor variable **[27]**. Unsupervised learning algorithms are those used where there is no outcome variables like the

supervised one to predict or classify. Association rules, data reduction methods and clustering techniques are all unsupervised learning methods**.**

2.4. Classification and Prediction

Databases are rich with hidden information that can be used for intelligent decision making processes. Classification and prediction are two forms of data analysis that can be used to extract models describing important data classes or to predict future data trends. Such analysis can help provide us with a better understanding of the data at large **[28]**.

According to **[16],** classification is a process of learning a function that maps a data item into one of several predefined classes. As a result the goal of learning is formulate to create a classification model, known as a *classifier*, which will predict, with the values of its available input attributes, the class for some entity (a given sample) will be analyzed. In other words, classification is the process of assigning a discrete label value (class) to an unlabeled record, and a classifier is a model (a result of classification) that predicts one attribute-class of a sample-when the other attributes are given. In doing so, samples are divided into predefined groups**.**

Many classification and prediction methods have been proposed in machine learning, pattern recognition, and statistics **[28]**. Nowadays, data mining research has built on such work, developing scalable classification and prediction techniques capable of handling large disk-resident data and have numerous applications, including fraud detection, target marketing, performance prediction, manufacturing, medical diagnosis, and disease outbreak prediction**.**

Categories in classification can be represented by discrete values, whereas the ordering among values has no meaning **[16]**. For example, the values 1, 2, and 3 may be used to represent treatments A, B, and C, where there is no ordering implied among this group of treatment regimes. It was also stated in **[16]** the data analysis task is an example of numeric prediction, where the model constructed predicts a *continuous-valued function*, or *ordered* value, as opposed to a categorical label. This model is then a predictor model. In reality, we may expect many more attributes to be considered. In the first step, a classifier is built, describing a predetermined set of data classes or concepts.

2.4.1. Comparison between Classification and Prediction Methods

As mentioned above in the previous section, classification and prediction methods can be compared and evaluated. As mentioned in **[26],** the following criteria are used for the process of evaluation:

Accuracy: The accuracy of a classifier refers to the ability of a given classifier to correctly predict the class label of new or previously unseen categorical data. Similarly, the accuracy of a predictor refers to how well a given predictor can guess the value of the predicted attribute for new or previously unseen continuous data. **Robustness:** This is the ability of the classifier or predictor to make correct predictions given noisy data or data with missing values.

Scalability: This refers to the ability to construct the classifier or predictor efficiently given large amounts of data. That means whatever the amount of data given to the system; it must works all the time.

Interpretability: This refers to the level of understanding and insight that is provided by the classifier or predictor and which is subjective and therefore more difficult to assess. Recently, data mining research has contributed to the development of scalable algorithms for classification and prediction and exploration of mined "associations" between attributes and their use for effective classification **[28].**

2.5. Methods of Data preparation for data mining algorithms

As described in **[26]**, the preprocessing steps may be applied to the data to help the improvement of the accuracy, efficiency, and scalability of the classification or prediction process. The following are activities taken for the preprocessing phase:

Data cleaning: This refers to the preprocessing of data in order to remove or reduce noise (by applying smoothing techniques or binning) and the treatment of missing values (e.g., by replacing a missing value with the most commonly occurring value for that attribute, or with the most probable value based on statistics).

Relevance analysis: Many of the attributes in the data may be redundant. Therefore correlation analysis can be used to identify whether any two given attributes are statistically

related or not. For example, a strong correlation between attributes A1 and A2 would suggest that one of the two could be removed from further analysis. A database may also contain irrelevant attributes. As a result; Attribute subset selection, concept hierarchy generation and best feature generation can be used in these cases to find a reduced set of attributes such that the resulting probability distribution of the data classes is as close as possible to the original distribution obtained using all attributes. This is done by the relevance analysis which can be used to detect attributes that do not contribute to the classification or prediction task. It is because; including such attributes may otherwise slow down, and possibly mislead, the learning step. Hence, such analysis can help improve classification efficiency and scalability.

Data transformation and reduction: The data we are obtaining from different sources may be transformed by normalization particularly when neural networks or methods involving distance measurements are used in the learning step. With Normalization we can scale all values for a given attribute so that they fall within a small specified range, such as between -1:0 to 1:0, or 0:0 to 1:0. Data transformation can also be done by generalizing it to higher-level concepts (i.e. Concept hierarchies may be used for this purpose). This is particularly useful for continuous valued attributes. For example, numeric values for the attribute number of epidemic typhus can be generalized to discrete ranges, such as disease outbreak occurrences and none occurrences. Similarly, categorical attributes, like epidemic weeks, can be generalized to higher-level concepts, like months. Because of generalization compresses the original training data, fewer input/output operations may be involved during learning and details may be hidden.

2.6. Association Rules Mining

Association rule mining and its sub problem, frequent itemsets mining, is one of the most popular and well researched areas in data mining **[21].** It is also used for finding relationships (or association rules, patterns) between different variables in the data.

The process of fining association rule mining falls within the preview of database mining, also called knowledge discovery in database. It is used to infer functional dependencies and is resulted probabilistic in nature. It is shown by the presence of a dependency rule such that Y➔ A does not necessarily mean that X+Y➔A also because the later may not have

minimum support. Similarly, the presence of rules X➔Y and Y➔Z does not necessary mean that X➔Z holds because the later may not have minimum confidence **[29].**

Association rules are one of the major techniques of data mining and it is perhaps the most common form of local-pattern discovery in unsupervised learning systems. It is a form of data mining that most closely resembles the process that most people think about when they try to understand the data-mining process; namely, "mining" for gold through a vast database. The gold in this case would be a rule that is interesting, that tells you something about your database that you didn't already know and probably weren't able to explicitly articulate **[16]**.

Association rules provide information about relationships between different variables in data. They are sometimes described as if-then type of rules. This can better be understood with an example: (malaria; epidemic typhus) ➔ typhoid fever is a typical association rule discovered from a database. In a rule such as *A* ➔ *B*, the left hand side, A, is called the antecedent and the right hand side, B, is called the consequent. A and B in the rule are item-sets containing items existing in the database, and their intersection is the empty set **[21]**.

There are two important measures of association rules mining **[21]**. The first is the Support value and the second one is the confidence value. The Support value is used to prune the search space mining algorithms will be run on. It is used with a minimum support value, also called support threshold, and items with support values lower than the set minimum are pruned. These items, by the definition of support, are not frequently observed in the database, and thus not taken into consideration. The Confidence metric is a little more controversial than the support metric. As in the support case, a minimum confidence value, also called confidence threshold, is used to prune those rules having less confidence value than the minimum. A high confidence value shows a high correlation between the antecedent and the consequent**.**

Another measure for association rules mining is the lift of a rule which measures the correlation among the antecedent and consequent predicates of a rule **[30]**. It is also stated that a lift value of 1 would indicate that the predicates expressed in the rule are independent of each other while a value greater than 1 would indicate a positive correlation among the antecedent and consequent predicates, and a value less than 1 would indicate a negative correlation among the antecedent and consequent predicates of the rule**.**

It is often desirable to pay attention to only those rules with high confidence and strong support, which is referred to as strong rules **[21][29]**. In addition to this, the task of mining association rules is essentially to discover strong association rules in large databases. Here we have two important phases in this process:

1. Discover the large itemsets, i.e., the sets of items that have transaction support S above a predetermined minimum threshold. This determines the overall performance of mining association rules.
2. Use the large item-sets to generate the association rules for the database that have confidence C above a predetermined minimum threshold**.**

2.7. Data Clustering Applications

Clustering is a division of data into groups of similar objects. Representing the data by fewer clusters necessarily loses certain fine details, but achieves simplification. It models data by its clusters **[31].** An operational definition of clustering can be stated as follows: Given a representation of n objects, find K groups based on a measure of similarity such that the similarities between objects in the same group are high while the similarities between objects in different groups are low. I.e. for good clustering activities, the intra-cluster similarities between objects are high and the inter-cluster similarities between objects are low. The notion of similarity may be due to their density, size, number, and types of cluster assignments **[32].**

As stated in **[33],** Clustering refers to the grouping of records, observations, or cases into classes of similar objects from unsupervised datasets. Therefore; a cluster is a collection of records that are similar to one another and dissimilar to records in other clusters. Clustering differs from classification in that there is no specified class or target variable for clustering. The clustering task does not try to classify, estimate, or predict the value of a target variable. Instead, clustering algorithms seek to segment the entire data set into relatively homogeneous subgroups or clusters, where the similarity of the records within the cluster is maximized, and the similarity to records outside this cluster is minimized.

The goal of data clustering, also known as cluster analysis, is to discover the *natural* grouping(s) of a set of patterns, points, or objects **[32]**. It is also defined that cluster analysis as "a statistical classification technique for discovering whether the individuals of a population fall into different groups by making quantitative comparisons of multiple characteristics." Therefore, clustering is to develop an automatic algorithm that will discover the natural groupings in the unlabeled data.

As stated in **[34],** many clustering methods exist to partition a data set by some natural measure of similarity. This similarity measure places similar objects close to one another forming a group, thus several clusters related to objects are formed.

Clustering algorithms can be broadly divided into two groups **[32],** hierarchical and partitional clustering. The first one is Hierarchical clustering algorithms in which the algorithm recursively find nested clusters either in agglomerative mode (starting with each data point in its own cluster and merging the most similar pair of clusters successively to form a cluster hierarchy) or in divisive (top-down) mode (starting with all the data points in one cluster and recursively dividing each cluster into smaller clusters). The second is partitional clustering algorithms, which is Compared to hierarchical clustering algorithms, it finds all the clusters simultaneously as a partition of the data and do not impose a hierarchical structure. Partitioning clustering algorithm can find similarity between data objects in such a way that the distance between cluster assignments will be calculated in order to cluster the objects.

2.8. Application of data mining in Health Care

As stated in **[35],** Medical data mining has great potential for exploring the hidden patterns in the datasets of the medical domain. These patterns can be utilized for clinical diagnosis, pattern analysis, disease investigation, clinical decision making, and disease outbreak investigation and so on. However, the available raw medical data are widely distributed, heterogeneous in nature, and voluminous in size. These data need to be collected in an organized form and then integrated to form a hospital information system or medical

information system. Hence, Data mining technology provides a user oriented approach to novel and hidden patterns in the medical data**.**

Data mining outcomes are seen as generalizations and are not verified for medical validity or accuracy **[36].** Nowadays, it is also known that medicine is a highly complex domain for which data mining processes were not designed well. It is because of medicine has requirements which are outside the original scope of the technology, and to be applicable to a science which is concerned with critical decision making. There is a need to modify the technology to reflect this different environment and careful consideration of the language used when reporting results of data mining applications in medical science.

Medical diagnosis is regarded as one of an important yet complicated task that needs to be executed accurately and efficiently **[35]**. The automation of this system would be extremely advantageous. Regrettably all doctors do not possess expertise in every sub specialty and moreover there is a shortage of resource persons at certain places. Therefore, an automatic medical diagnosis system would probably be exceedingly beneficial by bringing all of them together. It is added that, appropriate computer-based information and/or decision support systems can aid in achieving clinical tests at a reduced cost. It is an efficient and accurate implementation of automated system which needs a comparative study of various techniques available. Here, in this research the aim was to analyze the PHEM sector disease outbreak datasets by the application of data mining algorithms and techniques for various predictive and descriptive data mining results proposed to the outbreak surveillance and response system of Ethiopia **[35].**

Nowadays the importance of decision support system in the delivery of managed healthcare can hardly be overemphasized **[25].** A variety of decision support capabilities will be necessary to increase the productivity of medical personnel, analyze care outcomes, and continually refine care delivery processes to remain profitable while holding the line on costs and maintaining quality of care. Healthcare decision support is faced with the challenges of complex and diverse data and knowledge forms and tasks, the lack of standardized terminology compared to basic sciences, the stringent performance and accuracy requirements and the prevalence of legacy systems. As a result data mining framework enables specialists to create customized models that can be shared throughout the

organization, making the application attractive to skilled modelers in a pharmaceutical company's and bioinformatics division **[25].**

As stated in **[36],** there is a scientific data mining process which is applied with the language of information presentation. But sometimes it is misleading and the resultant reporting is inaccurate and medically invalid. For example, in some cases where a mining outcome showed that smoking does not have a direct link with skin cancer. However the resulting media story reported that smoking is not linked with cancer generally. Therefore; is it misleading to make a decision about the outcome because it doesn't have a direct link won't show that it doesn't cause at all. More over medicine is especially sensitive to this form of information distortion and the consequences have the potential to be life threatening, politically sensitive, costly and persistent which is rarely the case in other domains**.**

As described in **[27],** Data mining in medicine is most often used to complement and expand the work of the clinician and researcher by qualifying or expanding knowledge rather than providing new knowledge, as has been the trend in other domains. Very little health data mining is purely exploratory and hence it is generally not applied to provide novel knowledge, that is, to identify new patterns hidden within the data. One of the difficulties in providing new knowledge in the health domain is the need to sufficiently cross reference and validate the results. It is not sufficient to provide a standard rule in the form of A gives B in the presence of C without substantiating the information held therein**.**

Finally, as stated in **[38],** the healthcare environment is generally perceived as being 'information rich' yet 'knowledge poor'. Therefore; there is a wealth of data available within the healthcare systems. However, there is a lack of effective analysis tools to discover hidden relationships and trends in data. Since, Knowledge discovery and data mining have found numerous applications in business and scientific domain, valuable knowledge can be discovered from application of data mining techniques in healthcare system also. That means the healthcare industry collects huge amounts of healthcare data which, unfortunately, are not "mined" to discover hidden knowledge**.** Hence, there are vast potential for data mining application in health care to provide effective and efficient decision making processes. Generally, these can be grouped as the evaluation of treatment effectiveness; management of

healthcare; customer relationship management; and detection of fraud, outbreak and abuse and soon **[39]**.

Treatment effectiveness: Data mining applications can be developed to evaluate the effectiveness of medical treatments. By comparing and contrasting causes, symptoms, and courses of treatments, data mining can deliver an analysis of which courses of action prove effective. For example, the outcomes of patient groups treated with different drug regimens for the same disease or condition can be compared to determine which treatments work best and are most cost-effective. Similarly, data mining can help identify successful standardized treatments for specific diseases.

Healthcare management: - To aid healthcare management, data mining applications can be developed to better identify and track chronic disease status and high-risk patients, design appropriate interventions, and reduce the number of hospital admissions and claims. Recently, data mining are used extensively to identify areas for health care quality improvements, including treatment guidelines, disease management groups, and cost management. The Lightweight Epidemiological outbreak emergency management system is also one of the applications of data mining techniques to enhance health care provisions.

Customer relationship management*:* - While customer relationship management is a core approach in managing interactions between commercial organizations typically banks and retailers and their customers, it is no less important in a healthcare context. Customer interactions may occur through call centers, physicians' offices, billing departments, inpatient settings, and ambulatory care settings. As in the case of commercial organizations, data mining applications can be developed in the healthcare industry to determine the preferences, usage patterns, and current and future needs of individuals to improve their level of satisfaction. These applications also can be used to predict other products that a healthcare customer is likely to purchase, whether a patient is likely to comply with prescribed treatment or whether preventive care is likely to produce a significant reduction in future utilization.

***Fraud and abuse:* -** Data mining applications that attempt to detect fraud and abuse often establish norms and then identify unusual or abnormal patterns of claims by physicians, laboratories, clinics, or others. Among other things, these applications can highlight

inappropriate prescriptions or referrals and fraudulent insurance and medical claims. Detection and investigation of disease outbreak cases are also the major application of data mining techniques **[39].**

2.9. Limitations of Data Mining Applications

As discussed in **[16],** despite many success stories often presented by vendors and consultants to show the benefits that data mining provides, this technology has also several pitfalls. These pitfalls are detected more when used improperly; as a result, data mining can generate lots of "garbage". "Given enough time, enough attempts, and enough imagination, almost any set of data can be teased out of any conclusion". It is also added that there are four pitfalls of data mining: The first is tempting to develop a theory to fit an oddity found in the data. The second One can find evidence to support any preconception if you let the computer churn long enough. The third is finding makes more sense if there is a plausible theory for it. But a beguiling story can disguise weaknesses in the data. The fourth and the more factors or features in a data set the computer considers, the more likely the program will find a relationship, valid or not. It is crucial to realize that data mining can involve a great deal of planning and preparation. Just having a large amount of data alone is no guarantee of the success of a data-mining project. In the words of one senior product manager from Oracle said: "Be prepared to generate a lot of garbage until you hit something that is actionable and meaningful for your business".

As already mentioned above in the document data mining applications can greatly benefit the health care industry without limitation. However; according to **[40][39],** Health care data mining is limited by about four major factors: the first is accessibility of compiled health care related data, because the raw input for data mining can be exist in different settings and systems, such as administration, clinics and laboratories and more. Hence the data have to be collected and integrated before data mining can be done on health care related activities. It is also suggested that a data warehouse should be built before data mining is attempted, that can be a costly and time consuming project. Secondly other data problems may arise. This includes missing, corrupted, inconsistent and none standardize data Such as pieces of information recorded in different formats in different data sources. In particular the lack of clinical vocabulary is a serious hindrance to data mining. Thirdly, the successful application

of data mining requires knowledge of the domain area as well as in data mining methodology and tools. Without sufficient knowledge of data mining the user may not be aware of or be able to avoid the pitfalls of data mining. Collectively, the data mining in medical application users should possess domain knowledge of health related activities, statistical and research expertise, and IT and data mining knowledge and skills.

As stated in **[39],** a data warehouse has been successfully built for different sources of data to apply data mining in clinical data repository, acute care case-mix system, laboratory information system, ambulatory case-mix system, and health plans database and used to find and implement better evidence-based clinical solutions. In particular, the lack of a standard clinical vocabulary is a serious hindrance to data mining**.**

Finally, health care organizations developing data mining applications must also make a substantial investment of human and material resources, particularly: time, effort, and money, and IT as well as Health related professionals. Since health care data mining projects can fail for a variety of reasons such as lack of management support, unrealistic user expectation, poor project management, and inadequate data mining expertise and more **[40].** Therefore, information technology professionals and public health professionals have to work cooperatively to enhance the decision support systems of health related cares.

2.10.Related works

A research conducted in **[42],** on the utilization of Data- Mining Techniques for Evaluation of Patterns of Asthma Drugs Use by Ambulatory Patients in a Large Health Maintenance Organization in 2007. The aim was to identify outlier patients continuously using a combination of several medications. They construct a prediction model for asthma drug utilization by applying novel methods of knowledge discovery in time-series databases to a multi-year asthma drug utilization data set. The methods of mining utilization patterns combine clustering algorithms, clustering validity measures, and decision-tree classification algorithms. This methodology is applied to a regional patients' database maintained in 'Clalit Health Services' HMO, Beer-Sheva, Israel between January 2000 and November 2002. The clustering results reveal that 274 asthma patients who received 9,319 prescriptions during that period can be partitioned into three groups of utilization patterns, where ten patients (3.6%) who used 1,333 prescriptions (14.3%) are classified as outliers. The classification results show that the use of corticosteroids medications (oral or by inhalation) and the age of a patient can be considered as the main predictive factors in the induced models.

Health researchers also have seen how data mining application is used for medical research like influenza surveillance. In **[43],** the researchers used the Royal Thai Army (RTA) influenza surveillance dataset based on 578 respiratory illness cases during March 2007 – April 2008 from 6 RTA hospitals. The data set included the following information: Demographic: age, date, sex, site, travel history, vaccination history; Symptom: oral or axillary temperature chills, cough, fatigue, headache, malaise, snuffy nose, sore throat. Laboratory test results: rapid diagnostic tests and reverse transcription polymerase chain reaction (RT-PCR) for influenza A or B Out of the 578 cases, 106 were influenza A, 27 influenza B and 445 negative for influenza by RTPCR and this was used as the gold standard for comparison to all other variables. They used the Waikato Environment for Knowledge Analysis (Weka) software tool from the University of Waikato for the Association Rule: the apriori algorithm depicts attributes that occur together frequently in the RTA influenza dataset with a minimum confidence = 0.9. The 20 best pattern rules for a confirmed Influenza A case were the following attributes which occurred together: fatigue, headache, malaise, snuffy nose and positive for influenza A by rapid diagnostic test. For classification: using the C4.5 algorithm, they used a cross validation and ran the classifier 10 times with 10 folds.

This model generated main predictive variables: demographics (age, site, date, vaccination status); symptoms (chills, cough, fatigue, headache, oral or axiliary temperature); and laboratory test results (rapid diagnostic tests). The model correctly classified variables in 83.4% of all cases. Sensitivity was 48.1% for a positive influenza case using the model variables, but had 94.6% specificity. Finally, the research showed that data mining can be applied to influenza datasets to extract predictive information about influenza epidemics, including demographics, symptoms, and laboratory results. Decision tree models could also be useful for influenza surveillance and strategic planning in control of influenza. The analysis can assist in promising an investigation of a case that has symptoms related to Influenza.

As justified in **[40],** the potential application of data mining techniques in support of predicting health seeking behavior of households in the district of Butajira has been explored in 2010. To do that CRISP_DM methodology was used. A total of 60446 records were used for the experiments. The J48 decision tree classifier with C4.5 algorithms was used. On the result decision tree classification for the household health seeking behavior on BRHP dataset were applied. The J48 decision tree classification algorithm classified records with the accuracy rate of 89.9017%. Even if the highest accuracy value of the model was not achieved, the results in the study has used for decision support aids to predict the health seeking behavior of households. Finally, based on the above results the researcher concluded that data mining applications were used for the prediction of the household health seeking behavior.

2.11. Methods and Techniques used for outbreak surveillance data mining applications

There are several data mining techniques which are applied in medical data to extract unseen knowledge or patterns. Some of the techniques used here in this research are predictive or classification and descriptive data mining techniques. Classification algorithms like decision tree (DT) and Naïve Bayes(NB) were used to forecast the future outbreak occurrences of epidemic typhus disease outbreak cases and descriptive data mining modeling like association rule mining specifically Apriori algorithms to show the co-occurrence of diseases outbreak cases and the disease segmentation using clustering analysis method specifically with Simple-Kmeans algorithm. As per the research objectives the techniques are implemented accordingly.

2.11.1. Simple K-Means clustering algorithm

In simple K-Means there are a variety of partition-based clustering algorithms available with the quickest and most simple algorithms for data clustering techniques **[44].** This algorithm partition objects in a datasets in to fixed number of k disjoint subset. Here, for each cluster, the partitioning algorithm maximizes the homogeneity within the cluster by minimizing the square error. The simple K-means algorithm has a rich and diverse history as it was independently discovered in different scientific fields **[32].** It was also added that, even though K-means was first proposed over 50 years ago, it is still one of the most widely used algorithms for clustering because of its ease of implementation, simplicity, efficiency, and empirical success which are the main reasons for its popularity. First of all, we determine number of cluster K and we assume the centroid or center of these clusters. We can take any random objects as the initial centroids or the first K objects can also serve as the initial centroids. Then the K means algorithm will do the three steps below until convergence. Iterate until the objects are *stable* (= no object move from the group): those activities which are done to implement clustering simple K-Means are as follows:

1. Determine the centroid coordinate.
2. Determine the distance of each object to the centroids

3. Group the object based on minimum distance (find the closest centroid)

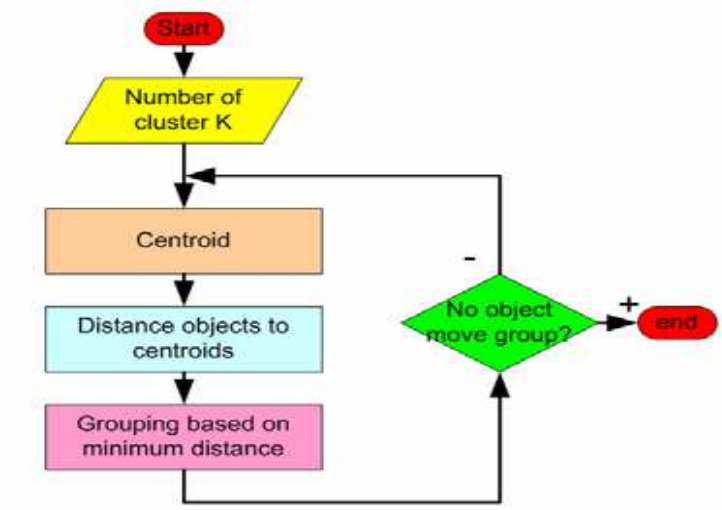


Figure3: Simple K-Means clustering algorithm implementation steps

The implementation of the algorithm can be done as follows: Given that $\boldsymbol{X = \{x_i, i = 1... n\}}$ be the set of *n d*-dimensional points to be clustered into a set of *K* clusters, $\mathbf{C= \{ C_k, k=1,...,K \}}$. The simple K-means algorithm finds a partition such that the squared error between the empirical mean of a cluster and the points in the cluster is minimized. Let μ_k be the mean of the cluster C_k The squared root between μ_k and the point in cluster c_k is defined as:

$$J(C_k) - \sum_{X_i \varepsilon C_k} \| X_i - \mu_k \|2$$

As shown in **[18],** mostly the simple-KMeans algorithm is executed by the use of Euclidean Distance which is the most common use of distance in clustering. In most cases when people said about distance, they will refer to Euclidean distance. Euclidean distance or simply 'distance' examines the root of square differences between coordinates of a pair of objects.

$$\mathbf{d_{ij}} = \sqrt{\sum_{k=1}^{n}(Xik - Xjk)2}$$

Euclidean distance is a special case of Minkowski distance with λ = 2. Furthermore, the algorithm automatically normalizes numerical attributes when doing distance computations. The WEKA Simple K-Means algorithm uses Euclidean distance measure to compute distances between instances and clusters.

As stated in **[32],** simple determinations of the number of clusters have been one of the most difficult problems in data clustering. Most methods for automatically determining the number

of clusters cast it into the problem of model selection. Most of the time clustering algorithms was run with different values of number of clusters *K*; the best value of *K* was then chosen based on a predefined criterion. In addition to that, simple K-means algorithm requires three user-specified parameters: number of clusters K, cluster initialization, and distance metric. The most critical choice is how to define K. While no perfect mathematical criterion exists, a number of heuristics and discussion therein were available for choosing K. Typically, K-means was run independently for different values of K and the partition that appeared the most meaningful to the domain expert was selected**.**

As we know that different initializations can lead to different final clustering because K-means only converges to local minima. One way to overcome the local minima was to run the K-means algorithm, for a given K, with multiple different initial partitions and choose the partition with the smallest squared error. K-means is typically used with the Euclidean metric as mentioned here for computing the distance between points and cluster centers. In general the goal of K-means clustering is to minimize the sum of the squared error from the overall K clusters, That means, minimizing the sum of square errors will makes the intra-cluster similarity high and inter-cluster similarity low. One way to overcome the local minima is to run the K-means algorithm, for a given K as mentioned above, with multiple different initial partitions and choose the partition with the smallest squared error.

2.11.2. Association rule mining with Apriori algorithms

In this research, the researcher used association rule mining Apriori algorithm which is becoming the standard approach to mine association rules. It generates frequent item sets in the same way. It is because that the idea behind in association rules mining is that an itemsets (patterns). An itemsets is called frequent when its support is above a user's predefined minimum threshold and confidence with the user specified minimum threshold.

2.11.2.1. Apriori Algorithm

Apriori Association rule mining is used to mine the frequent patterns in database. Support & confidence are the normal method used to measure the quality of Apriori association rule mining **[45].**

According to **[21],** the algorithm Apriori computes the frequent itemsets in the database through several iterations. Iterations i computes all frequent i-itemsets (itemsets with i elements). In this algorithm, every iteration has two steps: the first is candidate generation and the second is candidate counting and selection. In the first phase of the first iteration, the generated set of candidate itemsets contains all i-itemsets (i.e, all items in the database). In the counting phase, the algorithm counts their support searching again through the whole database. Finally, only i-itemsets (items) with a minimum support s which is above the required threshold will be selected as frequent. Thus, after the first iteration, all frequent i-itemsets will be known.

As mentioned in **[45],** using the rule below one can facilitate to generate the association rule from frequent itemsets:

- For each frequent itemset L, find all nonempty subset of L
- For each nonempty subset of L, write the association rule S ➔ (L-S) if support count of L/support count of S >= Minimum Confidence.

As mentioned before the two important measures of association rule mining interestingness and usefulness are the support and confidence of association rules. For instance, given a transaction with an itemset i that contain items A and C the confidence of the rule is calculated as:

Conf (A → C) = sup (A ∪ C) /sup (A) which can show the strength of the rule and the support value is also calculated as follows:

Sup (A➔C) = P (AnC) / P (AuC) = the ratio of the number of transaction which contains both A and C to the total number of transactions in the datasets. Support shows the presence of an itemsets in the transaction and measure the probability of the existence of an item frequently in the transaction. Here the researcher has to define the minimum support and the minimum confidence. The higher the confidence the better the rule with a user specified minimum support threshold.

2.12.Classification with Naïve Bayes and Decision Tree classifier

2.12.1. Naïve Bayes Classifier

Naive Bayes is one of the simplest density estimation methods from which we can form one of the standard classification methods in machine learning applications **[47].** It represents a supervised learning method as well as a statistical method for classification.

As stated in **[48],** a Naive Bayes classifier is a simple probabilistic classifier based on applying Bayes' theorem with strong independence assumptions. The traditional naive Bayes classifier is still widely used as a popular learning algorithm for data mining applications due to its simplicity. However, we should be aware that many real-world problems are represented by data which often do not satisfy all the assumptions of this technique.

The algorithm assumes an underlying probabilistic model and it allows us to capture uncertainty about the model in a principled way by determining probabilities of the outcomes. This algorithm also provides practical learning algorithms and prior knowledge and observed data can be combined. It calculates explicit probabilities for hypothesis and it is robust to noise in input data. Bayesian reasoning is applied to decision making and inferential statistics that deals with probability inference. For the prediction of future events, it uses the knowledge of prior events. Example: Predicting the occurrence of future outbreak cases from the existing surveillance datasets.

As explained in **[26],** Naïve Bayesian classifier is to be comparable in performance with decision tree. As a result it has also exhibited high accuracy and speed when applied to large databases. Naïve Bayesian classifiers the assumption is that class conditional independency will be developed since the effect of an attribute value on a given class is independent of the values of the other attributes.

2.12.1. Naïve Bayes algorithm

As mentioned above the Naive Bayes classifier was designed for use when features were independent of one another within each class, but it appeared to work well in practice even when that independence assumption was not valid. It classifies data in two steps:

1. Training step: Using the training samples, the method estimates the parameters of a probability distribution, assuming features are conditionally independent given the class.
2. Prediction step: For any unseen test sample, the method computes the posterior probability of that sample belonging to each class. The method then classifies the test sample according to the largest posterior probability.

Parameter estimation for Naive Bayes models uses the method of maximum likelihood. The advantage of these techniques is to model a prediction with the use of a small amount of training data to estimate the parameters. It is also used to calculate the probability of each of the possible classifications, and then choose the classification model with the largest values. The implementation process is as follows: Given a set of "**k"** mutually exclusive and exhaustive classes $\mathbf{c_1, c_2, c_3, \ldots c_k}$ in which they have prior probabilities $\mathbf{p(c_1), p(c_2), p(c_3) \ldots p(c_k)}$ respectively, and the datasets with "m" number of attributes $\mathbf{a_1, a_2, a_3, \ldots a_m}$ and each of the attributes with "n" number of values are $\mathbf{v_1, v_2, v_3, \ldots v_n}$ respectively. Therefore the posterior probability of the class $\mathbf{c_i}$ occurring for the specified attribute can be shown to be proportional to:

$$P(c_i) * p(a_1 = v_1 \text{ and } a_2 = v_2 \text{ and } a_3 = v_3, \ldots a_m = v_m / c_i).$$

Assuming that the attributes are independent, the above expression can be calculated using the product formula as follows:

$$P(c_i) * p(a_1 = v_1 | c_i) * p(a_2 = v_2 | c_i) * p(a_3 = v_3 | c_i) * \ldots * p(a_m = v_m / c_i).$$

Finally the naïve bayes classifier calculates the above product for each value **"i"** from one to **"k"** and the classification with the highest value will be used for prediction activities. Therefore; the prior probability of $\mathbf{C_i}$ in the formula will be combined with the value of "**m**" with the possible conditional probabilities involving a test in the algorithm to the value of the single attribute.

This is formulated as follows:

$$p(ci)\prod_{i=1}^{m} p(aj = vj|class = ci)$$

When a new datasets arrives in the database, measure the probabilities using the most likely class that the Bayes theorem generates as a class. The conditional probability with the Naïve Bayes theorem will be calculated to formulate the probability of a given tuple in the datasets. The best way using Naïve Bayes classification is that unknown variables will be classified from the very beginning by calculating all the prior probabilities and also all the conditional probabilities involving on one attribute, though not all the attributes are required for the classification of a particular attribute.

2.12.2. Decision Trees and Decision Rules

Decision trees and decision rules are data-mining methodologies applied in many real-world applications as a powerful solution to classification problems **[16]**. Here the process is the learning algorithms in which the goal is to create a classification model, known as a *classifier*, which will predict, with the values of its available input attributes, the class for some entity (a given sample). Decision trees and decision rules are typical data-mining techniques that belong to a class of methodologies that give the output in the form of logical models.

Decision tree is an efficient method for producing classifiers from a huge datasets and most widely used methods of learning that construct trees from a set of input-output samples. A typical decision-tree learning system adopts a top-down strategy and it consists of *nodes* where attributes are testes that are the outgoing *branches* of a node correspond to all the possible outcomes of the test at the node.

As shown below in the figure, all samples with feature values **X = yes** and Y=A belongs to class1, Y= B belong to Class2, and Y= C belongs to Class3 while the samples with values **X = no** belong to Class4, whatever the value for feature Y. In general, any restriction on a particular tree representation can significantly restrict the functional form and thus the approximation power of the model. The greedy search methods, which involve growing and

pruning decision-tree structures, are typically employed in J48 algorithms to explore the exponential space of possible models.

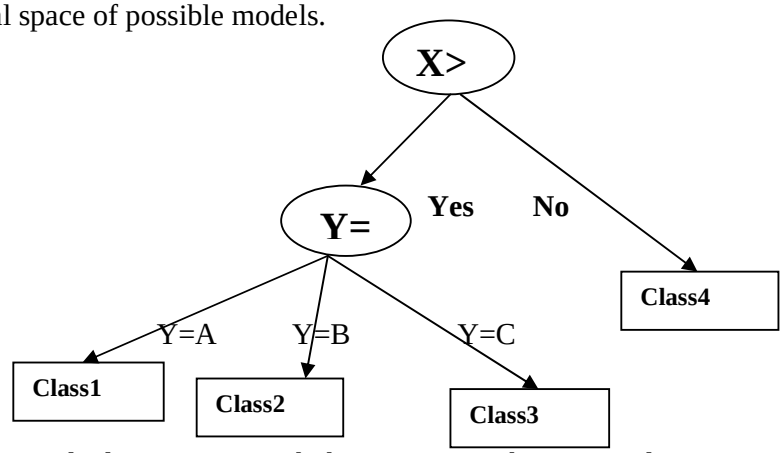


Figure4: A simple decision tree with the tests on attributes X and Y

As we can see from the figure above every path to the leaf in the decision tree represents a classification rule. Attribute selection in J48 and C4.5 algorithms are based on minimizing an information entropy measure applied at a node.

The approach based on information theory insists on minimizing the number of tests that will allow a sample to classify in a database. Since the attribute-selection part of J48 is based on the assumption that the complexity of the decision tree is strongly related to the amount of information conveyed by the value of the given attribute. Since the information gain is used to measure the best attribute selection, the attribute that minimizes the information in the resulting sub tree to classify the sample is needed **[16].**

As mentioned above decision tree algorithm used for this research is J48 which is the implementation of C4.5 algorithm. The C4.5 algorithm recursively visits each decision node, selecting the optimal split, until no further splits are possible. For categorical attributes, the C4.5 algorithm by default produces a separate branch for each value of the categorical attribute. This may result in more "bushiness" than desired, since some values may have low frequency or may naturally be associated with other values. The C4.5 algorithm uses the concept of *information gain* or *entropy reduction* to select the optimal split of the attributes.
Suppose that we have a variable *X* whose *k* possible values have probabilities $p_1, p_2, \ldots, p_j$. Here the question is, what is the smallest number of bits, on average per symbol, needed to transmit a stream of symbols representing the values of *X* observed? The answer is called the *entropy of X* and is defined as:

$$H(X) = -\sum_{j} p_j \log_2(p_j)$$

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:

$$H(X) = -\sum_{j} p_j \log_2(p_j)$$

Here what we can observe is C4.5 uses this concept of entropy in the following ways. Suppose that we have a candidate split *S,* which partitions the training data set *T* into several subsets, *T*1, *T*2, . . . ,*Tk* . As a result the mean information requirement can then be calculated as the weighted sum of the entropies for the individual subsets, as follows:

$$H_S(T) = \sum_{i=1}^{k} P_i H_S(T_i)$$

Where *Pi* represents the proportion of records in subset *i*. We may then define our *information gain* to be gain(*S*) = *H* (*T*) – *Hs* (*T*), that is, the increase in information produced by partitioning the training data *T* according to this candidate split *S.* At each decision node, C4.5 chooses the optimal split to be the split that has the greatest information gain, gain(*S*). As we have seen above in the C4.5 algorithm about information gain the J48 algorithm uses the same establishment for decision tree construction **[33]**.

2.12.2.1. Minimizing Model Over-fitting on decision tree

A classification algorithm is said to *overfit* to the training data if it generates a decision tree (or any other representation of the data) that depends too much on irrelevant features of the training instances, with the result that it performs well on the training data but relatively poorly on unseen instances **[49].** Realistically, overfitting will always occur to a greater or lesser extent simply because the training set does not contain all possible instances. Therefore; it only becomes a problem when the classification accuracy on unseen instances is significantly downgraded. Therefore; we always need to be aware of the possibility of significant overfitting and to seek ways of reducing it. Let us consider a typical rule such as IF *a* = 1 and *b* = yes and *z* = red THEN class = OK Adding an additional term to this rule will

specialize it, for example the augmented rule IF *a* = 1 and *b* = yes and *z* = red and *k* = green THEN class = OK will normally refer to fewer instances than the original form of the rule (possibly the same number, but certainly no more). In contrast, removing a term from the original rule will *generalize* it, for example the depleted rule IF *a* = 1 and *b* = yes THEN class = OK will normally refer to more instances than the original form of the rule (possibly the same number, but certainly no fewer). The principal problem with Top-Down Induction of Decision Trees and other algorithms for generating classification rules is that of over-fitting.

When a decision tree is built, many of the branches will reflect anomalies in the training data due to noise or outliers. Tree pruning methods address this problem of *over-fitting* the data. Such methods typically use statistical measures to remove the least reliable branches of the decision tree **[26].**

2.12.2.2. Classification Metrics for predictive model development

After a model has been developed for outbreak prediction using the PHEM datasets, it has to be checked for its performance for the data that will be collected in the surveillance system in the future. The future outbreak data which is not seen today will be predicted accurately with a certain level of accuracy. As mentioned before the researcher has used two types of classification or prediction model to forecast the future outbreak. The first one is decision tree and the second is Naive Bayes Classifier. The performance of the predictive model will be evaluated by the use of confusion matrix. As described in **[50], t**he classification model is evaluated by the conventional machine learning metrics such as Precision, Recall, F-Measure, TP Rate, FP Rate, and ROC Area and soon.

As stated in **[26],** the *confusion matrix* is a useful tool for analyzing how well your classifier can recognize tuples of different classes. If there is *m* classes in the data set, a table of at least size *m* by *m confusion matrix will be formulated.* 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*. As stated in **[51],** the confusion matrix is more commonly named contingency table. It states that the number of correctly classified instances is the sum of diagonals in the matrix; all others are incorrectly classified.

The model has to have a classifier which is having good accuracy; ideally most of the tuples would be represented along the diagonal of the confusion matrix, from entry *CM*1, 1 to entry *CMm, m,* with the rest of the entries being close to zero. The following table shows a sample two by two confusion matrix.

Confusion Matrix		**True Classes**	
		Positive	**Negative**
Predicted Classes	**Positive**	**True Positive Count(TP)**	**False Positive Counts(FP)**
	Negative	**False Negative Counts(FN)**	**True Negative Counts(TN)**

Table1. Confusion matrix for two classes

As we have seen from the table above the confusion matrix with a certain class A is described as follows:

True Positive (TP): Total percentage of members (in our case flows) classified as Class A, which belongs to Class A.

False Positive (FP): Total percentage of members of Class A, which does not belong to Class A.

False Negative (FN): Total percentage of members of Class A, incorrectly classified as not belonging to Class A.

True Negative (TN): Total percentage of members which do not belong to Class A and are classified as not a part of Class A. It can also be given as (100% - FP). These metrics will help us form the overall accuracy, precision and recall of the classification model. We define these terms as follows:

Precision is given as the ratio of True Positives to True and False Positives. This value helps us understand the classification capability to identify objects correctly.

Precision = TP / (TP + FP)

Recall is given as the ratio of True Positives to True and False Negatives. This value helps us understand the classification capability to determine misclassified members are something else.

Recall = TP/ (TP + FN)

Overall Accuracy is given as the ratio of all True Positives plus all True Negatives to the total Positives and Negatives.

Overall Accuracy = TP+TN / (TP+TN+FP+FN)

F-Measure is simply a combined measure for precision and recall. These measures are useful for comparing classifiers

F-Measure = 2*Precision*Recall / (Precision + Recall)

Chapter Three

3. Public Health Emergency Management

3.1. Public health surveillance and response system

The public health system is continually challenged by recurrent and unexpected disease outbreaks and is facing the challenges of managing health consequences of natural and human made disasters, emergencies, crisis and conflicts. As mentioned in **[52],** these problems continue to disrupt the health care system, and successful detection and response to these challenges is becoming increasingly complicated. It is also added that the rapidly expanding global economy, the convergence of people in large urban areas, the ease with which people and goods travel around the world, emergence of new infectious agents, the wide distribution of manufactured foods and the changing nature of our environment are some of the factors challenging the public health system to quickly recognize and respond to widely dispersed public health events.

As stated in **[53],** the goals of surveillance often differ at the various administrative levels of the public health system. Surveillance data are used to allocate resources and evaluate the impact of control and prevention strategies and programs at all levels. However, at the local level the use of surveillance to trigger basic public health investigations and implement specific control activities predominates for infectious diseases and environmental hazards. In contrast, monitoring for trends, measuring the effectiveness of specific interventions, and conducting more complicated analysis to elucidate risk factors predominate at the national level. At the local level analytic capacity is usually much more limited than at the national level. At the state level public health agencies typically share both perspectives.

As defined in **[54],** the term ‘outbreak’ has been defined as “an epidemic limited to a localized increase in the incidence of a disease, e,g, in a village, town, or closed institution” and an epidemic is defined as “the occurrence in a community or region of cases of an illness, specific health-related behavior, or other health-related events clearly in excess of normal expectancy.” Sometimes we can see an overlap between the term ‘outbreak’, as defined here, and the term ‘cluster’. Both can be used to describe an aggregation of diseases

or events grouped in space, time, or both. The main distinction is that clusters generally refer to groupings of diseases or events that are relatively uncommon, whereas this is not a condition for defining an outbreak.

WHO asserts that 'effective communicable disease control relies on effective response systems and effective response systems rely on effective disease surveillance'. As defined in **[52] [55],** Health surveillance can be described as "the tracking and forecasting of any health event or health determinant through the continuous collection of high quality data, the integration, analysis and interpretation of data results into surveillance products and the dissemination of surveillance products to those who need to know [to address] a specific public health purpose or policy objective. It is also defined as "Information for Action**" [52].** In general, Health surveillance is an essential component of evidence-based decision making practices in public health system.

As discussed in **[55],** Humanitarian emergencies, whether they are caused by conflict, natural disaster or food insecurity, are frequently characterized by the displacement of large numbers of people. Those affected are often settled in temporary locations with high population densities, inadequate food security and shelter, unsafe water, poor sanitation and infrastructure that has been compromised or destroyed. These circumstances can increase the risk of transmission of communicable diseases and other conditions, and can thus lead to increased mortality (death). Diseases that have a tendency to become epidemic (referred to as epidemic-prone diseases), in particular, can be a major cause of morbidity (disease) and mortality (death) during emergencies. Rapid detection and prompt response to epidemics among the affected population is a key priority during humanitarian crises.

3.2. Overview of PHEM sector in Ethiopia

The general definition of Public health emergency management is the process of anticipating, preventing, preparing for, detecting, responding to, controlling and recovering from consequences of public health threats in order that health and economic impacts are minimised **[56]**.

The International Health Regulations (2005), (also referred to as the IHR or the Regulations) entered into force on 15 June 2007 for public health emergence management activities. The Regulations are binding on all 46 WHO Member States in the African Region as all have agreed in 2005 to be bound by these Regulations. The IHR have a broad scope and the Regulations apply to "any emergency with international repercussions for health, including outbreaks of emerging and reemerging epidemic-prone diseases, outbreaks of food-borne disease, natural disasters, and chemical or radio-nuclear events, whether accidental or caused deliberately''. The IHR take into account lessons learnt in past decades in detecting and responding to disease outbreaks. The IHR aim at protecting global health security while avoiding unnecessary interference with international travel and trade and the purpose and scope is to "prevent, protect against, control and provide a public health response to the international spread of disease in ways that are commensurate with and restricted to public health risks, and which avoid unnecessary interference with international traffic and trade''. In the Ethiopian context IHR would be implemented within the frame work of newly established PHEM Sector.

The importance of disease surveillance in guiding health planning and interventions was recognized for a long time. In Ethiopia, "Quarantine "rules were proclaimed in 1947 with emphasis on surveillance. Another legal notice was issued in 1951, binding all public health practitioners in the country to report communicable diseases. The PHEM under the sponsorship of EHNRI has submitted legislation to Council of Ministers for approval. The "Public Health Proclamation No.200/2000" orders that any individual who knows the existence of communicable diseases in his/her vicinity to report immediately to the nearest health institution and the institution receiving the report to take the necessary measures and report to the appropriate health authority.

In 1948 an anti-epidemic service was established to deal with prevention and control of communicable diseases. In 1951, 35 priority diseases were selected and classified into first

and second class to be notified to MoH, immediately or weekly as necessary. In the mid 1970`s the anti-epidemic unit was converted to epidemic control and surveillance unit under communicable diseases control division and vertical programs were conducting their own disease specific surveillance.

After the health system reform in 1994 nineteen diseases were selected for surveillance. Though, Ethiopia has started implementing public health surveillance strategy since 2000, there were impediments in the implementation of the public health surveillance strategy. These include non-recording, failure to report diseases of epidemic potential in time, incomplete and late reporting of notifiable diseases, inadequate data analysis especially at peripheral level, failure to use available information to follow disease trends, inadequate laboratory involvement in case detection, poor feedback to health workers and communities, duplication of resources meant for surveillance and response, under utilization of surveillance information in decision-making, shortage of technical personnel, high turn-over of technical staff, lack of motivation, and under developed communication systems.

Recently the Federal Ministry of Health has finalized BPR process for health related activities. One of the core processes is PHEM which becomes operational since mid 2009. PHEM core process is designed, as much as possible, to reflect a multi hazard view giving due emphasis to disease outbreaks and nutritional emergencies. It is envisaged that PHEM core process will be responsible to detect unusual health events timely and institute appropriate response measures promptly.

3.3. Ethiopian Public Health surveillance and Response System

As mentioned above nowadays, investigations of diseases are more complex in nature than they were in the past because of a variety of new pathogens, risk factors and outbreaks, which cross jurisdictions and national boundaries, often raising political and economic concerns. Periodic infectious disease outbreaks and recurrent natural disasters serve to remind the importance of the public health system which encompasses the government and private sector, academia, NGOs, associations and development partners as a whole.

However, there is tremendous task to be addressed in order to narrow the gaps between these actors in order to maintain adequate responses to emerging diseases and health consequences of natural and human made disasters **[52].**

Cognizant of the aforementioned factors, the health sector in Ethiopia has developed a fully integrated, adaptable, all-hazards national preparedness and response system within PHEM core process. This Core process is comprised of four sub processes which are Public Health Emergency Preparedness, Early Warning, Response and Recovery. It is believed that the PHEM core process will provide the health sector with a system that is effective and efficient; and its implementation shall be in an accountability basis **[52].**

As mentioned in **[55],** the process of surveillance activities for different diseases involves similar functions and very often uses the same structures, processes and personnel. A multi-disease approach to disease surveillance aims at establishing well co-ordinate action-oriented surveillance systems that seek opportunities for integration of core and support surveillance functions when appropriate, maximize synergies, take advantage of new tools, build on existing resources, and benefit from successful initiatives**.**

A functional disease surveillance system is essential for defining problems and taking action. Proper understanding and use of this essential epidemiological tool (public health surveillance) would help health workers at the Woreda and health units to set priorities, plan interventions, mobilize and allocate resources, detect epidemics early, initiate prompt response to epidemics, and to evaluate and monitor health interventions. It also helps to assess long term disease trends for better implementations **[52].**

As mentioned above, Ethiopia is facing the challenges of managing the health consequences from natural and human made disasters, emergencies, crises and conflicts. These problems continue to disrupt the health care system. Successful detection and response to these challenges is becoming increasingly complicated. Ethiopia's public health infrastructure requires adequate attention and allocation of resources to be sufficiently prepared, enable early detection, and to respond and recover rapidly from the impacts of these challenges **[58]**.

Health departments investigate suspected outbreaks for a variety of reasons. These include the need to institute control and prevention measures; the opportunity for research and training; program considerations; and public relations, political concerns, and legal

obligations **[59]**. Additionally, most outbreaks come to the attention of health departments in one of two ways. One way is by regular analysis of surveillance data. That means, unusual rises or patterns of disease occurrence can be detected promptly if surveillance data collection and analysis are timely. The second, and probably more common, way is through calls from a health care provider or citizen who knows of "several cases." For example, a member of the public may report three infants born with birth defects within a 1-month period in the same community. This aggregation of cases *seems* to be unusual, but frequently the public does not know the denominator--e.g., the total number of births--or the expected incidence of birth defects.

For the most part, public health surveillance is still very much disease or program specific. It works how data are gathered, aggregated, analyzed, and summarized may be quite different for different types of diseases cases **[60]**. It is also different for influenza than for enteric illnesses, and different again for sexually transmitted diseases. In addition to that many disease specific surveillance systems still rely on the manual collection of data from the providers, mail and fax (and now email) to deliver data to a centralized location, manual execution of basic statistical procedures to summarize the data, and mail and fax (and email) to distribute the results.

Chapter Four

4.1. Business understanding and data preprocessing

4.1.1. Business understanding

The MOH of Ethiopia and its development partners assessed the country's surveillance system and used the results to adapt a five-year national plan**.** As mentioned in **[6],** In Ethiopia, infectious and communicable diseases account for about 60-80% of the health problems in the country. Ethiopia developed a five-year strategic plan and started implementing the IDSR consisting of 23 priority diseases that are epidemic-prone, have public health importance and are targeted for eradication and elimination. “Later three diseases were reduced and 20 of them were only followed up in the sector”. The systematic evaluation of the first five years of implementation indicated that the country has successfully expanded its surveillance system up to the health centre and hospital level, improving the early detection, reporting, epidemic preparedness and response activities**.**

As mentioned before the objectives of communicable disease outbreak surveillance and response are usually take care of patients, prevent further cases, end the disease outbreak cases quickly and prevent its recurrence**.** As defined in the literature part of this paper, Surveillance is the systematic and continuous collection, analysis and, interpretation of data related to disease occurrence and giving feedback to those who need it to take action, it is important to have quality and consistence datasets.

As stated before the PHEM sub processes of the MoH of Ethiopia are designed to ensure rapid detection of any public health threats, preparedness related to logistic and fund administration, and prompt response to and recovery from various public health emergencies, which range from recurrent epidemics, emerging infections, and nutritional emergencies to chemical spills and bioterrorism. In the sector there are 20 diseases which are epidemic prone and reported in timely and weekly bases. But the decision making process was not supported with machine learning data mining tools and techniques to enhance the system functionalities by increasing its effective and efficient. Hence the researcher wanted to analyze the system with these applications whether to see its function was good or not and to develop a better

model on the existing datasets collected in the sector. These datasets were consisting of the following disease outbreak cases shown in the table below:

Surveillance data set reporting	Epidemic Disease lists
Mandatory notification(Daily reportable diseases)	**1. AFP/Polio**
	2. AHI
	3. Cholera(AWD)
	4. Guinea Worm(Dracunculiasis)
	5. NNT(Neonatal Tetanus)
	6. RVF
	7. SARS
	8. Small Pox
	9. Plague
	10. Yellow fever
	11. Anthrax
	12. VHF(Viral hemorrhagic fever)
Routine surveillance(Weekly reportable diseases)	**All the above+**
	13. Meningitis
	14. Malaria
	15. Measles
	16. Typhoid fever
	17. Typhus
	18. Relapsing fever
	19. Rabies
	20. Malnutrition

Table2. Epidemic prone reportable disease lists

The researcher used weekly reportable (routine surveillance) disease outbreak cases to test with data mining techniques for description and prediction data mining modeldevelopment techniques. Finally, the EMoH stated that the PHEM core process activities were to be implemented by appropriately trained and capable professionals for quality data management activities and accurate decision making processes for better decision making activities. But the data sets were still in a poor quality and have missing variables which were important for data mining applications. The PHEM sector used traditional data collection mechanisms and the data analysis tools were also statistical for the purpose of description capability of disease occurrence on the bases of time, place and person.

4.1.2. Data Understanding and Preprocessing

As stated in the methodology part of this research, the disease outbreak surveillance data was incorporated and used from PHEM. In the surveillance datasets, data had been collected in timely (daily) and weekly bases form different parts or reporting sites of the country. Here all the regions, zones, districts and respected reporting sites were expected to send the data as per the requirements of the sector and the MoH on timely and weekly bases. Since disease outbreaks cases were investigated in person, time and place setting, the datasets were consisting of attributes like *Regions*, *Zones*, and *Woredas* of the country for place setting, *Year*, *Month* and *Epidemic week* for time setting, *number of cases* and *deaths* of the population in the country for person and finally *number of NGO's, governmental organization* and *health facilities* for data reporting activities.

The researcher used only weekly collected disease outbreak cases (routine surveillance data) from the surveillance data store. As per the research design methodology, datasets from 2004-2012 G.C were incorporated. These data were a collection of records from the IDSR datasets which was collected from 2004-2008G.C. and having 4715 number of records and 42 number of attributes. Whereas the data in the current PHEM core process dataset were having a collection of 14509 number of records and 55 number of attributes collected from 2009-2012G.C.

These datasets were analyzed in a separate manner and later they were integrated or merged together for some model development as a comparison of experimental results from the two isolated datasets in association rule mining to show the correlation of disease cases from the three datasets individually and altogether. For the purpose of prediction activities of disease outbreak cases, only the new PHEM sector datasets were used and for the purpose of data clustering or disease segmentation the researcher used the integrated (merged) datasets of the two systems, because of time and budget constraints as mentioned in the limitation part of this research.

Finally, the data preprocessing activities, such as data cleansing, best feature generation (attribute selection), data transformation, and data reduction were implemented step wise as follows.

4.1.3. Data Preparation

As shown on the methodology parts of this research, data preparation is the firsts step in data mining applications after the data collection phase has been completed. Therefore, this process was used to get quality data. It was because; data quality is a prerequisite and crucial part of data mining techniques. Therefore high quality data was required to implement data mining algorithms for better results. Prior to data modeling, the data needed to be prepared for better quality in the required format and level. According to **[59],** the objective at this stage was two-fold; to obtain data prepared in the form required by the data mining algorithms and to expose as much information as possible for data modeling.

Data preprocessing activities are the most important parts of the process for data quality. In general there are four types of data preprocessing techniques to transform the data and improve its quality to suit the requirements of the learning algorithms. The first part was data cleansing. It was applied to the data by filling in missing values, removing or smoothing noise, identifying or removing outliers and inconsistencies in the data. The second part was data integration. It was used to merge the data from multiple sources into coherent data stores. The third is data transformation. It was used to transform and consolidate data into forms appropriate for mining activities, and the forth is data reduction. It was used to reduce the data size by aggregation, elimination of redundant features, or clusters.

As described above after the data have been collected from the respected data stores, the process of data cleaning and preparation was followed. This process was highly dependent on the specific data mining algorithm and tools chosen for the data mining activities. But all data mining applications need the preprocessed and cleaned data for better prediction and description purpose. The researcher has attempted to prepare the data accordingly to the requirements of the selected data mining software, Weka and selected data mining algorithm such as Decision Tree and Naïve Bayes classification algorithm, Apriori algorithm for association rule mining and simple-K-Means algorithm for data clustering. It was because; Weka is multi-functional data mining software which is used for the process of classification, clustering and association rule mining applications.

Hence, the major data mining functionalities incorporated in the software are data preprocessing, classification, association, clustering, attribute selection and visualizing. To

use the surveillance datasets for data mining applications from PHEM sector to a specific algorithm the researcher checked the quality of data. Since the selected data mining tool, Weka, does not allow any inconsistency among attribute values and their definition, considerable amount of time was spent in checking this consistency.

In IDSR dataset attribute values were written in small and capital letter for the same data element for example in the region attribute like *AMHARA, OROMIA, TIGRAY* in some records and *Amhara, Oromia, Tigray* in other records and sometimes miss spelt of the word itself like *Somali* and *Somalia*. There were also cases like *BENISHANGUL GUMUZ* and *B-GUMUZ* and so on which was the name for one region in the datasets with different naming styles, therefore the researcher with the discussion of the domain experts had given the same naming style for the values of the attribute to correct the inconsistency among data elements throughout the datasets. For instance *B-Gumuz* was assigned for the whole *BENISHANGUL GUMUZ* naming styles.

The Algorithms chosen to generate patterns from the selected dataset, predict the future occurrence and to show disease outbreak segmentation and the density of the outbreak cases in a certain location with time variant were implemented in the research. The researcher determined to remove the missing data with the threshold of 10% missing elements and eliminated all the attributes from the datasets.

As mentioned above the missing data elements were dropped if the percentage of missing elements were more than 10% as shown in Appindex1 and Appindex2 for IDSR and PHEM datasets attribute description and missing values calculation respectively. Therefore, attributes such as *Woreda, HealthF, NumReportingSites, NumSitesReportedTimely, NumSites ReportedLate, UncomplicatedLabMalOutptcases, MeningitisInpt, Meningitisdeaths, DateSent Report, and DateRecReport* have been removed from the datasets since the percentage in the missing values were higher than the user specified threshold. Additionally, some attributes have been integrated together and changed to one attribute values because of the objectives of the research and the importance of the attributes.

For instance attributes *like MalL5yOutPtCases, MalL5yInptCases, MalGE5yOutPtCases, Mal GE5yInPtCases, MalGE5yInPtCases, MalL5ySevAnemiaInPtCases, MalG5SevAnemiaInptCa ses, MalariaInPregnancyOutPtCases, MalariaInPregnancyInptCases, MalL5yLabOutPtCase s, UncomplicatedMalPVivax, UncomplicatedLabMalOutptcases, and uncomplicatedMalG5L*

abConfPVivax were merged together into one attribute value called malaria occurrence which was later represented with the binary values of 0 or 1 for clustering and association rule mining and Yes or No values for classification algorithm for the occurrence and non-occurrence of the disease cases.

The indication of 0 or "No" meant the non-occurrence of disease outbreak cases and 1 or "Yes" meant the occurrence of the outbreak diseases cases and the other remaining malaria related
deaths like *MalL5yInptDeaths, MalGE5yInPtDeaths, MalL5ySevAnemiaInptDeaths, MalG5y WithSevAnemiaInptDeaths, MalariaInPregnancyInptDeaths* had no use in model preparation for diseases occurrence or future incidences of disease cases. The attributes for deaths of all the disease cases were not included in the model development to run on the algorithm since it was not the cause of an outbreak rather they were only the consequence of the disease occurrence.

4.1.4. Data Quality Measures

The quality assurance mechanisms of the data start at the early stage of the disease outbreak surveillance data reporting and collection activities. It implies the timeliness and completeness of data in the data store as well as the inclusion and exclusion of attributes important to disease outbreak surveillance system. The PHEM data center registered only the data that the sector staffs required for statistical descriptive analysis as mentioned in the problem. Therefore, the problem of getting quality data decreases as the efficiency and effectiveness of the reporting mechanism or system decreases as well.

As stated in **[62],** it was also important to classify the disease cases by *age* group and *sex* category to take measurable action accordingly. For a better predictive model preparation, disease outbreak cases datasets have to include the determinants of outbreak cases in place and time setting. Some other determinant factors like *environmental condition(Temperature, relative humidity, Latitude, rainfall, drought), social condition (income statues, hygiene system, settlement condition, population density, nearness of health center, urban or rural setting, immune system of the community, vaccination used), disease condition (specimen collection, result confirmation, natural history of the disease description, way of*

transmission) and the existence of other natural as well as manmade hazards(forest burning, war, flooding, earth quake, and so on) were missing from the dataset which had a greater effect for disease occurrences.

As per the discussion with the domain experts, the problem was the uniformity of data reporting formats, activities and the burden of data for disease outbreak cases which were reported from different locations of the country. Since the most important thing in the surveillance data quality is the best use of data collection tools, use of appropriate reporting formats, integration of data elements from different reporting sources and data formats, use of appropriate transformation mechanisms to prepare data warehouse or data marts, collecting the data on timely bases, filling the data as much as possible or reducing the missing values, removing noise in the data, reducing inconsistencies, and so on. But the surveillance data store lacks most of the above features, which were important for the best use of data mining techniques.

4.1.5. Data transformation

Data transformation is the process of converting the data in to appropriate format in which data mining algorithms are implemented correctly. Data transformation started from the conversion of data file format for better data mining applications. For Apriori algorithm in association rules mining the data should be in nominal format and the class of the decision tree algorithm should also be categorical. Since data mining application for association rule mining application the attribute values of numeric data types needed be discretized or categorized in nominal data type before the algorithm is applied to it. That was why the researcher had converted the attributes of diseases outbreak cases from the number of total people infected cases in to non-occurrence for zero number of people infected and occurrence for none zero infected persons with the specified outbreak cases. Additionally for the purpose of disease outbreak classification there were also a need of categorization in to occurrences and non occurrences of disease case.

Data mining have different types of tools and algorithms which were mentioned in chapter1 and Chapter2 of this research work, the researcher used WEKA3.6 and WEKA3.7 tools to analyze the data after it had been converted in WEKA data mining tools' acceptable format.

The input of the data into data mining applications proved to be simple with the conversion of an Excel spread sheet datasets into a CSV file format and then an ARFF file format.

Since the surveillance data was stored in Microsoft excel worksheet, the researcher converted or transformed into CSV text format or ARFF file format to be processed accordingly with appropriate data mining algorithms. As mentioned above, researcher had tried to convert from the excel datasets format first CSV file format and then ARFF file format for better processing in Weka data mining tools acceptable format. Another issue for data transformation was the data types needed by each data mining algorithm. For example one of the limitations of association rule mining with Apriori algorism was inability to handle numeric data values. Therefore the researcher was forced to transform the data from numeric attribute values into nominal values by changing number of disease cases to occurrence and non-occurrence.

4.1.6. Data integration, data reduction and feature selection process

As mentioned in the previous parts of this paper, the country disease outbreak surveillance data were collected previously by IDSR system and currently by PHEM core process. The researcher had done analysis on the two datasets separated and merged together for further investigation. Here, the data integration processes were performed by analyzing between IDSR and PHEM surveillance system after the two datasets had been preprocessed and cleaned separately.

After the preprocessing phase of data mining application 4093 number of records from PHEM and 4703 number of records from IDSR system datasets were obtained. Therefore, the total of nine year of 8796 records were prepared for clustering and association rules mining model development as well the four years data of 4093 records from PHEM only were used for epidemic typhus disease classification purposes. But in association rule mining for the sake of result comparison and to see the new patterns in the two datasets and in their integration i.e. in ISDR, PHEM and the combined datasets were investigated separately by the Apriori algorithm.

The researcher tried to analyze the data deeply for data integration process, and considered the following concepts and terms in the preprocessing phase for better integration. These

were the consistency of the naming system in each datasets, the appropriateness of data type used, the number of attributes involved, the number of records needed, the importance of the attributes, the number of missing values or percentages of missing values for each datasets and the redundancy of records and attributes finally making the two datasets to have the same number of attributes and attribute naming system.

According to **[63]**, when processing large databases, one faces two major obstacles such as numerous samples and high dimensionality of the feature space. For example, the documents are represented by several thousands of words; images are composed of millions of pixels, where each word or pixel is here understood as a feature. That currently, processing capabilities are often not able to handle such high dimensional data, mostly due to numerical difficulties in processing, requirements in storage and transmission within a reasonable time. Smaller dimensions are also advantageous when visualizing and analyzing the data **[61]**. Thus, in order to extract desirable information, dimensionality reduction methods were often applied. The overall idea was to determine the coordinate system where the mapping would create low-dimensional compact representation of the data whilst maximizing the information contained within by removing unnecessary variables and records.

Since the datasets were from different data sources with different types of attribute names and number of attributes, the researcher integrated the data with the appropriate transformation mechanisms. For example, in the PHEM surveillance datasets, the attribute *EpidemicWeeks* was found but it was missed from the IDSR system dataset. Therefore; the researcher removed the attributes which were missing from one dataset and existing in the other to make the two datasets consistence. But the problem was how to drop this attribute or variables in such a way that the datasets would not loss its original meaning or information and records. The process was using concept hierarchy.

In the case of concept hierarchy, the lower concept attributes are included in to the higher level concepts. For instance derived attributes such as the sum of four sequence (four weeks are one month) *EpidemicWeek* values were merged together to form a month for a given reporting site. To do this the problem was how could we merge weeks together in a month, for this the researcher planned to count the total number of weeks which were reported from each reporting sites of the country in a month. Since the reports were done on the weekly bases, every reporting sites of the country were expected to send reports al least three times

and appropriately four times in a month. I.e. all the sites were expected to send the report in a weekly base every time.

Another issue was the problem of redundant attributes like *Year* and *Month* which were recorded twice in datasets in Gregorian and Ethiopian colander systems. The researcher decided to use Gregorian calendar format. It was because of the weeks for data collection processes were taken on the bases of WHO epidemic week data reporting format. Some attributes were also merged together when they were representing about one thing like *inpatient, outpatient* cases which were simply telling us the occurrence of the disease cases. Since the objectives of the research, the researcher tried to combine the total cases as the occurrence of disease outbreak cases whatever the numbers of infected persons. All disease cases were formulated to be occurrences when there were existence of disease cases and non-occurrence's for no disease cases.

4.2. Data quality assessment based on timely data reporting activities

As mentioned in the methodology parts of this research specifically on the methods of data collection part of the first chapter, the researcher used routine surveillance data or weekly reportable disease surveillance datasets for analysis and model development purpose. Prediction and description model were also developed from this datasets. The analysis was done according to the format that was used by the sector personnel's and WHO data reporting format for weekly reparable diseases outbreak cases. Here the researcher with the discussion of the domain experts, weekly data reports were converted in a monthly manner with the following week to month conversion processes. The conversion was done to make the IDSR and the PHEM data sets having the same format for analysis purpose.

Since it was mentioned in the data understanding parts of this research work, the IDSR system didn't have attribute values in the ***numberOfWeek*** variable. Therefore the researcher has performed the conversion to eliminate the inconsistencies of data elements (attribute values) from the two datasets. Weekly data reporting quality measures were done according to the criteria set by the domain experts and the researcher for the weekly reporting system evaluation process with respect to time and the sequence of reporting activities:

- If the number of weeks reported in a month were less than three, then that month was said to be under reported
- If the number of weeks reported in a month were three or four, then that month was said to be correctly reported and
- If the number of weeks reported in a month were more than four, then that month was said to be over reported(unexpected report)

The above criteria were set with the standard of the PHEM sector effective data reporting activities. These scenarios were described in the table below for quality data reporting measures with regard to the timely measures. Therefore, weekly Reports of disease outbreak data which were sending from different places of the country were evaluated statistically with the following tabular forms for each month to the PHEM dataset.

No.	Months	Under reported	Under %	Correct reported	Correct %	Over reported	Over %	Total counts of all weeks	Remark
1	January	635	71.43	224	25.20	30	3.37	889	
2	February	630	70.87	231	25.98	28	3.15	889	
3	March	626	70.42	232	26.10	31	3.49	889	
4	April	512	76.65	86	12.87	70	10.48	668	
5	May	501	75.00	158	23.65	9	1.35	668	
6	June	497	74.40	75	11.23	96	14.37	668	
7	July	501	75.00	162	24.25	5	0.75	668	
8	August	507	75.90	156	23.35	5	0.75	668	
9	September	487	72.90	74	11.08	107	16.02	668	
10	October	485	72.60	179	26.80	4	0.60	668	
11	November	501	75.00	169	25.30	3	0.45	673	
12	December	456	68.26	113	16.92	99	14.82	668	
Total number		6338		1859		487		8684	
Average %			73.20		21.06		5.80		

Table 3: The percentage of reporting efficiency in monthly bases of the PHEM dataset

From the table above we could understand that 73.2% of the total reports on weekly reporting bases of the PHEM sector datasets were under reported from the expected report amount. Around 21.06% were correctly reported and 5.80% were reported over expected or the datasets contains unexpected report. Therefore the reporting system of the current PHEM surveillance system was not having timely data collection mechanism for effective and efficient surveillance system. As the discussion with the sector experts, the problem was the absence of enough infrastructures like: web-base data collection tools, fully integrated mobile computing technology, the shortage of skilled human resources, the lack of integration among information technology and public health professionals at every places of the country and soon. The above problem led to the problem of late, incomplete, inconsistence, unexpected (over or under) reports of disease occurrences and its impacts to the sector.

Chapter Five

5.1. Experimental Process for MD

It is known that accessing and processing the real emerging disease outbreak data for experimentation was a very difficult activity due to national data security and confidentiality problems. Here the researcher has discussed about the data with the domain experts for analysis purpose. They allowed giving only the data collected from routine surveillance on weekly bases. Those data consisted of five selected epidemic prone disease cases such as: *Malaria, Typhoid Fever, Epidemic Typhus, Relapsing Fever, and Meningitis.* Therefore, the researcher has decided to use the available disease data on hand for model development purpose.

For this phase, the researcher tried to analyze and experiment the preprocessed nine years of data(from 2004-2012G.C.) with 8796 number of records from the integration of the total datasets for clustering and association rules mining and the currently PHEM system four years data (2009-2012G.C.) of 4093 number of records were used for disease classification purposes. But in association rule mining for the sake of best rule generation purpose both the old and the new datasets and their integration were involved. I.e. in ISDR, PHEM and the combined datasets were investigated separately by the Apriori algorithm. But for classification purpose the researcher would only use the new PHEM surveillance datasets for Epidemic Typhus disease occurrences prediction. It was because of the change in the surveillance system of the country from IDSR to PHEM by the new implementation of BPR in the sector.

For a better prediction model preparation purpose, the researcher used Tenfold (10 fold) cross-validation which was used to test the significance of the decision tree and the naive's Bayes classification algorithm with better performance. The researcher had also tried to show disease segmentation or clustering for disease outbreak cases to group the disease cases in to different regions of the country and with the time period of the nine years datasets. Simple K-means algorithm was used to cluster the prevalence of disease cases in to appropriate level of segmentation. All the preprocessed and prepared 8796 records of the nine years data were involved in this algorithm since the datasets were expected to be grouped according to the time and place setting of the disease occurrence.

5.1.2. Experimentation with Apriori Association-Rule Mining algorithm

As mentioned in the methodology parts of this research Apriori algorithms are efficient way of detecting important patterns from a large datasets. The researcher used weka.gui.GenericObjectEditor window's default upperBoundMinSupport = 1.0, lowerBoundMinSupport = 0.1 and the first Experiment's numbers of rules = 10. To do so, In every iteration the support was decreased by some specified values (delta) till the production of constant or identical Min-Sup, Cycles Performed and Numbers of large itemsets. The number of cycles indicated the time to generate some sorts of rules with the researchers specified Min-Sup values.

As mentioned before in the methods of association rule mining application the researcher used support and confidence rule interestingness measure. They respectively reflect the usefulness and certainty of discovered rules **[63**]. Association rules are considered interesting if they satisfy both a minimum support threshold and a minimum confidence threshold. Such thresholds can be set by researchers or domain experts. Here, the researcher defined the minimum support as 20% and minimum confidence as 90%. The different experiments were performed to uncover interesting correlations between associated disease cases. The minimum support was defined to be small, it was because the researcher with the discussion to the domain experts, there were disease cases which were occurring infrequently like meningitis. Therefore; to include such rarely occurring disease cases to Apriori algorithm association rules mining, the researcher had defined that small value of minimum support and minimum confidence.

As mentioned before, the rules that satisfy both a minimum support threshold (*min sup*) and a minimum confidence threshold (*min conf*) were called strong. By convention, we wrote support and confidence values so as to occur between 0% and 100%, rather than 0 to 1.0 for this research work. It was because the percentage value is more preferable to discuses about the results. The three datasets mentioned in the preprocessing phase would be analyzed below one by one for better rule generation and comparison purpose. Finally the two datasets were merged or integrated together to see whether there were different and additional interesting rules or not. After the generation of rules from the individual datasets of the two surveillance system, the researcher also generated rules from the combined datasets of the whole nine years country surveillance system.

As mentioned above, after the data have been loaded to Weka data mining tools, the researcher used the default values of the *weka.gui.GenericObjectEditor* window components for the first 10 rules to be generated by loading the datasets in the stepwise process. The algorithm outputs 10 rules, which were ranked according to the confidence measure. The confidence was given in parentheses after each rule. In the rules, there were numbers following a rule's antecedent which shows how many instances satisfy the antecedent and the number following the consequence or the conclusion shows how many instances satisfy the entire rule (i.e. the rule's "support"). Practically, it could be very exhaustive to find minimum support and minimum confidence values that would give interesting rules.

To generate interesting rules, the researcher run *Apriori algorithm* several times with different numbers of rules for each iteration with 5 additional rules for the next experimentation to see additional interesting and different patterns. The interval between number of rules were taken as 5 by the researcher because of time constraint to analyze all possible rules which were generated only by adding 1 rule for every experiments. The researcher used the same specified minimum confidence value which is 0.9 or 90% throughout the different experimentations, which was given by the *minMetric* parameter of *weka.gui.GenericObjectEditor* window as default value.

As mentioned in the chapter 2 of this paper, the support of the rule was expressed as a proportion of the total number of instances or of a given diseases outbreak cases to the total number of disease outbreak cases in the datasets. As mentioned before for Apriori algorithm to take place, there are many database scans in the development of interesting rules with different large i-temsets in each experimentation shown below. The result of the rules was analyzed accordingly for each datasets in five experimentation and the experimentation results were described in tabular format for better visualization purpose.

5.1.2.1. Association rules mining by Apriori Algorithm for PHEM datasets

The rules were generated from Weka3.6 and weka7.4 tool with Apriori association rules mining algorithm were shown below. As mentioned in data preparation phase of this research, the PHEM datasets had a total number of 4093 records which were well prepared in the required format. The total numbers of disease cases were five with occurrence and non-occurrence of disease cases. The experimentation results were described in the table below. There were repetitions of rules from the first Experiment to the next Experiment and so on. Therefore; the researcher identified in each iteration and discarded the repeated rules from the next following experiment. Here the researcher stopped developing the rules after the 5^{th} experiment or the generation of 30 rules. It was because of the algorithm generated the same values of Min-Sup, No. of instances for Min-Sup (number of frequent itemsets), No. of Cycles Performed, L(1), L(2) L(3), L(4), and L(5), most of the rules were also repeated and found in the previous experimentation and the required rules were also obtained.

Experiment No.	No. of instance	No. attributes involved	Min-Sup	No.of instances for Min-Sup	Min-Conf	No.Cycles performed	L(1)	L(2)	L(3)	L(4)	L(5)	L(6)	Best rules performed
E1	4093	5	50%	2046	90%	10	5	8	5	-	-	-	10
E2	4093	5	35%	1433	90%	13	5	10	8	2	-	-	15
E3	4093	5	30%	1228	90%	14	7	12	11	4	-	-	20
E4	4093	5	25%	1023	90%	15	8	17	15	6	1	-	25
E5	4093	5	25%	1023	90%	15	8	17	15	6	1	-	30
E6	4093	5	25%	1023	90%	15	8	17	15	6	1	-	35
E7	4093	5	25%	1023	90%	15	8	17	15	6	1	-	40

Table 4: Tabular description of Apriori Algorithm to Association rules generation for PHEM datasets.

Descriptions of terms in the table above: Min-Sup=minimum support, Min-Conf=minimum confidence, L (1) = the first large 1-temsets, L (2) = the second large 2-temsets, L (3) = the third large 3-temsets, L (4) = the forth large 4-temsets and L (5) = the fifth large 5-temsets.

Apriori Algorithm Experimentation Result Analysis

Here the researcher discussed only the selected rules since the purpose of the research to get the more interesting rules and time constraint to discuses all the generated rules.

Experiment 1:

=== Run information ===

Scheme: weka.associations.Apriori -N 10 -T 0 -C 0.9 -D 0.05 -U 1.0 -M 0.1 -S -1.0 -c -1

Relation: PHEM Datasets for prediction activities:

weka.filters.unsupervised.attribute.Remove-R1-5-
weka.filters.unsupervised.attribute.NumericToNominal-Rfirst-last

Instances: 4093

Attributes: 5

MalariaOutbreak

MeningitesOutbreak

TyphoidFeverOutbreak

RelapsingFeverOutbreak

EpidemicTyphusOutbreak

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

Apriori

=======

Minimum support: 0.5 (2046 instances)

Minimum metric <confidence>: 0.9

Number of cycles performed: 10

Generated sets of large itemsets:

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

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

Size of set of large itemsets L(3): 5

Best rules found:

1. TyphoidFeverOutbreak =1 2960 ==> MalariaOutbreak=1 2879 conf:(0.97)

2. MeningitesOutbreak=0 TyphoidFeverOutbreak =1 2306 ==> MalariaOutbreak=1 2232 conf:(0.97)

3. MeningitesOutbreak=0 3393 ==> MalariaOutbreak=1 3149 conf:(0.93)

4. RelapsingFeverOutbreak=0 EpidemicTyphusOutbreak=0 2371 ==> MeningitesOutbreak=0 2193 conf:(0.92)

5. RelapsingFeverOutbreak=0 2827 ==> MalariaOutbreak=1 2605 conf:(0.92)

6. MeningitesOutbreak=0 RelapsingFeverOutbreak=0 2567 ==> MalariaOutbreak=1 2359 conf:(0.92)

7. RelapsingFeverOutbreak=0 EpidemicTyphusOutbreak=0 2371 ==> MalariaOutbreak=1 2159 conf:(0.91)

8. EpidemicTyphusOutbreak=0 2787 ==> MalariaOutbreak=1 2536 conf:(0.91)

9. RelapsingFeverOutbreak=0 2827 ==> MeningitesOutbreak=0 2567 conf:(0.91)

10. MeningitesOutbreak=0 EpidemicTyphusOutbreak=0 2498 ==> MalariaOutbreak=1 2268 conf:(0.91)

==========

1. TyphoidFeverOutbreak=1 2960==> MalariaOtbreak=1 2879 conf:(0.97)

The meaning of the above rule has shown that if there was the occurrence of Typhoid Fever disease case then the occurrence of malaria disease cases in a certain area and time period was 97% confidence with the support level of 70.34%. The support of the rule was calculated from the number of items on the right hand side of the rule 2879 with the total number of items in the transaction (datasets of disease cases) which was 4703. The number on the right hand side of the rule 2879 indicates that the number of items which were covered by its antecedent. As we can see from above there were two disease cases involved in the rule. Therefore this rule was generated from two frequently occurring disease cases. We can also calculate the confidence of the rule by the division of the number on the right hand sides of the arrow or rule to the left hand side of the arrow or the rule. The number of items on the consequence side was also shown after the rule.

Experiment 2:

Additional Best rules found after manually removal of the repeated rules:

3. TyphoidFeverOutbreak =1 RelapsingFeverOutbreak=0 1739 ==> MalariaOutbreak=1 1683 conf:(0.97)

4. MeningitesOutbreak=0 TyphoidFeverOutbreak =1 RelapsingFeverOutbreak=0 1519 ==> MalariaOutbreak=1 1468 conf:(0.97)

5. TyphoidFeverOutbreak =1 EpidemicTyphusOutbreak=0 1667 ==> MalariaOutbreak=1 1597 conf:(0.96)

8. MalariaOutbreak=1 RelapsingFeverOutbreak=0 EpidemicTyphusOutbreak=0 2159 ==> MeningitesOutbreak=0 1995 conf:(0.92)

13. MeningitesOutbreak=0 RelapsingFeverOutbreak=0 EpidemicTyphusOutbreak=0 2193 ==> MalariaOutbreak=1 1995 conf:(0.91)

Description of the rules generated:

5. TyphoidFeverOutbreak=1 Epidemic TyphusOutbreak=0 1667==> MalariaOtbreak=1 1597 conf:(0.96)

The above rule is a three itemsets (disease cases) rule. i.e three disease cases were involved. The rule has shown that if there were occurrences of Typhoid Fever Outbreak and with no occurrence of relapsing fever then the occurrence of malaria disease cases in a certain area and with some time period was 96% confidence with the support value of 39.02%.

8. MalariaOutbreak=1 RelapsingFeverOutbreak=0 EpidemicTyphusOutbreak=0 2159 ==> MeningitesOutbreak=0 1995 conf:(0.92)

The above rule is a four itemsets(disease cases) rule. i.e four disease cases were involved. The rule has shown that if there were occurrences of Malaria disease cases but with no occurrence of relapsing fever and epidemic typhus disease cases then meningitis disease cases were not occurring in a certain area and with some time period was 92% confidence with the support value of 48.74%.

3. TyphoidFeverOutbreak =1 RelapsingFeverOutbreak=0 1739 ==> MalariaOutbreak=1 1683 conf:(0.97)

The above rule is a three itemsets(disease cases) rule. i.e. three disease cases were involved. The rule has shown that if there were an occurrence of typhoid fever but no occurrences of Relapsing Fever then the occurrence of Malaria disease cases was with 97% confidence and 41.12% support level.

Experiment 3:

Additional Best rules found after manually removal of the repeated rules:

1. MalariaOutbreak=1 EpidemicTyphusOutbreak=1 1292 ==> TyphoidFeverOutbreak =1 1282 conf:(0.99)

2. TyphoidFeverOutbreak =1 EpidemicTyphusOutbreak=1 1293 ==> MalariaOutbreak=1 1282 conf:(0.99)

3. EpidemicTyphusOutbreak=1 1306 ==> TyphoidFeverOutbreak =1 1293 conf:(0.99)

4. EpidemicTyphusOutbreak=1 1306 ==> MalariaOutbreak=1 1292 conf:(0.99)

5. EpidemicTyphusOutbreak=1 1306 ==> MalariaOutbreak=1 TyphoidFeverOutbreak =1 1282 conf:(0.98)

10. TyphoidFeverOutbreak =1 RelapsingFeverOutbreak=0 EpidemicTyphusOutbreak=0 1292 ==> MalariaOutbreak=1 1244 conf:(0.96)

12. MeningitesOutbreak=0 TyphoidFeverOutbreak =1 EpidemicTyphusOutbreak=0 1424 ==> MalariaOutbreak=1 1361 conf:(0.96)

Description of some interesting rules:

1. MalariaOutbreak=1 EpidemicTyphusOutbreak=1 1292 ==> TyphoidFeverOutbreak =1 1282 conf:(0.99)

The above rule is a three itemsets(disease cases) rule. i.e. three disease cases were involved. The rule has shown that if there were occurrences of both malaria and epidemic typhus disease cases then the occurrence of typhoid fever disease cases was with 99% confidence and 31.32% support.

2. TyphoidFeverOutbreak =1 EpidemicTyphusOutbreak=1 1293 ==> MalariaOutbreak=1 1282 conf:(0.99)

The above rule is a three itemsets rule. I.e. three disease cases were involved. The rule has shown that if there were occurrences of typhoid fever and epidemic typhus then the occurrence of Malaria disease cases was with 99% confidence and 31.32% support.

3. EpidemicTyphusOutbreak=1 1306 ==> TyphoidFeverOutbreak =1 1293 conf:(0.99)

The above rule is a two itemsets (disease cases) rule. i.e. two disease cases were involved. The rule has shown that if there were occurrences of Epidemic Typhus disease cases then the occurrence of Typhoid Fever disease cases in a certain area and with some time period was 99% confidence with the support level of 31.59%.

4. EpidemicTyphusOutbreak=1 1306 ==> MalariaOutbreak=1 1292 conf:(0.99)

The above rule is a two itemsets (disease cases) rule. i.e. two disease cases were involved. The rule has shown that if there was an occurrence of epidemic typhus disease cases then the occurrence of malaria disease cases in a certain area and with some time period was 99% confidence with the support level of 31.57%.

5. EpidemicTyphusOutbreak=1 1306 ==> MalariaOutbreak=1 TyphoidFeverOutbreak =1 1282 conf:(0.98)

The above rule is a three itemsets (disease cases) rule. i.e. three disease cases were involved. The rule has shown that if there were occurrences of Epidemic Typhus disease cases then the occurrences of Typhoid Fever and malaria disease cases were with 98% confidence and support level of 31.32%.

Experiment 4:

Additional Best rules found after manually removal of the repeated rules:

2. TyphoidFeverOutbreak =0 RelapsingFeverOutbreak=0 1088 ==> EpidemicTyphusOutbreak=0 1079 conf:(0.99)

4.MeningitesOutbreak=0 TyphoidFeverOutbreak =0 RelapsingFeverOutbreak=0 1048 ==> EpidemicTyphusOutbreak=0 1039 conf:(0.99)

7. TyphoidFeverOutbreak =0 1133 ==> EpidemicTyphusOutbreak=0 1120 conf:(0.99)

8. MeningitesOutbreak=0 TyphoidFeverOutbreak =0 1087 ==> EpidemicTyphusOutbreak=0 1074 conf:(0.99)

10. TyphoidFeverOutbreak =1 RelapsingFeverOutbreak=1 1221 ==> MalariaOutbreak=1 1196 conf:(0.98)

11. MalariaOutbreak=1 RelapsingFeverOutbreak=1 1223 ==> TyphoidFeverOutbreak =1 1196 conf:(0.98)

13. MeningitesOutbreak=0 TyphoidFeverOutbreak =1 2306 ==> MalariaOutbreak=1 2232 conf:(0.97)

15. MeningitesOutbreak=0 TyphoidFeverOutbreak =0 EpidemicTyphusOutbreak=0 1074 ==> RelapsingFeverOutbreak=0 1039 conf:(0.97)

17. RelapsingFeverOutbreak=1 1266 ==> MalariaOutbreak=1 1223 conf:(0.97)

18. RelapsingFeverOutbreak=1 1266 ==> TyphoidFeverOutbreak =1 1221 conf:(0.96)

19. MeningitesOutbreak=0 TyphoidFeverOutbreak =0 1087 ==> RelapsingFeverOutbreak=0 1048 conf:(0.96)

20. TyphoidFeverOutbreak =0 EpidemicTyphusOutbreak=0 1120 ==> RelapsingFeverOutbreak=0 1079 conf:(0.96)

21. TyphoidFeverOutbreak =0 RelapsingFeverOutbreak=0 1088 ==> MeningitesOutbreak=0 1048 conf:(0.96)

22. TyphoidFeverOutbreak =0 RelapsingFeverOutbreak=0 EpidemicTyphusOutbreak=0 1079 ==> MeningitesOutbreak=0 1039 conf:(0.96)

24. MeningitesOutbreak=0 TyphoidFeverOutbreak =1 RelapsingFeverOutbreak=0 EpidemicTyphusOutbreak=0 1154 ==> MalariaOutbreak=1 1111 conf:(0.96)

25. TyphoidFeverOutbreak =0 1133 ==> RelapsingFeverOutbreak=0 1088 conf:(0.96)

Description of some interesting rules:

2. TyphoidFeverOutbreak =0 RelapsingFeverOutbreak=0 1088 ==> EpidemicTyphusOutbreak=0 1079 conf:(0.99)

The above rule is a three itemsets (disease cases) rule. i.e. three disease cases were involved. The rule has shown that if there were no occurrences of both typhoid fever and relapsing fever disease cases then there was no occurrences of epidemic typhus disease cases in a certain area and within some time period is 99% confidence with the support value of 26.36%. This rule has shown less support value. It was because of that the occurrence of either one or two or three of the disease cases were infrequent. But it was interesting since the confidence of the rule was high.

11. MalariaOutbreak=1 RelapsingFeverOutbreak=1 1223 ==> TyphoidFeverOutbreak =1 1196 conf:(0.98)

The above rule is a three itemsets (disease cases) rule. i.e. three disease cases were involved. The rule shows that if there are the occurrences of both Malaria and relapsing fever disease cases then the occurrences of typhoid fever disease cases are with 98% confidence and support level of 29.22%.

19. MeningitesOutbreak=0 TyphoidFeverOutbreak =0 1087 ==> RelapsingFeverOutbreak=0 1048 conf:(0.96)

The above rule is a three itemsets (disease cases) rule. i.e. three disease cases are involved. The rule shows that if there were no occurrences of both meningitis and typhoid fever disease cases then there was no occurrences of relapsing fever disease cases with 96% confidence and support level of 25.60%.

10. TyphoidFeverOutbreak =1 RelapsingFeverOutbreak=1 1221 ==> MalariaOutbreak=1 1196 conf:(0.98)

The above rule is a three itemsets (disease cases) rule. i.e. three disease cases were involved. The rule has shown that if there were the occurrences of both typhoid fever and relapsing fever disease cases then the occurrences of Malaria disease cases was with 98% confidence and support level of 29.22%.

Experiment 5:

Additional Best rules found after manually removal of the repeated rules:

26. TyphoidFeverOutbreak =0 1133 ==> MeningitesOutbreak=0 1087 conf:(0.96)

27. TyphoidFeverOutbreak =0 EpidemicTyphusOutbreak=0 1120 ==> MeningitesOutbreak=0 1074 conf:(0.96)

29. MeningitesOutbreak=0 TyphoidFeverOutbreak=0 1087 ==> RelapsingFeverOutbreak=0 EpidemicTyphusOutbreak=0 1039 conf:(0.96)

Description of some interesting rules:

26. TyphoidFeverOutbreak =0 1133 ==> MeningitesOutbreak=0 1087 conf:(0.96)

The above rule is a two itemsets (disease cases) rule. i.e. two disease cases were involved. The rule has shown that if there was a nonoccurrence of typhoid fever disease cases then

there was no occurrences of meningitis disease cases with 96% confidence and support level of 26.56%.

N:B: - All rules with less support but high confidence level was selected as an important rules with researcher specified minima. It was because of the goal of the research with respect to include the rarely occurring or infrequent disease cases from the datasets and disease cases can also be investigated even if they occurred infrequently if we provided the minSup values with small percentage. As mentioned and specified above to assign the values of the minimum support and minimum confidence level only the rules which had at least minSup = 20% and MinConf = 90% were stated as interesting. All the other rules were not considered from the view point of this research work to be best rules.

5.1.2.2. Association rules mining by Apriori Algorithm for IDSR datasets

The different experimentation of the association rule mining results with Apriori Algorithm were shown below in Table 3. The researcher showed that whenever there was an increasing in the number of rules the number of cycles were also increasing. But at the 4th and 5th experiments were showing an equal number of cycles. As the number of rule increased with constant confidence the support decreases with increasing rule numbers but it reached at constant level at the 4th and 5th experiments like in the PHEM dataset. Here in the rules generated from the datasets, important rules were detected while increasing the number of rules even if it was minimizing the support threshold level. I.e. the researcher set 20% min-sup threshold level. Apriori algorithm results for the 5 experimentations for IDSR datasets in tabular representation.

Exp No.	No. of instance	No. attributes involved	Min-Sup	No.of instances for Min-Sup	Min-Conf	No. Cycles performed	L(1)	L(2)	L(3)	L(4)	L(5)	L(6)	Best rules performed
E1	**4703**	**5**	**50%**	**2351**	**90%**	**10**	**5**	**9**	**6**	**-**	**-**	**-**	**10**
E2	**4703**	**5**	**35%**	**1646**	**90%**	**13**	**5**	**10**	**10**	**4**	**-**	**-**	**15**
E3	**4703**	**5**	**20%**	**941**	**90%**	**16**	**9**	**15**	**13**	**5**	**1**	**-**	**20**
E4	**4703**	**5**	**15%**	**705**	**90%**	**17**	**9**	**24**	**24**	**11**	**2**	**-**	**25**
E5	**4703**	**5**	**15%**	**705**	**90%**	**17**	**9**	**24**	**24**	**11**	**2**	**-**	**30**
E6	**4703**	**5**	**15%**	**705**	**90%**	**17**	**9**	**24**	**24**	**11**	**2**	**-**	**35**
E7	**4703**	**5**	**15%**	**705**	**90%**	**17**	**9**	**24**	**24**	**11**	**2**	**-**	**40**

Table5: The Run information to the five experimentation of Apriori association rule mining algorithm for IDSR datasets.

Descriptions of terms in the table are: Min-Sup=minimum support, Min-Conf=minimum confidence, L (1) = the first large 1-temsets, L (2) = the second large 2-temsets, L (3) = the third large 3-temsets, L (4) = the forth large 4-temsets and L (5) = the fifth large 5-temsets. The rules of each Experiment were described as follows:

Experiment 1:

=== Run information ===

Scheme: weka.associations.Apriori -N 10 -T 0 -C 0.9 -D 0.05 -U 1.0 -M 0.1 -S -1.0 -c -1

Relation: IDSR datasets for Apriori Algorithm
weka.filters.unsupervised.attribute.NumericToNominal-Rfirst-last

Instances: 4703

Attributes: 5

MalariaOtbreak

TyphoidFeverOutbreak

RelapsingFeverOutbreak

Epidemic TyphusOutbreak

MeningitOutbreak

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

Apriori

=======

Minimum support: 0.5 (2351 instances)

Minimum metric <confidence>: 0.9

Number of cycles performed: 10

Generated sets of large itemsets:

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

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

Size of set of large itemsets L(3): 6

Best rules found:

1. TyphoidFeverOutbreak=1 Epidemic TyphusOutbreak=0 2732 ==> MalariaOtbreak=1 2624 conf:(0.96)

2. Epidemic TyphusOutbreak=0 3698 ==> MalariaOtbreak=1 3479 conf:(0.94)

3. RelapsingFeverOutbreak=0 Epidemic TyphusOutbreak=0 2769 ==> MalariaOtbreak=1 2590 conf:(0.94)

4. Epidemic TyphusOutbreak=0 MeningitOutbreak=0 3020 ==> MalariaOtbreak=1 2818 conf:(0.93)

5. RelapsingFeverOutbreak=0 3065 ==> MalariaOtbreak=1 2819 conf:(0.92)

6. MalariaOtbreak=1 RelapsingFeverOutbreak=0 2819 ==> Epidemic TyphusOutbreak=0 2590 conf:(0.92)

7. RelapsingFeverOutbreak=0 MeningitOutbreak=0 2619 ==> MalariaOtbreak=1 2400 conf:(0.92)

8. TyphoidFeverOutbreak=1 3723 ==> MalariaOtbreak=1 3392 conf:(0.91)

9. RelapsingFeverOutbreak=0 MeningitOutbreak=0 2619 ==> Epidemic TyphusOutbreak=0 2380 conf:(0.91)

10. RelapsingFeverOutbreak=0 3065 ==> Epidemic TyphusOutbreak=0 2769 conf:(0.9)

Description of some interesting rules:

1. TyphoidFeverOutbreak=1 Epidemic TyphusOutbreak=0 2732 ==> MalariaOtbreak=1 2624 conf:(0.96)

The meaning of the above rule has shown that if there was the occurrence of Typhoid Fever disease cases with no occurrence of Epidemic Typhus disease cases then the occurrence of malaria outbreak in a certain area and time period was 96% confidence with the support value of 55.79%. The support of the rule was calculated from the number of items on the right hand side of the rule 2624 with the total number of items (disease cases) in the transaction (datasets) 4703. The number on the right hand side of the rule 2624 indicates that the number of items which were covered by its antecedent. As we could see above there were three outbreaks involved in the rule. Therefore this rule was generated from three frequently occurring disease cases. We could also calculate the confidence of the rule by the division of the number on the right hand sides of the arrow or rule to the left hand side of the arrow. The number of items on the consequence side was also shown after the rule. This calculation would also work for the rest of the rules.

2. Epidemic TyphusOutbreak=0 3698 ==> MalariaOtbreak=1 3479 conf:(0.94)

The above rule is a two itemset rule. i.e only two outbreak cases were involved. The rule has shown that if there was no occurrence of Epidemic Typhus disease cases then the occurrence of malaria disease cases in a certain area and with some time period was 94% confidence with the support value of 73.97%.

8. TyphoidFeverOutbreak=1 3723 ==> MalariaOtbreak=1 3392 conf:(0.91)

The above rule is a two itemset rule. i.e only two outbreak cases were involved. The rule shows that if Typhoid Fever Outbreak is occurring then the occurrence of malaria outbreak in

a certain area and with some time period is 91% confidence with the support value of 72.12%. The number on the right hand side of the rule 3479 which indicates that the number of items which are covered by its antecedent.

Experiment 2:

Additional interesting rules:

2. TyphoidFeverOutbreak=1 RelapsingFeverOutbreak=0 Epidemic TyphusOutbreak=0 1889 ==> MalariaOtbreak=1 1814 conf:(0.96)

3. TyphoidFeverOutbreak=1 Epidemic TyphusOutbreak=0 MeningitOutbreak=0 2143 ==> MalariaOtbreak=1 2040 conf:(0.95)

6. TyphoidFeverOutbreak=1 RelapsingFeverOutbreak=0 2177 ==> MalariaOtbreak=1 2035 conf:(0.93)

8. RelapsingFeverOutbreak=0 Epidemic TyphusOutbreak=0 MeningitOutbreak=0 2380 ==> MalariaOtbreak=1 2217 conf:(0.93)

9. TyphoidFeverOutbreak=1 RelapsingFeverOutbreak=0 MeningitOutbreak=0 1806 ==> MalariaOtbreak=1 1679 conf:(0.93)

10. MalariaOtbreak=1 RelapsingFeverOutbreak=0 MeningitOutbreak=0 2400 ==> Epidemic TyphusOutbreak=0 2217 conf:(0.92)

Description of some interesting rules:

3. TyphoidFeverOutbreak=1 Epidemic TyphusOutbreak=0 MeningitOutbreak=0 2143 ==> MalariaOtbreak=1 2040 conf:(0.95)

The above rule is a four itemsets rule. i.e four outbreak cases were involved. The rule has shown that if there were Typhoid Fever disease cases occurrence with no occurrence of Epidemic Typhus and meningitis disease cases then the occurrence of malaria outbreak in a certain area and with some time period was 95% confidence with the support value of 43.38%. The number on the right hand side of the rule 2040 indicated that the number of items (disease cases) which were covered by its antecedent.

2. TyphoidFeverOutbreak=1 RelapsingFeverOutbreak=0 Epidemic TyphusOutbreak=0 1889 ==> MalariaOtbreak=1 1814 conf:(0.96)

The above rule is a four itemsets rule. I.e. four diseases cases were involved. The rule has shown if there were an occurrence of Typhoid Fever disease cases with no occurrence of Epidemic Typhus and meningitis disease cases then the occurrence of malaria outbreak in a certain area and with some time period was 96% confidence with the support value of 38.57%.

Experiment 3:

Additional interesting rules:

1. Epidemic TyphusOutbreak=1 1005 ==> TyphoidFeverOutbreak=1 991 conf:(0.99)

2. TyphoidFeverOutbreak=0 980 ==> Epidemic TyphusOutbreak=0 966 conf:(0.99)

5. TyphoidFeverOutbreak=1 RelapsingFeverOutbreak=0 Epidemic TyphusOutbreak=0 MeningitOutbreak=0 1572 ==> MalariaOtbreak=1 1501 conf:(0.95)

7. RelapsingFeverOutbreak=1 1638 ==> TyphoidFeverOutbreak=1 1546 conf:(0.94)

8. MalariaOtbreak=1 RelapsingFeverOutbreak=1 1442 ==> TyphoidFeverOutbreak=1 1357 conf:(0.94)

11. RelapsingFeverOutbreak=1 MeningitOutbreak=0 1139 ==> TyphoidFeverOutbreak=1 1065 conf:(0.94)

Description of some interesting rules:

Here the researcher found that there are other important rules to be described.

1. Epidemic TyphusOutbreak=1 1005 ==> TyphoidFeverOutbreak=1 991 conf:(0.99)

The above rule is a two itemsets rule. i.e. two disease cases were involved. The rule has shown that if there was an occurrence of Epidemic Typhus disease cases then the occurrence of Typhoid Fever disease cases in a certain area and with some time period was 99% confidence with the support level of 21.07%. This rule shows less support values. It is because of that the occurrence of the two disease cases or one of the disease cases were infrequent manner but the rules are interesting since the confidence value of the rules was very high and the supports were higher than the researcher specified minimum threshold.

2. TyphoidFeverOutbreak=0 980 ==> Epidemic TyphusOutbreak=0 966 conf:(0.99)

The above rule is a two itemsets rule. i.e. two outbreak cases were involved. The rule has shown that if there was no occurrence of Typhoid Fever disease cases then Epidemic Typhus disease cases were not occur with 99% confidence and support level of 20.54%. This rule has shown less support values but higher confidence. It was because of that the occurrence of the two disease cases was not so much together but interesting since the value of the confidence was high.

7. RelapsingFeverOutbreak=1 1638 ==> TyphoidFeverOutbreak=1 1546 conf:(0.94)

The above rule is a two itemsets rule. I.e. two outbreak cases were involved. The rule has shown that if there was the occurrence of Relapsing Fever disease cases then the occurrence

of Typhoid Fever disease cases in a certain area and with some time period was 94% confidence with the support level of 32.87%.

8. MalariaOtbreak=1 RelapsingFeverOutbreak=1 1442 ==> TyphoidFeverOutbreak=1 1357 conf:(0.94)

The above rule is a three itemsets rule. i.e. three disease cases were involved. The rule has shown that if there are the occurrences of Malaria and Relapsing Fever disease cases then the occurrence of typhoid fever disease cases was with 94% confidence and 38.85% support level.

Experiment 4:

Additional interesting rules:

2. MalariaOtbreak=1 TyphoidFeverOutbreak=0 RelapsingFeverOutbreak=0 MeningitOutbreak=0 721 ==> Epidemic TyphusOutbreak=0 716 conf:(0.99)

3. TyphoidFeverOutbreak=0 RelapsingFeverOutbreak=0 888 ==> Epidemic TyphusOutbreak=0 880 conf:(0.99)

4. MalariaOtbreak=1 TyphoidFeverOutbreak=0 RelapsingFeverOutbreak=0 784 ==> Epidemic TyphusOutbreak=0 776 conf:(0.99)

5. TyphoidFeverOutbreak=0 MeningitOutbreak=0 887 ==> Epidemic TyphusOutbreak=0 877 conf:(0.99)

6. MalariaOtbreak=1 TyphoidFeverOutbreak=0 MeningitOutbreak=0 788 ==> Epidemic TyphusOutbreak=0 778 conf:(0.99)

7. Epidemic TyphusOutbreak=1 MeningitOutbreak=0 738 ==> TyphoidFeverOutbreak=1 728 conf:(0.99)

10. MalariaOtbreak=1 TyphoidFeverOutbreak=0 869 ==> Epidemic TyphusOutbreak=0 855 conf:(0.98)

11. MalariaOtbreak=1 Epidemic TyphusOutbreak=1 782 ==> TyphoidFeverOutbreak=1 768 conf:(0.98)

12. TyphoidFeverOutbreak=1 RelapsingFeverOutbreak=1 Epidemic TyphusOutbreak=0 843 ==> MalariaOtbreak=1 810 conf:(0.96)

15. TyphoidFeverOutbreak=1 MeningitOutbreak=1 852 ==> MalariaOtbreak=1 816 conf:(0.96)

16. RelapsingFeverOutbreak=1 Epidemic TyphusOutbreak=0 929 ==> MalariaOtbreak=1 889 conf:(0.96)

19. MeningitOutbreak=1 945 ==> MalariaOtbreak=1 897 conf:(0.95)

Description of some interesting rules:

11. MalariaOtbreak=1 Epidemic TyphusOutbreak=1 782 ==> TyphoidFeverOutbreak=1 768 conf:(0.98)

The above rule is a three itemsets rule. i.e. three disease cases were involved. The rule has shown that if there were the occurrences of malaria and epidemic typhus disease cases then

the occurrence of Typhoid Fever Outbreak in a certain area and within some time period was 98% confidence with the support value of 16.33%. This rule has shown less support value. It was because of that the occurrence of either one or two or three of the disease cases are infrequent. However, as defined in the criteria to define the min-sup of disease cases, this rule was not selected due to the min-sup given by the researcher even if it has high minimum Confidence level.

10. MalariaOtbreak=1 TyphoidFeverOutbreak=0 869 ==> Epidemic TyphusOutbreak=0 855 conf:(0.98)

The above rule is a three itemsets rule. I.e. three disease cases were involved. The rule has shown that if there were the occurrence of Malaria disease cases with nonoccurrence of typhoid fever disease cases then the occurrence of Epidemic Typhus disease cases was with 98% confidence and support level of 18.18%. This rule has shown less support value. It was because of that non occurrence or occurrence of either one or two or three of the disease cases were infrequent. However, as defined in the criteria to define the min-sup of disease cases, this rule was not selected a better rule due to the min-sup threshold given by the researcher.

Experimentation 5:

1. MalariaOtbreak=1 RelapsingFeverOutbreak=1 MeningitOutbreak=0 964 ==> TyphoidFeverOutbreak=1 897 conf:(0.93)

Descriptions of the rules:

In the rule above, if malaria disease cases and relapsing fever disease cases occurred with no meningitis disease cases occurrence in a certain place with specified time period then the occurrence of typhoid fever disease cases was with 93% confidence and support level of 19.07%. This rule has shown less support value. It was because of that non occurrence or occurrence of either one or two or three of the disease cases were infrequent. However, as defined in the criteria to define the min-sup of disease cases, this rule was not selected a better rule due to the min-sup threshold given by the researcher.

N: B: - the rules with less than the specified min-sup and min-conf were not interesting for this research since the predefined min-sup and min-conf threshold set by the researcher. Some rule with less than the minimum support but higher than the researcher specified minimum were more interesting for the rare diseases and more occurring disease cases to

show the correlation or the association among disease cases in the datasets. As stated above, it was also done that the repeated rules from the different experimentations were discarded manually to reduce the number of rules for analysis activities and to show only the new rules generated from the Apriori algorithm by adding some number of rules to the next run (i.e. 5 additional rules for the next experimentation).

5.1.2.3. Association rules mining by Apriori Algorithm for combined Dataset (2004-2012g.C).

Here the researcher also tried to analyze both datasets one by one separately and finally merged them together to analyze the combined effects of the two datasets. Therefore; the researcher has used the same fixed minimum support and minimum confidence thresholds which was fixed for the two separate datasets. The results of the two individual datasets and the combined datasets were compared for interesting rules and best rules were selected accordingly. Rules generation process was the same with the previous individual datasets. The table below showed that the rules are described in relation to the predefined min-sup and min-conf threshold.

Exp No.	No. of instance	No. attributes involved	Min-Sup	No.of instances for Min-Sup	Min-Conf	No. Cycles performed	L(1)	L(2)	L(3)	L(4)	L(5)	L(6)	Best rules performed
E1	**8796**	**5**	**45%**	**3958**	**90%**	**11**	**5**	**9**	**6**	**1**	**-**	**-**	**10**
E2	**8796**	**5**	**30%**	**2639**	**90%**	**14**	**6**	**12**	**10**	**5**	**-**	**-**	**15**
E3	**8796**	**5**	**25%**	**2199**	**90%**	**15**	**7**	**13**	**11**	**5**	**1**	**-**	**20**
E4	**8796**	**5**	**20%**	**1759**	**90%**	**16**	**8**	**19**	**17**	**6**	**1**	**-**	**25**
E5	**8796**	**5**	**20%**	**1759**	**90%**	**16**	**8**	**19**	**17**	**6**	**1**	**-**	**30**
E6	**8796**	**5**	**20%**	**1759**	**90%**	**16**	**8**	**19**	**17**	**6**	**1**	**-**	**35**
E7	**8796**	**5**	**15%**	**1319**	**90%**	**17**	**9**	**24**	**25**	**12**	**2**	**-**	**40**

Table6: The Run information to the five experimentation of Apriori association rule mining algorithm for IDSR datasets.

Experiment 1:

=== Run information ===

Scheme: weka.associations.Apriori -N 10 -T 0 -C 0.9 -D 0.05 -U 1.0 -M 0.1 -S -1.0 -c -1

Relation: 2005-2008 IDSR and 2009-2012 PHEM DATASETS
weka.filters.unsupervised.attribute.NumericToNominal-Rfirst-last-
weka.filters.unsupervised.attribute.NumericToNominal-Rfirst-last

Instances: 8796

Attributes: 5

MalariaOutbreak

TyphoidFever Outbreak

RelapsingFeverOutbreak

EpidemicTyphusOutbreak

MeningitOutbreak

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

Apriori

=======

Minimum support: 0.45 (3958 instances)

Minimum metric <confidence>: 0.9

Number of cycles performed: 11

Generated sets of large itemsets:

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

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

Size of set of large itemsets L(3): 6

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

Best rules found:

1. TyphoidFever Outbreak=1 EpidemicTyphusOutbreak=0 4399 ==> MalariaOutbreak=1 4221 conf:(0.96)

2. TyphoidFever Outbreak=1 6683 ==> MalariaOutbreak=1 6271 conf:(0.94)

3. TyphoidFever Outbreak=1 MeningitOutbreak=0 5177 ==> MalariaOutbreak=1 4808 conf:(0.93)

4. EpidemicTyphusOutbreak=0 6485 ==> MalariaOutbreak=1 6015 conf:(0.93)

5. RelapsingFeverOutbreak=0 EpidemicTyphusOutbreak=0 5140 ==> MalariaOutbreak=1 4749 conf:(0.92)

6. EpidemicTyphusOutbreak=0 MeningitOutbreak=0 5518 ==> MalariaOutbreak=1 5086 conf:(0.92)

7. RelapsingFeverOutbreak=0 EpidemicTyphusOutbreak=0 MeningitOutbreak=0 4573 ==> MalariaOutbreak=1 4212 conf:(0.92)

8. RelapsingFeverOutbreak=0 5892 ==> MalariaOutbreak=1 5424 conf:(0.92)

9. RelapsingFeverOutbreak=0 MeningitOutbreak=0 5186 ==> MalariaOutbreak=1 4759 conf:(0.92)

10. MeningitOutbreak=0 7151 ==> MalariaOutbreak=1 6513 conf:(0.91)

Description of some interesting rules:

1. TyphoidFeverOutbreak=1 Epidemic TyphusOutbreak=0 4399 ==> MalariaOtbreak=1 4221 conf:(0.96)

The meaning of the above rule has shown that if there were occurrences of Typhoid Fever disease cases and with no occurrences epidemic typhus disease cases then the occurrence of malaria disease cases in a certain area and time period was 96% confidence with the support level of 47.98%. The support of the rule is calculated from the number of items (disease cases) on the right hand side of the rule 4221 with the total number of items (disease cases) in the transaction (datasets of disease cases) 8796.

2. TyphoidFever Outbreak=1 6683 ==> MalariaOutbreak=1 6271 conf:(0.94)

The above rule is a two itemset(disease cases) rule. i.e only two disease cases were involved. The rule has shown that if there was occurrence of typhoid fever disease cases then the occurrence of malaria disease cases in a certain area and with some time period is 94% confidence with the support value of 71.29%.

Experiment 2:

Additional interesting rules:

1. TyphoidFever Outbreak=1 RelapsingFeverOutbreak=0 EpidemicTyphusOutbreak=0 3181 ==> MalariaOutbreak=1 3058 conf:(0.96)

3. TyphoidFever Outbreak=1 EpidemicTyphusOutbreak=0 MeningitOutbreak=0 3567 ==> MalariaOutbreak=1 3401 conf:(0.95)

4. RelapsingFeverOutbreak=1 2904 ==> TyphoidFever Outbreak=1 2767 conf:(0.95)

5. TyphoidFever Outbreak=1 RelapsingFeverOutbreak=0 3916 ==> MalariaOutbreak=1 3718 conf:(0.95)

6. TyphoidFever Outbreak=1 RelapsingFeverOutbreak=0 MeningitOutbreak=0 3325 ==> MalariaOutbreak=1 3147 conf:(0.95)

13. RelapsingFeverOutbreak=0 5892 ==> MalariaOutbreak=1 5424 conf:(0.92)

Description of some interesting rules:

1. TyphoidFever Outbreak=1 RelapsingFeverOutbreak=0 EpidemicTyphusOutbreak=0 3181 ==> MalariaOutbreak=1 3058 conf:(0.96)

The above rule is a four itemsets (disease cases) rule. I.e. four diseases cases were involved. The rule has shown that if there were occurrences of typhoid fever disease cases and no occurrence of relapsing fever and epidemic typhus disease cases then the occurrence of malaria disease cases was with 96% confidence and with the support level of 34.77%.

4. RelapsingFeverOutbreak=1 2904 ==> TyphoidFever Outbreak=1 2767 conf:(0.95)

The above rule is a two itemsets (disease cases) rule. i.e two outbreak cases were involved. The rule has shown that if there was an occurrence of relapsing fever disease cases then the occurrence of Typhoid Fever disease cases in a certain area and with some time period was 95% confidence with the support value of 31.46%.

6. TyphoidFever Outbreak=1 RelapsingFeverOutbreak=0 MeningitOutbreak=0 3325 ==> MalariaOutbreak=1 3147 conf:(0.95)

The above rule is a four itemsets (disease cases) rule. I.e. four disease cases were involved. The rule has shown that if there were an occurrence of Typhoid Fever disease cases and neither occurrences of relapsing fever nor meningitis disease cases then the occurrence of malaria disease cases in a certain area and with some time period was 95% confidence with the support value of 35.78%.

Experiment 3:

Additional interesting rules:

1. EpidemicTyphusOutbreak=1 2311 ==> TyphoidFever Outbreak=1 2284 conf:(0.99)

4. TyphoidFever Outbreak=1 RelapsingFeverOutbreak=0 EpidemicTyphusOutbreak=0 MeningitOutbreak=0 2726 ==> MalariaOutbreak=1 2612 conf:(0.96)

5. MalariaOutbreak=1 RelapsingFeverOutbreak=1 2665 ==> TyphoidFever Outbreak=1 2553 conf:(0.96)

14. TyphoidFever Outbreak=1 RelapsingFeverOutbreak=1 2767 ==> MalariaOutbreak=1 2553 conf:(0.92)

Description of some interesting rules:

1. EpidemicTyphusOutbreak=1 2311 ==> TyphoidFever Outbreak=1 2284 conf:(0.99)

The above rule is a two itemsets (disease cases) rule. i.e. two disease cases were involved. The rule has shown that if there was occurrence of Typhoid Fever disease cases then the occurrence of Epidemic Typhus disease cases in a certain area and with some time period was 99% confidence with the support level of 25.97%.

4. TyphoidFever Outbreak=1 RelapsingFeverOutbreak=0 EpidemicTyphusOutbreak=0 MeningitOutbreak=0 2726 ==> MalariaOutbreak=1 2612 conf:(0.96)

The above rule is a five itemsets (disease cases) rule. i.e. five disease cases were involved. The rule has shown that if there were occurrences of typhoid fever and no occurrences of relapsing fever, epidemic typhus and meningitis disease cases then the occurrence of Malaria disease cases was with 96% confidence and 29.70% support level.

5. MalariaOutbreak=1 RelapsingFeverOutbreak=1 2665 ==> TyphoidFever Outbreak=1 2553 conf:(0.96)

The above rule is a three itemsets (disease cases) rule. i.e. three disease cases were involved. The rule has shown that if there were the occurrences of both Malaria and relapsing fever disease cases then the occurrence of typhoid fever disease cases was with 96% confidence and support level of 29.02%. This rule shows less support value. It was because of that none occurrences of either one or two of the disease cases were infrequent but the other one or two were frequent.

14. TyphoidFever Outbreak=1 RelapsingFeverOutbreak=1 2767 ==> MalariaOutbreak=1 2553 conf:(0.92)

The above rule is a three itemsets (disease cases) rule. i.e. three disease cases were involved. The rule has shown that if there were the occurrences of both typhoid fever and relapsing fever disease cases then the occurrences of Malaria disease cases was with 92% confidence and support level of 29.02%. This rule shows less support value. It was because of that none occurrence of either one or two or three of the disease cases were infrequent.

Experiment 4:

Additional interesting rules:

1.TyphoidFever Outbreak=0 RelapsingFeverOutbreak=0 MeningitOutbreak=0 1861 ==> EpidemicTyphusOutbreak=0 1847 conf:(0.99)

2.TyphoidFever Outbreak=0 RelapsingFeverOutbreak=0 1976 ==> EpidemicTyphusOutbreak=0 1959 conf:(0.99)

3. MalariaOutbreak=1 EpidemicTyphusOutbreak=1 2074 ==> TyphoidFever Outbreak=1 2050 conf:(0.99)

4. TyphoidFever Outbreak=0 MeningitOutbreak=0 1974 ==> EpidemicTyphusOutbreak=0 1951 conf:(0.99)

6. TyphoidFever Outbreak=0 2113 ==> EpidemicTyphusOutbreak=0 2086 conf:(0.99)

7. MalariaOutbreak=1 TyphoidFever Outbreak=0 1818 ==> EpidemicTyphusOutbreak=0 1794 conf:(0.99)

15.TyphoidFever Outbreak=0 EpidemicTyphusOutbreak=0 MeningitOutbreak=0 1951 ==> RelapsingFeverOutbreak=0 1847 conf:(0.95)

17.TyphoidFever Outbreak=0 RelapsingFeverOutbreak=0 EpidemicTyphusOutbreak=0 1959 ==> MeningitOutbreak=0 1847 conf:(0.94)

18. TyphoidFever Outbreak=0 MeningitOutbreak=0 1974 ==> RelapsingFeverOutbreak=0 1861 conf:(0.94)

19. RelapsingFeverOutbreak=1 MeningitOutbreak=0 1965 ==> TyphoidFever Outbreak=1 1852 conf:(0.94)

20. TyphoidFever Outbreak=0 RelapsingFeverOutbreak=0 1976 ==> MeningitOutbreak=0 1861 conf:(0.94)

21.TyphoidFever Outbreak=0 EpidemicTyphusOutbreak=0 2086 ==> RelapsingFeverOutbreak=0 1959 conf:(0.94)

23.TyphoidFever Outbreak=0 MeningitOutbreak=0 1974 ==> RelapsingFeverOutbreak=0 EpidemicTyphusOutbreak=0 1847 conf:(0.94)

24.TyphoidFever Outbreak=0 EpidemicTyphusOutbreak=0 2086 ==> MeningitOutbreak=0 1951 conf:(0.94)

25. TyphoidFever Outbreak=0 2113 ==> RelapsingFeverOutbreak=0 1976 conf:(0.94)

Description of some interesting rules:

2.TyphoidFever Outbreak=0 RelapsingFeverOutbreak=0 1976 ==> EpidemicTyphusOutbreak=0 1959 conf:(0.99)

The above rule is a three itemsets (disease cases) rule. i.e. three disease cases were involved. The rule has shown that if there were none occurrences of both typhoid fever and relapsing fever disease cases then the non occurrence of Epidemic Typhus disease cases was with 99% confidence and with the support level of 22.27%. This rule has shown less support value. It was because of that the occurrence of either one or two or three of the disease cases were infrequent. But it was interesting since the confidence of the rule was high and the minimum support value was above the specified minimum support threshold.

3. MalariaOutbreak=1 EpidemicTyphusOutbreak=1 2074 ==> Typhoid Fever Outbreak=1 2050 conf: (0.99)

The above rule is a three itemsets (disease cases) rule. i.e. three disease cases were involved. The rule showed that if there were the occurrences of both Malaria and relapsing fever disease cases then the occurrences of typhoid fever disease cases was with 99% confidence and support level of 23.31%. This rule shows less support value. It was because of that none occurrence of either one or two or three of the disease cases were infrequent.

19. RelapsingFeverOutbreak=1 MeningitOutbreak=0 1965 ==> TyphoidFever Outbreak=1 1852 conf:(0.94)

The above rule is a three itemsets (disease cases) rule. i.e. three disease cases were involved. The rule showed that if there were occurrences relapsing fever and no occurrences of meningitis disease cases then the occurrences of typhoid fever disease cases was with 94% confidence and support level of 21.06%. This rule showed less support value but more than that of the researcher specified threshold. It was because of that the occurrence of either one or two or three of the disease cases were infrequent.

Experiment 5:

Additional interesting rules:

26. TyphoidFever Outbreak=0 RelapsingFeverOutbreak=0 1976 ==> EpidemicTyphusOutbreak=0 MeningitOutbreak=0 1847 conf:(0.93)

27. TyphoidFever Outbreak=0 2113 ==> MeningitOutbreak=0 1974 conf:(0.93)

30. TyphoidFever Outbreak=0 2113 ==> RelapsingFeverOutbreak=0 EpidemicTyphusOutbreak=0 1959 conf:(0.93)

Description of some interesting rules:

26.TyphoidFever Outbreak=0 RelapsingFeverOutbreak=0 1976 ==> EpidemicTyphusOutbreak=0 MeningitOutbreak=0 1847 conf:(0.93)

The above rule is a four itemsets (disease cases) rule. i.e. four disease cases were involved. The rule has shown that if there were occurrences of neither typhoid fever nor relapsing fever disease cases then there are also occurrences of neither epidemic typhus nor meningitis disease cases with 93% confidence and support level of 21.00%. This rule has shown less support value. It was because of that the occurrence of either one or two or three of the disease cases are infrequent but the remaining cases are frequent.

27. TyphoidFever Outbreak=0 2113 ==> MeningitOutbreak=0 1974 conf:(0.93)

The above rule is a two itemsets (disease cases) rule. i.e. two disease cases are involved. The rule shows that if there are no occurrences of Typhoid Fever disease cases then there are no occurrences of meningitis disease cases in a certain area and with some time period is 93% confidence with the support level of 22.44%. This rule has shown less support value. It was because of that the occurrence of either of the two disease cases was infrequent but the occurrence of the other disease cases was frequent.

30. TyphoidFever Outbreak=0 2113 ==> RelapsingFeverOutbreak=0 EpidemicTyphusOutbreak=0 1959 conf:(0.93)

The above rule is a three itemsets (disease cases) rule. i.e. three disease cases were involved. The rule shows that if there were no occurrences of typhoid fever disease cases then the occurrences of neither relapsing fever nor epidemic typhus disease cases were with 93% confidence and support level of 22.27%. This rule has shown less support value. It was because of that the occurrence of either one or two or three of the disease cases are infrequent but the remaining cases are frequent.

5.1.2.4. Summary of association rules mining results from Apriori Algorithm

Rule comparison summary

The comparison of Apriori algorithm results for the repetition of rules and the selecting interesting rules from the three datasets (IDSR, PHEM, and the combination of IDSR with PHEM) were shown on table below.

Exp No.	Total numbers rules in each experiment	IDSR datasets		PHEM datasets		The whole combined datasets of IDSR and PHEM	
		Repeated Rules	Rules left after removal of repeated rules	Repeated Rules	Rules left after removal of repeated rules	Repeated Rules	Rules left after removal of repeated rules
E1	**10**	**-**	**10**	**-**	**10**	**-**	**10**
E2	**15**	**9**	**6**	**10**	**5**	**9**	**6**
E3	**20**	**14**	**6**	**13**	**7**	**16**	**4**
E4	**25**	**12**	**13**	**9**	**16**	**10**	**15**
E5	**30**	**29**	**1**	**27**	**3**	**27**	**3**
Total rules	**100**	**64**	**36**	**59**	**41**	**62**	**38**

Table7: Comparison among the three datasets for Apriori association rules mining result

As shown on the table above, the total numbers of interesting rules after the removal of the repeated rules was identified. That means, the PHEM datasets produced 41 interesting rules, lastly IDSR dataset produced about 36 interesting rules and the combination of the two datasets produced 38 interesting rules on the five experimentation. From the above statistics we could see the number of interesting rules were more in the PHEM datasets than the two datasets. This justification didn't show that the PHEM dataset was better than the other. Rather we could say that the PHEM dataset could produce more number of rules with the specified researcher's min-sup and min-conf threshold. It was also true that rule interestingness is not a measure of number of rules produce rather number of interesting rules produced from a certain datasets.

Result Analysis Summary

Since the research was for educational purpose and there was also time constraint to describe all interesting rules generated from Apriori algorithm, the researcher tried to show only some of the interesting rules from each experiment. Here the results of association rules mining showed that those rules which were selected as interesting from the three datasets were compared and contrasted below in the table8 given: - since the transaction of disease outbreak

cases were recorded according to the time and place setting then these transaction was investigated further in the following discussions for the selected rules. As mentioned in chapter2 of this research, association rule didn't show any causality between disease cases. Therefore it could only be expected which disease occurred with the occurrence of which types of disease outbreak. Best common rules were selected for the Comparison purpose among the whole three datasets for association rules mining with Apriori Algorithm (Rule numbers were given here manually for the sake of discussion purpose) to some selected rules which were getting more acceptance with the domain experts.

Dataset Name	Rule No.	Best Rules Produced				Measure of Association		Remark
		Antecedence		Consequence		Confidence	Support	
		Occur	Not occur	Occur	Not occur			
IDSR	1.		TF, RF		ET	99%	18.71%	Not accepted
	2.	ET		TF		99%	21.01%	Accepted
	3.	Mal, ET		TF		98%	18.86%	Not accepted
	4.	Mal, RF		TF		94%	28.86%	Accepted
	5.	RF		TF		94%	32.87%	Accepted
	6.	TF		Mal		91%	72.12%	Accepted
PHEM	**1.**		TF, RF		ET	99%	26.34%	Accepted
	2.	ET		TF		99%	31.57%	Accepted
	3.	Mal, ET		TF		99%	31.3%	Accepted
	4.	Mal, RF		TF		98%	29.12%	Accepted
	5.	RF		TF		96%	29.83%	Accepted
	6.	TF		Mal		97%	70.29%	Accepted
Combined datasets	1.		TF, RF		ET	99%	22.27%	Accepted
	2.	ET		TF		99%	25.97%	Accepted
	3.	Mal, ET		TF		99%	23.31%	Accepted
	4.	Mal, RF		TF		96%	29.02%	Accepted
	5.	RF		TF		95%	31.46%	Accepted
	6.	TF		Mal		94%	71.29%	Accepted

Table8: Selected rules and their respected support and confidence level

N: B: - ET=Epidemic Typhus disease case, Mal=Malaria disease case, Men=Meningitis disease case, RF=Relapsing Fever disease case, TF=Typhoid Fever disease case.

The remark showed that rules were accepted or not accepted with regard to the rules interestingness measures which were defined according to the minimum support and minimum confidence threshold of the rules (since the researcher defined it for rule interestingness by considering the rarely or infrequently occurring disease cases).

Rules acceptance only indicated interestingness and not acceptance indicated as rules were not interesting with the researcher's specified or predefined threshold. The table below showed the average support and average confidence of the rules given above for the whole three datasets with the predefined rules with its sequence numbers given manually for each dataset for the same rule. I.e. the first rule which was rule number 1 has the same variables for the three datasets and soon.

Rule numbers	Average support	Average confidence	Remark
1.	22.44%	99%	Accepted
2.	26.16%	99%	Accepted
3.	23.85%	98.67%	Accepted
4.	29%	96%	Accepted
5.	31.39%	95%	Accepted
6.	55.97%	94%	Accepted

Table9: selecting rule interestingness from the average support and confidence of the three datasets

Finally, when the researcher considered the datasets separately, the IDSR datasets were having two uninteresting rules from the selected rules but the PHEM as well as the combined datasets had shown that all the selected rules were interesting according to the threshold given. Therefore the researcher had decided to take the average values of the same rule from the three datasets and decided the rule interestingness accordingly. Hence the average support and confidence of the six rules had a greater average than that of the minimum required threshold (for Min-Sup=20% and Min-Conf=90%). There were repetitions of some rules to the next run while adding 5 additional rules on the first run of Apriori algorithm. Therefore; those repeated rules were not important to further analysis then the researcher manually removed those repeated rules and made analysis on the remaining rules.

5.1.3. Model development for disease clustering with a simple-KMeans Algorithm

As stated in the methodology part of this research work, the goal of data clustering, also known as cluster analysis, is to discover the *natural* grouping(s) of a set of patterns, points, or objects. Since the data sets were cleaned for missing values, noise, inconsistencies, and outlier's specification in the preprocessing phase of this research work then the simple K-means algorithm was performed simply.

As per the research objectives, the researcher analyzes the results to determine which regions and Zones of the country would be affected by disease cases with respect to time (i.e. the year and the month). Then disease segmentation was performed within time and place distributions. It was also important to know that cluster analysis is a descriptive or exploratory data mining applications; therefore the output of clustering algorithms can only lead to suggest hypotheses and plan of action for decision making activities.

Minimum number of iteration and minimum min-square error for simple K-means were chosen for disease case clustering. After the data preprocessing phase, the total numbers of instances or records involved were 8796 and the numbers of attributes were 9. Consequently, the attributes were: *RegionName, ZoneName, Year, Month, MalariaOutbreak, TyphiodFeverOutbreak, RelapsingFeverOutbreak, EpidemicTyphusOutbreak, and MeningitesOutbreak.* Here the most difficult task was to determine the number of cluster elements K for disease case data segmentation as mentioned above. However; the researcher with the discussion to the domain experts had decided to generate appropriate number of clusters to group the disease cases data in to the total number of years involved for experimentation for this research and number of total regions in the country taken for the clustering purpose.

Finally, we agreed up on the segmentation of diseases cases first with a minimum of 9 clusters for the 9 years datasets to show the time variation with the occurrence and nonoccurrence of disease cases and then 11 clusters for the 11 region of the country. Here the researcher also tried to find more clusters according to the sum of square errors for simple K-means algorithm and tried the possible clusters in between. 10 cluster segments which were

in between of the 9 years and the 11 regions of the country where tried to see whether there are some cluster elements with different possible outcomes or not.

The researcher then analyzed the within cluster sum of squared errors (min-square error), the number of iteration performed, and the percent (%) of instances involved in each clusters of the three experiments with the values of K = 9, 10, and 11 number of clusters. Then the minimum within cluster sum of squared errors and less number of iteration performed were taken as good cluster.

As shown in the Table below, since this experiment 3 had lower value of within cluster sum of squared errors which was equal to 26,503 and number of iteration taken = 11. Even though the number of iteration was more than that of experiment 2 which had only 10 iteration performed, experiment 2 was having greater cluster sum of squared errors and less numbers of clusters. Then the researcher decided to use and discus about this experiment 3.

Performed Experiments	Within cluster sum of squared errors	Number of iterations performed	No. of instances involved	No. of attributes taken	No. of clusters(K) Defined
Experiment 1	27,311	13	8796	9	9
Experiment 2	26,780	10	8796	9	10
Experiment 3	26,503	11	8796	9	11

Table10: Clustering Experimentation analysis for the comparison of within cluster sum of squared errors and Number of iterations performed for each experimentation.

The researcher also wanted to show below the percentage of instances in each of the experiments done to see the distribution of data elements in each of the clusters in the whole datasets.

The following table showed that the three experimentation were having varying proportions of instances in each clusters elements. But as we could see from the table below, the third experiment showed many clusters which were somehow proportional to each cluster instances and it provided with more useful clusters for discussion.

Cluster number	Experiment 1(K=9)		Experiment 2(K=10)		Experiment 3(K=11)	
	Number of instances for each cluster	%of instances involved	Number of instances for each cluster	%of instances involved	Number of instances for each cluster	%of instances involved
C0	1311	15%	1256	14%	1257	14%
C1	1180	13%	886	10%	887	10%
C2	826	9%	816	9%	798	9%
C3	1072	12%	865	10%	773	9%
C4	1166	13%	787	9%	683	8%
C5	561	6%	510	6%	344	4%
C6	1246	14%	1228	14%	1135	13%
C7	909	10%	820	9%	767	9%
C8	525	6%	495	6%	480	5%
C9	-	-	1133	13%	1110	13%
C10	-	-	-	-	562	6%

Table11: cluster result of the three experimentations for cluster instance distribution

At this step it was possible to decide and discus about the chosen experiments which was experiment 3. The simple-KMeans clustering algorithm run information would be presented below.

=== Run information ===

Scheme: weka.clusterers.SimpleKMeans -N 11 -A "weka.core.EuclideanDistance -R first-last" -I 500 -S 10

Relation: Combined datasets of IDSR and PHEM data for simple-KMeans algorithm

Instances: 8796

Attributes: 9

RegionName

ZoneName

Year

Month

MalariaOutbreak

TyphoidFever Outbreak

RelapsingFeverOutbreak

EpidemicTyphusOutbreak

MeningitOutbreak

Test mode: evaluate on training data

=== Model and evaluation on training set ===

kMeans

======

Number of iterations: 11

Within cluster sum of squared errors: 26503.284507283723

Missing values globally replaced with mean/mode

Cluster centroids:

Attribute	Full data	Cluster #										
		0	1	2	3	4	5	6	7	8	9	10
RegionName	**SNNPR**	**SNNPR**	**Tigray**	**Oromia**	**Addis Ababa**	**Oromia**	**Addis Ababa**	**Tigray**	**Oromia**	**Harari**	**Oromia**	**Addis Ababa**
ZoneName	**Central Tigray**	**Segen**	**South Tigray**	**Arsi**	**Arada**	**East Hararge**	**Kolfe Keranio**	**Central Tigray**	**West Shoa**	**Harari**	**West Hararge**	**Yekatit Hospital**
Year	2011	2011	2011	2010	2011	2005	2009	2010	2006	2011	2007	2008
Month	Mar	Jan	May	Sep	Jun	May	Jan	Jun	Apr	Jul	Oct	Dec
MalariaOutbreak	92%	98%	98%	100%	93%	99%	43%	99%	99%	63%	97%	70%
TyphoidFever Outbreak	76%	98%	73%	100%	99%	81%	97%	26%	96%	15%	83%	61%
RelapsingFeverOutbreak	33%	30%	4%	87%	72%	31%	76%	3%	66%	63%	11%	13%
EpidemicTyphusOutbreak	26%	25%	2%	87%	95%	3%	86%	2%	11%	0%	5%	12%
MeningitOutbreak	19%	20%	4%	58%	15%	43%	9%	5%	32%	10%	5%	7%
Total number of cluster instances involved	**8796**	**1257**	**887**	**798**	**773**	**683**	**344**	**1135**	**767**	**480**	**1110**	**562**
Percentage of instances	**100%**	**14%**	**10%**	**9%**	**9%**	**8%**	**4%**	**13%**	**9%**	**5%**	**13%**	**6%**

Table12: clustering with simple K-means algorithm run information results for better visualization purpose for experiment3

The above table showed that the 11 clusters were generated from the run information of Weka data mining tools and provided with detailed information about each of all the clusters and presenting with the percentage of instances as well as the percentage of specific disease occurrences in each of the regions which were mined with the simple-KMeans algorithm within the specified period of time. The proportion of disease occurrences therefore was explained for each disease below in the discussion. The percentage only showed us the occurrences of a specific disease to that specific cluster in place and time setting distribution of the results.

Discussion of Clusters from disease perspective:

In most of the clusters we could see the occurrence of malaria was more in every cluster with an average of more than 92% from the total instances most of the time. But only in areas like Kolfe Keranio in Addis Ababa region with 43%, Harari zone in Harari region with 63% and finally Yekatit hospital in Addis Ababa with 70%, the occurrence was less relatively.

Another disease case like typhoid fever outbreaks were more prevailing in most of the clusters with 76% an average of the instances most of the time. But in some clusters like south Tigray with 73%, central Tigray with 26%, Harari zone with 15% and finally Yekatit hospital in Addis Ababa with 61%, the occurrence was less relatively.

The prevalence of outbreak occurrences of another disease case called relapsing fever outbreaks were less than the total average of 33% of most of the time. But in some clusters like Arsi zone in Oromia region with 83%, Arada sub city in Addis Ababa region with 72%, Kolfe Keranio with 76%, west Shoa zone in Oromia region 66%, and in Harari zone with 63% the occurrence.

The other disease outbreak cases was the occurrence of the Epidemic disease cases which were occurring with a total average of 26% most of the time. But the prevalence of this disease outbreak occurrences in some places were high like Arsi zone in Oromia region with 87%, Arada sub city in Addis Ababa Region with 95%, and Kolfe Keranio sub city in Addis Ababa with 86%.

Lastly, meningitis disease outbreak occurrences were found with the total average of 19%. But the prevalence of this disease outbreak occurrences in some places were high like Segen zone in SNNPR region with 20%, Arsi zone in Oromia Region with 58%, East Harargie in

Oromia region with 43%, and west Shoa in Oromia with 32% but in other places the occurrence was less.

As mentioned above the experimentation3 has also produced more clusters with better proportions of instances found within it. Totally there were more clusters gained from this experiment with better instance proportions. For instance: - C0, C1, C2, C3, C4, C6, C7, and C9 others cluster with less proportion of instances like: - C5, C8, and C10. All cluster elements would be discussed below one by one for better investigation. Cluster selection activities were simply done by the numbers of instances found within the clusters. Therefore; the researcher decided to analyze experiment 3 run information from the simple K-means clustering data mining algorithm.

The summary of experimentation results for each cluster

The researcher also wanted to discus about each cluster elements below for better understanding and description purpose:-

Cluster0: In this cluster SNNPR region and Segen Zone in the year of 2011 and in the month of January, the occurrence of disease outbreak cases with the following proportion like: Malaria with 98%, Typhoid Fever with 98%, Epidemic Typhus with 30%, Relapsing Fever with 25%, and finally Meningitis with 20% proportion from 1257 total number of instances 14% of the total instances involved in the experimentation.

Cluster1: In this cluster the region Tigray and South Tigray Zone in the year of 2011 and in the month of May, the occurrence of disease outbreak cases with the following proportion like: Malaria with 98%, Typhoid Fever with 73%, Epidemic Typhus with 4%, Relapsing Fever with 2%, and finally Meningitis with 4% proportion from 887 total numbers of instances or 10% of the total instances involved in the experimentation.

Cluster 2: In this cluster the region Oromia and Arsi Zone in the year of 2010 and in the month of September, the occurrence of disease outbreak cases with the following proportion like: Malaria with 100%, Typhoid Fever with 100%, Epidemic Typhus with 87%, Relapsing Fever with 87%, and finally Meningitis with 58% proportion from 798 total numbers of instances or 9% of the total instances involved in the experimentation.

Cluster3: In this cluster the region Addis Ababa and Arada sub-city Zone in the year of 2011 and in the month of June, the occurrence of disease outbreak cases with the following proportion like: Malaria with 93%, Typhoid Fever with 99%, Epidemic Typhus with 72%,

Relapsing Fever with 95%, and finally Meningitis with 15% proportion from 773 total numbers of instances or 9% of the total instances involved in the experimentation.

Cluster4: In this cluster the region Oromia and East Hararge Zone in the year of 2005 and in the month of May, the occurrence of disease outbreak cases with the following proportion like: Malaria with 99%, Typhoid Fever with 81%, Epidemic Typhus with 31%, Relapsing Fever with 3%, and finally Meningitis with 43% proportion from 683 total numbers of instances or 8% of the total instances involved in the experimentation.

Cluster5: In this cluster Addis Ababa region and Kolfe Keranio Zone in the year of 2009 and in the month of January, the occurrence of disease outbreak cases with the following proportion like: Malaria with 43%, Typhoid Fever with 97%, Epidemic Typhus with 76%, Relapsing Fever with 86%, and finally Meningitis with 9% proportion from 344 total numbers of instances or 4% of the total instances involved in the experimentation.

Cluster6: In this cluster the Tigray region and Central Tigray Zone in the year of 2010 and in the month of June, the occurrence of disease outbreak cases with the following proportion like: Malaria with 99%, Typhoid Fever with 26%, Epidemic Typhus with 3%, Relapsing Fever with 2%, and finally Meningitis with 5% proportion from 1135 total numbers of instances or 13% of the total instances involved in the experimentation.

Cluster7: In this cluster the Oromia region and West Shoa Zone in the year of 2006 and in the month of April, the occurrence of disease outbreak cases with the following proportion like: Malaria with 99%, Typhoid Fever with 96%, Epidemic Typhus with 66%, Relapsing Fever with 11%, and finally Meningitis with 32% proportion from 767 total numbers of instances or 9% of the total instances involved in the experimentation.

Cluster8: In this cluster the Harari region and Harari Zone in the year of 2011 and in the month of July, the occurrence of disease outbreak cases with the following proportion like: Malaria with 63%, Typhoid Fever with 15%, Epidemic Typhus with 63%, Relapsing Fever with 0%, and finally Meningitis with 10% proportion from 480 total numbers of instances or 5% of the total instances involved in the experimentation.

Cluster9: In this cluster the Oromia region and West Hararge Zone in the year of 2007 and in the month of October, the occurrence of disease outbreak cases with the following proportion like: Malaria with 97%, Typhoid Fever with 83%, Epidemic Typhus with 11%, Relapsing Fever with 5%, and finally Meningitis with 5% proportion from 1110 total numbers of instances or 13% of the total instances involved in the experimentation.

Cluster10: In this cluster the Addis Ababa region and Yekatit Hospital as a zone level in the year of 2008 and in the month of December, the occurrence of disease outbreak cases with the following proportion like: Malaria with 70%, Typhoid Fever with 61%, Epidemic Typhus with 13%, Relapsing Fever with 12%, and finally Meningitis with 5% proportion from 562 total numbers of instances or 6% of the total instances involved in the experimentation.

The magnitude of disease case occurrences were analyzed from the point of view of time and place settings. Therefore the researcher had shown the more prevailing area of the country and the occurrence of disease cases with respect to the datasets which were taken and shown the segmentation with the 9 years time datasets. Here the most importance thing was to see the previous trends of disease occurrence and the surveillance improvement in time.

As per the objectives of the research, the researcher showed the more occurrences of disease cases with respect to time and place. Better results and interesting clusters were observed from the experiment. The research showed that data mining was applicable to the outbreak surveillance and response system of Ethiopia to enhance the betterments of the system strength with very low quality data in the system. The results was also interesting to develop a new hypothesis and do a better research on the field of data mining for disease outbreak cases in the country and providing a measurable plan of action for the decision makers to future preparedness and alarm on that specific area to see whether the occurrence of disease cases were deceasing or increasing with time.

As shown on the table above the year 2011 had pointed out more prevailing clusters since it appeared 4 times from the total clusters. Except the years 2004 and 2012 which didn't appear in any cluster, all the other years in the dataset appeared exactly one each. Therefore, the surveillance and response system had to check the prevalence of diseases cases against their plan of actions on that specified areas and time. As it was mentioned in the limitation parts of this research, here what should be noted was that the research was done only with the available datasets which were collected or reported from different places of the country within the 9 years of time from 2004-2012G.C. Therefore, all districts from all regions might not send all the expected reports on timely bases. The result showed only for the data involved for this research work.

5.1.4. Prediction MD for Epidemic Typhus disease

Experimentation for classification:

The Classification algorithm to classify disease cases according to the current and future prevalence of the disease occurrences to a certain area and within a certain seasonal time variation were investigated here in this part. Here the researcher wanted only to predict the occurrence of disease cases to show the future expected occurrence of cases to a certain area for simple planning and preparedness activities and to strength the surveillance system in that suspected areas.

As mentioned in the methodology part of this researcher work, the researcher used to classify disease cases by the use of decision tree J48 and naive bayes classification algorithm. Since the former surveillance system was reengineered by the new BPR national work reformation program for better surveillance system establishment. Therefore, only the new PHEM dataset was selected and analyzed for the purpose of classification because of the discussion with the domain experts. After the preprocessing phase of the dataset had a total of 4093 records and seven total attributes (*Region, Zone, Woreda, Year, Month, Epidemic Weeks)* and the disease cases called *Epidemic typhus*. But here the researcher discarded the *Year* and *EpidemicWeek* attributes since the years were already passed and the epidemic week attribute were having lower value of information gain.

As mentioned above the attributes were taken from the time and place setting of the dataset variables, one attribute had inheritance from the other attribute. The time attribute such as the Years are already passed as mentioned before then it was dropped out from the datasets because the prediction was done for the future occurrence of disease cases. From the information gain values of attribute selection process, the researcher obtained the results and described below in table13. As mentioned above for prediction purpose, the researcher used the attribute evaluator feature selector called ***Gain Ratio feature evaluator*** by the search method of ***Attribute Ranker***. The attributes are ranked according to the gain ratio of each attribute with respect to the class attribute. The Evaluation mode was performed to all the datasets.

The following table showed the description of some terms in the run information window of the attribute selection parts on Weka3.6 software tool for the process of attribute selection.

Parameters or terms	Description
Weka	The Data mining tool
Attributeselection	Attribute selection process
GainRatioAttributeEval	Evaluates the worth of an attribute by measuring the gain ratio with respect to the class.
Ranker	The attributes are selected according to the ranks to the gain ratio
Filters	The preprocessing step to change the numeric representation to nominal values
Unsupervised	The learning type
NumericToNominal-Rfirst-last	Changing the attribute values from numeric to nominal by ranking starting from the first attribute to the last attribute.
-T	The minimum threshold Gain Ratio given to each attribute.
-N	Ranks to be given for each attribute in numeric form

Table13: parameter description for attribute selecter or feature generator window in Weka

5.1.4.1. Model preparation For Prediction

The following table showed the Gain Ratio of attributes and the ranks they have to the class Epidemic Typhus from preprocessing attribute selection parts of this research. As could be seen from the table below month have a small value, it meant the prediction was less on the month attribute than the other place oriented attributes like *RegionName, ZoneName,* and *WoredaName*. But most importantly the region was the most important significant attribute and the root of the decision tree classifier since it had the greater value of gain ratio. The following table showed the importance of attributes with the ranking list against the gain ratio they had.

Attribute name	Gain Ratio value	Rank
RegionName	*0.110265*	1
ZoneName	*0.084928*	2
WoredaName	*0.078706*	3
Month	*0.000375*	4
Epidemic typhus	Class	Class

Table14: Gain ratio for best attributes selection process with their ranking lists

The researcher had done experimentation to develop a prediction model by the use of Naïve Bayes classifier and Decision Tree classification of J48 algorithm. The test option evaluation criteria were set as 10 folds cross-validation as mentioned before. The summary results of each algorithm were shown below and the comparison was done following the summary results of the two classifications algorithms on the table. The comparison between the results of the two classification data mining algorithms were done for Epidemic Typhus disease cases prediction modeling on place and time setting only. As could be seen from each of the results the values of disease occurrences were changed from the value of 1 to YES for the occurrence and from 0 to NO for nonoccurrence respectively for better and easy understanding for tree traversal on the run information of the results.

=== Summary Results of a Decision Tree J48 algorithm===

```
Time taken to build model: 0.08 seconds

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

Correctly Classified Instances        3588               87.6619 %
Incorrectly Classified Instances       505               12.3381 %
Kappa statistic                          0.7143
Mean absolute error                      0.166
Root mean squared error                  0.294
Relative absolute error                 38.1998 %
Root relative squared error             63.0666 %
Total Number of Instances             4093

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.795     0.085      0.814     0.795     0.804      0.935    Yes
                 0.915     0.205      0.905     0.915     0.91       0.935    No
Weighted Avg.    0.877     0.167      0.876     0.877     0.876      0.935

=== Confusion Matrix ===

    a    b   <-- classified as
 1038  268 |    a = Yes
  237 2550 |    b = No
```

=== Summary Results of the Naïve Bayes Classifier ===

```
Time taken to build model: 0.02 seconds

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

Correctly Classified Instances        3429               83.7772 %
Incorrectly Classified Instances       664               16.2228 %
Kappa statistic                          0.6589
Mean absolute error                      0.1708
Root mean squared error                  0.3401
Relative absolute error                 39.3044 %
Root relative squared error             72.9711 %
Total Number of Instances             4093

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.923     0.202      0.681     0.923     0.784      0.939    Yes
                 0.798     0.077      0.957     0.798     0.87       0.939    No
Weighted Avg.    0.838     0.117      0.869     0.838     0.843      0.939

=== Confusion Matrix ===

    a    b   <-- classified as
 1206  100 |    a = Yes
  564 2223 |    b = No
```

Then the researcher did the comparison between the classification performances of the two classifications algorithm results with the most important measures in the field of health related researches for disease predictions. The results were analyzed and shown below on the table for better comparison of the two algorithms.

	Measurements	Epidemic Typhus disease Case occurrence	Naïve Bayes	Decision Tree with J48 Algorithm	Remark
1	TP Rate	No	0.798	0.915	
		Yes	0.923	0.795	
		Average	**0.838**	**0.877**	
2	ROC Area	No	0.939	0.935	
		Yes	0.939	0.935	
		Average	**0.939**	**0.935**	
4	Sensitivity		**0.6814**	**0.8141**	
4	Specificity		**0.9570**	**0.9049**	

Table15: Measurement evaluation for the two decision tree classifiers for comparison

In the table above we could see that the occurrences of disease cases were measured by four important prediction performance measurements. The researcher wanted to use the best

classification algorithm to predict the occurrence of disease cases in a certain area and with some time period. The first performance evaluator was the average true positive (TP) rate to evaluate the performance of the two classification algorithm (i.e. Decision Tree with J48 classifier and the Naïve Bayes classifier). From the run information summary results of Weka3.6 tool, the average TP rates were 83.8% and 87.7% for Naïve Bayes and Decision Tree J48 classifier algorithm respectively. The better the average TP rate the better the performance. The second one was the ROC area which is the combination of the sensitivity and specificity measure. The ROC area average results of the two algorithms were 93.9% for Naïve Bayes and 93.5% for Decision Tree classifiers. Here the two algorithms had better performance which were both greater than 90% and nearly equal.

The researcher also tried to see the sensitivity and the specificity of the performance measurement for the occurrence and nonoccurrence of disease cases. Because these two measures were the most important model measurement techniques in the health related research. The result showed that the sensitivity of Naïve Bayes algorithm was 68.14% and that of the Decision tree classifier was 81.41%. Since the sensitivity of the test shows that the true occurrences of disease cases were to be classified as an occurrence. From this test, we could understand that decision tree had a better performance than Naïve Bayes classifier to classify the occurrence of disease cases. Finally the researcher tried to see the specificity of the two algorithms. The specificity of the test showed that the non occurrence of disease cases are classifies as non occurrence. The specificity of the test results were 95.7% and 90.49% for Naïve Bayes and decision tree classifiers respectively. Here both had high specificity but the Naïve Bayes classifier had better value.

As per the objectives of the researcher work, the researcher wanted to classify the disease occurrences than the nonoccurrence's for planning and taking a measurable actions to decision making process in more affected area of the country with the time slices. Then the researcher used the sensitivity, the TP rate and the ROC area for performance measurement to select the best performing algorithm. Therefore; Decision tree was selected by having a better performance than Naïve Bayes classifier.

It was also true that the Decision true classifier had classified instances correctly with 87.6619% percentage or accuracy rate but that of the Naïve Bayes classifier was with

83.7772% only. Finally, it was better to misclassify the occurrence of a disease cases than the nonoccurrence of cases. That was why in our investigation we took into consideration the recognition rate of disease cases occurrences than the nonoccurrence's. Since the Decision tree classifier with J48 algorithms were selected to develop a predictive model for the occurrence of Epidemic Typhus disease cases, then the results were discussed as follows from the run information of J48 classifier at Appendix 3.

Discussion of Decision Tree J48 algorithm results

In decision tree classifier the results were analyzed in the form of IF-THEN rules. Since the research was for educational purpose, all the generated rules were not explained here. The success ratio (rate) or the confidence of disease cases occurrence and nonoccurrence were calculated from the number after the leaf of the tree. As showed in appendix 3, the occurrence of disease cases were represented by Yes and the nonoccurrence were by No values. Success ratio is the ratio of correctly classified diseases cases with the total number of disease cases in that area.

Some rules were discussed below:

1. *RegionName = Addis Ababa*

    | *WoredaName = Kolfe Keraniyo: Yes (33.0)*

The above rule implies that:

The rule stated that, out of the total numbers of disease cases classified by the decision tree J48 classifier all the records i.e. the 33 records in the datasets were shown the disease cases occurrences, therefore the success ratio=33/33*100=100%. Likewise other rules would be calculated and the results were shown below the rules.

If region =Addis Ababa and WoredaName = Kolfe Keraniyo Then Epidemic Typhus occur with 100% success ratio. Yes implies Disease occurrence and 33.0 implies the total number of occurrences classified correctly.

2. *RegionName = Addis Ababa*

    |*WoredaName = Akaki Kaliti: Yes (31.0/1.0)*

The above rule implies that:

If region =Addis Ababa and WoredaName = Akaki Kaliti Then Epidemic Typhus occur with 96.88% success ratio. Yes implies Disease occurrence and 31.0 implies the total number of occurrences classified correctly and 1.0 implies incorrectly classified instances.

3. *RegionName = Addis Ababa*

    | *WoredaName = Bole: Yes (32.0/1.0)*

The above rule implies that:

If region =Addis Ababa and WoredaName = Bole Then Epidemic Typhus occur with 96.97% success ratio. Yes implies Disease occurrence and 32.0 implies the total number of disease occurrences classified correctly and 1.0 implies incorrectly classified disease occurrences.

4. *RegionName = Oromia*

    | *ZoneName = East Shewa: Yes (33.0/2.0)*

The above rule implies that:

If region = Oromia and ZoneName = East Shoa Then Epidemic Typhus occur with 94.29% success ratio.

5. *RegionName = Oromia*

    | *ZoneName = Adama Special Zone: No (29.0/6.0)*

The above rule implies that:

If region = Oromia and ZoneName = Adama Special Zone Then Epidemic Typhus do not occur with 82.86% success ratio.

6. *RegionName = Oromia*

    | *ZoneName = Bishoftu Town: No (29.0)*

If region = Oromia and ZoneName = Adama Special Zone Then Epidemic Typhus do not occur with 100% success ratio.

7. *RegionName = Oromia*

    | *ZoneName = Finfine Zuria: Yes (27.0/12.0)*

If region = Oromia and ZoneName = Adama Special Zone Then Epidemic Typhus occur with 69.23% success ratio.

8. *RegionName = Oromia*

    | *ZoneName = Jimma Spe Town: No (30.0/2.0)*

If region = Oromia and ZoneName = Adama Special Zone Then Epidemic Typhus do occur with 93.75% success ratio.

9. *RegionName = Oromia*

| *ZoneName = East Wellega: Yes (31.0/15.0)*

If region = Oromia and ZoneName = Adama Special Zone Then Epidemic Typhus occur with 67.39% success ratio.

10. *RegionName = Tigray: No (1180.0/26.0)*

If region = Tigray Then Epidemic Typhus will occur with 97.84% success ratio.

11. *RegionName = Gambella: No (26.0/2.0)*

If region = Gambella Then Epidemic Typhus will occur with 92.86% success ratio.

12. *RegionName = SNNPR*

| | ZoneName = Alaba: No (32.0/10.0)

If region = SNNPR and ZoneName = Alaba Then Epidemic Typhus do not occur with 96.97% success ratio.

13. *RegionName = SNNPR*

| | *ZoneName = Bench Maji: Yes (31.0/7.0)*

If region = SNNPR and ZoneName = Bench Maji Then Epidemic Typhus occur with 81.58% success ratio.

14. *RegionName = SNNPR*

| | *ZoneName = Dawuro: No (32.0/12.0)*

If region = SNNPR and ZoneName = Dawuro Then Epidemic Typhus do not occur with 72.73% success ratio.

15. *RegionName = SNNPR*

| | *ZoneName = Gamo Gofa: Yes (32.0/2.0)*

If region = SNNPR and ZoneName = Dawuro Then Epidemic Typhus do not occur with 94.12% success ratio.

16. *RegionName = SNNPR*

| | *ZoneName = Gedeo: Yes (32.0/11.0)*

If region = SNNPR and ZoneName = Dawuro Then Epidemic Typhus do not occur with 74.42% success ratio.

17. RegionName = SNNPR

| | ZoneName = Wolayita: No (32.0/2.0)

If region = SNNPR and ZoneName = Wolayita Then Epidemic Typhus do not occur with 94.12% success ratio.

18. RegionName = SNNPR

| WoredaName = Burgi: No (22.0)

If region = SNNPR and ZoneName = Burgi Then Epidemic Typhus do not occur with 100% success ratio.

19. RegionName = SNNPR

| WoredaName = Derashe: No (22.0/2.0)

If region = SNNPR and ZoneName = Derashe Then Epidemic Typhus do not occur with 91.67% success ratio.

20. RegionName = Dire Dawa: No (63.0/2.0)

If region = Dire Dawa Then Epidemic Typhus do not occur with 96.92% success ratio.

21. RegionName = Harari: No (161.0/1.0)

If region = Harari Then Epidemic Typhus do not occur with 99.38% success ratio.

22. RegionName = Amhara

| ZoneName = North Gonder: No (24.0/12.0)

If region = Amhara and ZoneName = North Gonder Then Epidemic Typhus do not occur with 66.67% success ratio.

23. RegionName = Amhara

| ZoneName = North Wollo: No (26.0/13.0)

If region = Amhara and ZoneName = North Wollo Then Epidemic Typhus do not occur with 66.67% success ratio.

24. RegionName = Amhara

| ZoneName = South Gonder: No (26.0/5.0)

If region = Amhara and ZoneName = South Gonder Then Epidemic Typhus will not occur with 83.87% success ratio.

25. RegionName = Amhara

| *ZoneName = South Wollo: Yes (28.0/3.0)*

If region = Amhara and ZoneName = South Wollo Then Epidemic Typhus occur with 90.32% success ratio.

26. RegionName = Amhara

| *ZoneName = West Gojjam: Yes (25.0/5.0)*

If region = Amhara and ZoneName = West Gojjam Then Epidemic Typhus occur with 83.33% success ratio.

27. RegionName = Amhara

| *ZoneName = Wag Himra: No (22.0/1.0)*

If region = Amhara and ZoneName = Wag Himra Then Epidemic Typhus occur with 95.65% success ratio.

28. RegionName = Amhara

| *ZoneName = Bahir Dar Liyu: No (23.0/6.0)*

If region = Amhara and ZoneName = Bahir Dar Liyu Then Epidemic Typhus occur with 79.31% success ratio.

29. RegionName = Amhara

| *ZoneName = Oromiya: Yes (25.0/10.0)*

If region = Amhara and ZoneName = Oromiya Then Epidemic Typhus occur with 71.43% success ratio.

30. RegionName = Afar: No (436.0/20.0)

If region = Afar Then Epidemic Typhus will not occur with 95.61% success ratio.

31. RegionName = B-Gumuz

| *ZoneName = Assossa: No (25.0/3.0)*

If region = B-Gumuz and ZoneName = Assossa Then Epidemic Typhus do not occur with 89.29% success ratio.

Summary of Epidemic Typhus disease Classification model:

However the above results were interesting and acceptable with the domain experts, prediction with less numbers of variables were not providing better prediction model. It was because of the importance of data mining which is more applicable and powerful for the multidimensional and huge amount of data in the datasets. With the use of quality data for better model development, it was known that, there should be a dependent and one or more independent variables to have the probability of getting a better model. In the case of this research, the dependent was disease cases and the independent was time and place variation. However, all important factors or determinants for diseases occurrences had to be included in the data collection format for better prediction model development and to enhance the probability of getting a better predictor.

Therefore the sector has to have a very sophisticated wired and wireless communication technology to enhance the ability of the surveillance system to function appropriately and correctly for timely and complete data collection and result dissemination techniques. Due to the limited use of appropriate infrastructure, better use of technology and skilled personnel, early identification of the determinants of disease occurrences, and the lack of quality data accumulated in the sector influenced the functionality of the system. The researcher has set recommendation about the future activities to the sector for better data quality to enhance the prediction performance with data mining applications.

5.2. General Summary of experimentation results:

The aim of this research work was to show the applicability of data mining techniques for the development of descriptive and predictive model to disease outbreak surveillance datasets in Ethiopia. To do that the three data mining applications such as classification, clustering and association rules mining were undertaken to explore the important applications to the datasets of the PHEM sector from different perspectives. A total of 18600 records were collected and assessed from the data store of the surveillance system from the year 2004-2012G.C. After the preprocessing phase of knowledge discovery in databases of data mining application a total of 8796 records were prepared for data mining algorithms. From the total records filtered and prepared for model preparation 4703 records were taken from the IDSR system data store and the remaining 4093 records were taken from the PHEM data store from the year 2004-2008G.C. and 2009-2012G.C. respectively.

The researcher analyzed two classification algorithms for the prediction of Epidemic typhus disease cases with decision tree J48 classifiers and Naïve Bayes classifiers. Using the two selected classification algorithm was due to the researcher's familiarity with those algorithms and ease of implementation. Finally the more performing algorithm has been taken for model development. From the results of the experiments done decision tree algorithm had a better performance to classify the disease cases in place and time setting. The accuracy rate of correctly classifying the Epidemic Typhus disease cases by the use of decision tree J48 algorithm was 87.44% whereas with Naïve Bayes classifier was 83.70%. The sensitivity and specificity test was also done for the two classifiers. Decision tree showed a better performance than Naïve Bayes algorithm for Epidemic Typhus prediction. Since the researcher tried to classify disease cases with less numbers of attributes on the defined algorithms, the probability of getting a better predictive model might be constrained with the given numbers of attributes and the importance of unseen variables. Though the two algorithms has shown hopeful results, the decision tree data mining technique was found to be more applicable with this experimentation for the prediction of Epidemic Typhus cases from the point of view of the attributes taken only from place and time setting.

The researcher also attempted to analyze the application of association rule mining to find some sort of correlation or patters among disease cases of the surveillance data. The attributes

were selected only from the disease cases for the occurrence and nonoccurrence, which were collected in time and place bases. Here, Apriori association rule mining algorithm was run to find interesting patterns among the co-occurrence of disease cases which were correlated together.

The attributes selected for association rules mining were *MalariaOutbreak, TyphoidFeverOutbreak, MeningitisOutbreak, EpidemicTyphusOutbreak* and *RelapsingFevereOutbreak.*

Pattern discovery in association rule mining for frequent itemsets were done from only the point of view of disease cases. A brief Analysis was done on the results of Apriori algorithm for all the two datasets and the integrated or combined datasets. Here the three datasets which were IDSR system, the PHEM and the combinations of the two datasets were involved in Apriori algorithm of association rule mining applications after the data preparation phase of KDD. The researcher used 20% for the minimum support and 90% for minimum confidence threshold before the application of the mining algorithm. In addition to this the researcher has defined the minimum support level to be small, because use of small minimum support is important to discover interesting rules for rarely or infrequently occurring disease cases.

From the experimentation result analysis, the researcher had found a lot of interesting and important rules. Even though the research was for educational purpose, the rules which were generated from Apriori algorithm had a value to the domain experts for comparison purpose with the existing knowledge and decision making activities. Due to time constraint and timely data collection problem from the sector, the researcher didn't do more experimentation and analyzing more association rule mining algorithms for better results.

The researcher also made another attempt with clustering algorithm for disease segmentation on the bases of place and time setting. The simple K-means algorithm was run to cluster the disease outbreak cases with respect to prevalence of disease in time and place of occurrence and nonoccurrence. The researcher took the combined (integrated) datasets for cluster analysis with the total numbers of 8796 records and 9 attributes. This technique has also shown the occurrences or nonoccurrence's of disease cases with a certain time and location in some extent. Clustering algorithm was done for the combined datasets since; the algorithm showed the grouping of disease cases with respect to the total years of time and place.

Here the challenge was to limit the number of clusters to be produced from this algorithm. But the researcher with the discussion to the domain experts decided to cluster disease cases with respect to the total numbers of years (nine clusters for the nine years dataset) from which the data was taken for the research work. Another option was raised to segment the disease cases to group the disease cases by taking the value of K to be 11. It was because of the numbers of regions in the country. Here, what the researcher has found was that clustering data mining application has a significant role on the division of places and times with the occurrence of disease cases to a better hypothesis development activity. From the experimentation results, the prevalence of disease cases was shown with time and place setting. The more affected and the less affected area of the country was also identified with the clustering techniques. In general data mining techniques were important and applicable in the classification, clustering and association rules model development for emerging and reemerging disease outbreak cases. The results of the research, apart from its education purpose, were also used for the planning, preparedness, decision making, and disease control and prevention activities of the domain experts. For instance the classification algorithm has shown the most affected areas of the country like *zones* and *Woredas* by *Epidemic Typhus* disease to domain experts for making measurable action.

The association rule also has shown the correlation of two or more disease to occur together. Therefore the domain experts will prepare all the necessary things, when there was an occurrence of a given disease, then the experts would also expected to prepare for the other disease case. For example if there existed an occurrence of Epidemic Typhus disease then the occurrence of typhoid fever disease cases were with 99% confidence level and 31.6% support level. The cluster analysis was also used for hypothesis development purpose, for the future model preparation activities.

With the better performing classification algorithm, decision tree J48 classifier with the performance accuracy rate of 87.66% for the processed dataset showed that, some regions were less susceptible than others like Afar region with 95.61%, Dire Dawa with 96.92%, Harari with 99.38%, Tigray with 97.84%, Somali with 100% and Gambella with 92.86% with less or no occurrence of incidence of the Epidemic Typhus disease. However the other regions of the country like *Oromia, Addis Ababa, Amhara, SNNPR,* and *B-Gumuz,* were having important differences with the variation in *Zones and Woredas* levels of those regions

for the occurrence and nonoccurrence of Epidemic Typhus disease. It was also stated in the problem statements of this researcher work that many attributes were needed. That was why the researcher had faced a challenging problem to develop a better model with limited numbers of attributes in the datasets.

Finally, as per the purpose of the research, it only tried to show how data mining techniques were easily implemented for description and prediction of disease outbreak cases with a better accuracy for planning, preparedness, detection, segmentation, prediction and association for outbreak disease surveillance and response system. The sector has also importantly engaged to prepare the stronger database system or data warehouses by including all the determinants of disease outbreak occurrences for better outbreak prediction model development by the use of data mining applications.

CHAPTER Six

6, CONCLUSIONS AND RECOMMENDATIONS

6.1. Conclusions

- The PHEM surveillance and response system were experienced with a serious shortage of quality data and data reporting mechanisms to perform data mining applications for model development. As well the numbers of attributes (lack of unseen variables) in the datasets of the PHEM sector were not enough.

- The absence of quality data in the PHEM sector had adversely affected the quality of predictive and descriptive model development for emerging and reemerging disease outbreak cases in the country.

- It was proved that data collection; collation, analysis, interpretation and presentation in the PHEM outbreak surveillance and response system were having a greater effect in strengthening its performance with a new model development.
- Disease prediction or classification has shown that some places of the country were more vulnerable than others for the incidence of Epidemic typhus disease cases.

- Decision tree J48 algorithm was better than Naïve Bayes classifier for the data at hand to prepare a prediction model with respect to time and place setting.

- There was an association among the disease cases to occur together in the PHEM sector dataset surveillance and response system of the country.
- Association rule mining with Apriori algorithm was important to show the real association of disease cases with time and place based transaction.

- The occurrence of a certain disease outbreak cases was associated with the occurrences of some other disease cases. But here the disease case association didn't show the causality effects.

- Apriori algorithm couldn't detect the infrequently occurring disease cases. For instance some rear disease occurrences like meningitis were not detected by the use of Apriori data mining algorithms for association rule mining.

- Some disease cases were more frequent to occur in the country throughout all the regions than others like Malaria and typhoid fever. Moreover, the incidences of malaria and typhoid fever were endemic in some places.

- The simple K-Means clustering algorithm for Disease clustering showed that Malaria and Typhoid fever disease cases were more prevalent in the country than others.

- There were more useful clusters of disease cases in the PHEM surveillance dataset than the IDSR dataset. I.e. most of the prevalence of disease cases like malaria and typhoid fever indicated that some of the cases were not properly controlled and managed with the current PHEM surveillance system to alleviate its occurrences than the former IDSR datasets.

- The current PHEM surveillance system and the combination of the whole datasets of the country was better than the former IDSR system in providing interesting rules production with Apriori algorithm for the available dataset on hand.

6.2. Recommendation

- The researcher recommended that the PHEM sector should have a better data collection and data reporting mechanism and have a consistent data collection format for all concerned reporting sites and organizations for emerging and reemerging disease outbreak cases.

- The researcher also added that data collection formats must include variables which can determine the occurrence of disease outbreak cases such as environmental variables (temperature, humidity, latitude, rainfall, existence of vector), social variables (hygiene system, sanitation, feeding habit, income status), demographic data(age, sex), settlement condition (population density, nearness of health facilities, urban or rural setting, nomad or settled, outmigration and immigration), disease condition (specimen collection, result confirmation, natural history of the disease description, existence of vector) and the existence of other natural as well as manmade hazards(forest burning, war, flooding, earth quake, and so on).

- The researcher has pointed out that the accuracy rate of the model developed tends to increase the desirable level of accuracy with more and more trials of different combinations of the variables by performing more tests to evaluate and analyze it. Therefore, more time should be assigned while developing such a model in the future.

- The researcher has also recommended that the PHEM sector under MoH need to establish a better data reporting mechanism for quality data management system. For this, the FMoH should install a fast and effective communication technology with trained professionals. To do that every reporting site has to have at least a dialup internet connection or high speed telecommunication wireless device, mobile computing application and a laptop computer administered by skilled professionals.

- As stated in **[64],** due to the widespread application of GIS and GPS technology and the increasingly mature infrastructure for data collection, sharing, and integration, more and more research domains have gained access to high-quality geographic data

and created new ways to incorporate spatial-temporal information and analysis in various studies. Therefore, the researcher has recommended the PHEM sector should have a Web-based and wireless technologies, such as GIS and mobile computing applications which can enhance the performance of the outbreak surveillance and response system and the process of decision making activities by accessing quality data from anywhere in the country to implement it with a better machine learning or data mining algorithms.

- Since data quality is affected by different factors such as skilled health professionals and technical database managers as well as lack of cooperation between these parties to accommodate the transition from the existing surveillance system to the newly emerging communication technology based surveillance system, the PHEM sector should work on these issues.
- The researcher recommended that the PHEM outbreak surveillance and response reporting system of the country should have an effective and efficient timely follow up mechanism.

- The researcher has also recommended a study to be done with other data mining tools like rapid miner and Tanagra with a variety of algorithms to develop a model on both studied and unstudied disease outbreak cases for better performance evaluation measures.

- Another research can also be done by changing the type of algorithm like neural network or CART for prediction and Expected Maximization for clustering and even by changing the used parameters for the applied algorithm with different values.

- The researcher has also left a place to other researchers for a better disease outbreak cases association probability by including other outbreak disease cases which are found in the PHEM surveillance datasets and also see the association of unseen influential variables left under study because of their absence from the datasets.

- Further research should be conducted so as to see better association among the whole disease outbreak cases in the country which are occurring in the PHEM sector whereby algorithms such as predictiveApriori, Tertius and Filtered Associator and Apriori itself are involved.

- Another recommendation set on the Apriori algorithm was other algorithms which can perform more efficiently and effectively on numeric values should be applied. Since Apriori algorism can’t support continuous numeric data type.

Reference

1. Redmond, WA: Microsoft Corporation, "Epidemic." Microsoft® Encarta® 2009 [DVD]; 2008.
2. Mulugeta A., Communicable Disease Control, Ethiopian public health training initiatives, Hawassa University, Ethiopia, 2004.
3. Abera K., Ahmed A., An overview of environmental health status in Ethiopia with particular emphasis to its organization, drinking water and sanitation, A literature survey, Department of Community Health, Medical Faculty, Addis Ababa University, Ethiopia, 2005.
4. Krusche, OtterWasse, Introduction of Ecological Sanitation for Large Scale Housing Programs in Ethiopia, GTZ, 2007.
5. Tsedaye Bezabeh, 2009, Epidemiology and Disease Surveillance, FMOH/TUTAPE.
6. WHO regional office for Africa, WHO country Cooperation strategy 2008–2011, Ethiopia, India; 2009.
7. Grant R, Spring S, A multidisciplinary approach for the early detection and response to disease outbreaks, USA, Armed Forces Health Surveillance Center; 2011.
8. WHO, Asian pacific strategy for Emerging Diseases, Library Cataloguing in Publication Data; 2010.
9. Richard G. W, Reviewing Ethiopia's Health System Development, International Medical Community, JMAJ 52(4): 279–286, 2009.
10. Freeman R, Evolution of healthcare-associated infection surveillance in England: initiatives, implementation and opportunities for innovation; 2011.
11. Communicable Disease Prevention and Control (CDC), Communicable Disease Prevention and Control (Focus Area Profile) plan, Wisconsin Department of Health, *Healthiest Wisconsin*, 2010,
12. Patz J. A., Githeko A. K., McCart J. P., Hussein S, Confalonieri U., de Wet N., Climate change and infectious diseases.
13. U.S. Department Of Health And Human Services, Principles Of Epidemiology, 2nd Edition, An Introduction To Applied Epidemiology And Biostatistics, Centers For Disease Control And Prevention (CDC), Epidemiology Program Office, Public Health Practice Program Office, Atlanta, Georgia 30333, 2005
14. Institute of Environmental Science & Research, Disease Outbreak Manual, Porirua, New Zealand, 2002.
15. Ghosh A, Nath B, Multi-objective rule mining using genetic algorithms, Information Sciences 163 (2004) 123–133; 2004.
16. Kantardzic M, Data Mining: Concepts, Models, Methods, and Algorithms, University of Louisville, The Institute of Electrical and Electronics Engineers, Inc.; 2003.
17. Chakrabarti S, Cox E, Frank E, Güting RH, Han J, Jiang X, et al., Data Mining knows it all, Morgan Kaufmann, Elsevier Inc.; 2009.

18. Refaat M, Schneider M., Teorey T J., Witten I H. , Data Mining knows it all, Morgan Kaufmann, Elsevier Inc.; 2009.

19. Omari A, Data Mining for Retail Website Design and Enhanced Marketing, Inaugural-Dissertation, Mathematisch-Naturwissenschaftlichen Fakult¨at der Heinrich-Heine-Universit¨at D¨usseldorf vorgelegt von; 2008.

20. Fayyad U, Piatetsky-Shapiro G, and Smyth P, From Data Mining to Knowledge Discovery in Databases, American Association for Artificial Intelligence, (AI Magazine Volume 17 Number 3 ,1996); 1996.

21. Tayfun .En, Parallel Closet+ Algorithm for Finding Frequent Closed Itemsets, Masters' thesis, Middle East Technical University; 2009.

22. Cios K.J., Pedrycz W., Swiniarski, R.W., Kurgan, L.A., Data mining, A knowledge discovery approach, 2007, Springer . http://www.springer.com/978-0-387-33333-5

23. Jackson J., DATA MINING: A CONCEPTUAL OVERVIEW, Communications of the Association for Information Systems (Volume 8, 2002) 267-296, Management Science Department , University of South Carolina, 2002.

24. Ponce J., Karahoca A., Data Mining and Knowledge Discovery in Real Life Applications, Croatia; 2009.

25. Ranjan J, Nagar R, Ghaziabad, Applications Of Data Mining Techniques In Pharmaceutical Industry, Information Management And Technology Area, Institute Of Management Technology, Journal Of Theoretical And Applied Information Technology, Uttar Pradesh, India; 2007.

26. Han J. and Kamber M, Data Mining: Concepts and Techniques, 2nd Edition, The University of Illinois at Urbana-Champaign, Morgan Kaufmann; 2006.

27. Shmueli G, Patel N R., Bruce P C., Data Mining In Excel: Lecture Notes and Cases, USA; 2005.

28. Ma1 L, Tsui F, William HR., Wagner MM., Haobo Ma, A Framework for Infection Control Surveillance Using Association Rules, RODS Laboratory, Center of Biomedical Informatics and Intelligent Systems Program, University of Pittsburgh, Pittsburgh; 2003.

29. Sivanandam S N, Sumathi S, Madhavan P K, Data Mining Concepts, Tasks and Technique; 2006.

30. Vinnakota S, Lam N S., Knowledge Discovery From Mining The Spatial Associations Between Cancer Mortality And Socioeconomic Characteristics, Department Of Geography And Anthropology, Louisiana State University, Baton Rouge, La 70803, USA.

31. Jain S, Aalam M Afshar, Doja M N, K-MEANS CLUSTERING USING WEKA INTERFACE, The 4th National Conference, New Delhi, India; 2010.

32. Jain A K., Data Clustering: 50 Years beyond K-Means, Department Of Computer Science & Engineering, Michigan State University, USA; *2009*.

33. Larose D T., Discovering Knowledge in Data, an Introduction to Data Mining, A John Wiley & Sons, Inc., Publication; 2005.

34. Nasser S, Alkhaldi R, Vert G, Department, A Modified Fuzzy K-means Clustering using Expectation Maximization, Department of Computer Science and Engineering, 171, University of Nevada Reno, Reno NV 89557, USA.

35. Soni J, Ansari U, Sharma D, Predictive Data Mining for Medical Diagnosis: An Overview of Heart Disease Prediction, International Journal of Computer Applications (0975 – 8887), (Volume 17– No.8, 2011); 2011.

36. Shillabeer A, Roddick J F., Establishing a Lineage for Medical Knowledge Discovery, Carnegie Mellon University, Australia; 2007.

37. Mei Yin L A., Prediction Model for HlNl Disease, University Utara, Malaysia; 2011.

38. Srinivas K, Kavihta B R, Govrdhan A, Applications of Data Mining Techniques in Healthcare and Prediction of Heart Attacks, International Journal on Computer Science and Engineering, (Vol. 02, No. 02, 2010, 250-255); 2010.

39. Koh H C and Tan G, Data Mining Applications in Healthcare, Journal of Healthcare Information Management, Vol. 19, No. 2; 2005.

40. Amanuel Dibaba Bedane, Application of data mining techniques to predict household health seeking behavior: the case of butajira rural health project (BRHP), Master's thesis, Addis Ababa University, Ethiopia; 2010.

41. Chapman W W., Dowling J N., Ivanov O, Gesteland P H., Olszewski R T., Jeremy, et al., Evaluating Natural Language Processing Applications Applied to Outbreak and Disease Surveillance, Pittsburgh, USA; 2004.

42. Last M., Carel R, Barak D, Utilization of Data- Mining Techniques for Evaluation of Patterns of Asthma Drugs Use by Ambulatory Patients in a Large Health Maintenance Organization; 2007.

43. Wongstitwilairoong T, Gaywee J, Sirisophana N, Mason C J., Pavlin J A, Mining Pattern Model of Influenza Surveillance, Armed Forces Research Institute of Medical Sciences, Bangkok, Thailand; 2008.

44. Erman J, Arlitt M., Mahanti A, Traffic Classification Using Clustering Algorithms, University of Calgary, Canada; 2006.

45. Aher S B., Lobo L.M.R.J, A Comparative Study of Association Rule Algorithms for Course Recommender System in E-learning, International Journal of Computer Applications , 2012, *(0975 – 8887) Volume 39– No.1, February 2012,*

46. Az eve PJ and Jorge AM., Comparing Rule Measures for Predictive Association Rules, department of informatics, University of Minho and University of Porto , Portugal

47. Shawe-Taylor J, Diethe T, Analysis of Complex Data; 2009.

48. Soria D, Garibaldi J M., Ambrogi F., Biganzoli E. M., Ellis I. O., A 'non-parametric' version of the naive Bayes classifier, Knowledge-Based Systems24; 2011.

49. Bramer M, Principles of Data Mining, Springer-Verlag, University of Portsmouth,UK; 2007.

50. Chaudhary K, Papapanagiotou I, and Devetsikiotis M, Flow Classification Using Clustering And Association Rule Mining, Electrical and Computer Engineering, North Carolina State University, Raleigh, USA; 2010.

51. Reutemann P, Seewald A, Scuse D, WEKA Manual for Version 3-6-0; 2008.

52. Federal Ministry of Health (FMoH), Public Health Emergency Management Guideline, Ethiopian Public Health Institute, Addis Ababa, Ethiopia; 2009.

53. Garcia-Abreu A, Halperin W, Danel I, Public Health Surveillance Toolkit, a guide for busy task managers, World Bank, 2002.

54. Institute of Environmental Science & Research (ESR), Disease Outbreak Manual, Porirua, New Zealand; 2002.

55. WHO, Outbreak surveillance and response in humanitarian emergencies, Geneva, Switzerland; 2012.

56. Federal Democratic Republic Of Ethiopia Ministry Of Health(FDREMoH), Public Health Emergency Management Core Process, Ethiopian Public Health Institute,Addis Ababa; 2008.

57. WHO, Protocol for the Assessment of National Communicable Disease Surveillance and Response Systems; 2001.

58. Tsehaynesh M, A Five Year, Balanced Score Card Based Strategic Plan (2010-2015 G.C), Ethiopian Health and Nutrition Research Institute, Addis Ababa, Ethiopia, 2010.

59. U.S. Department Of Health And Human Services, CDC, an Introduction to Applied Epidemiology and Biostatistics, 2nd Edition, Principles Of Epidemiology, Atlanta, Georgia, 2005.

60. Gao S, Advanced Health Information Sharing With Web-Based GIS, Department Of Geodesy And Geometrics Engineering University Of New Brunswick, Canada, 2010.

61. Srinivas K., Rao G.R , Govardhan A., Survey On Prediction Of Heart Morbidity Using Data Mining Techniques, International Journal Of Data Mining & Knowledge Management Process (Ijdkp) Vol.1, No.3, May 2011, India; 2011.

62. Johns Hopkins and Red Cross Red Crescent, Public health guide in emergencies, International Federation of Red Cross and Red Crescent Societies, 2nd edition, Geneva, Switzerland, 2008.

63. Kamber M, Han J, Data Mining: Concepts and Techniques 2nd edition, University of Illinois at Urbana-Champaign; 2006.

64. Diansheng G, Jeremy M, Spatial data mining and geographic knowledge discovery—an introduction, Journal of Computers, Environment and Urban Systems, USA; 2009.

Appendix 1:

Surveillance Datasets between the years 2004-2008G.C. for attribute description, variable type, missing data elements.

No.	Attribute	Attribute description	Variable type	Missing values	% of missing values
1.	Region	Regions of the country	Text	0	0.00
2.	Zone	Zone of the country	Text	4	0.08
3.	Woreda	Woredas of the country	Text	4714	99.98
4.	Health F	Health Facilities in the country	Text	4714	99.98
5.	Number of month	The numbers of the Month	numeric	5	0.11
6.	Year	The year of the outbreak occurrences	Numeric	0	0.00
7.	NumReportingSites	The number of sites which are reporting to the sector	Numeric	655	13.89
8.	NumSitesReportedTimely	The number of sites which are reporting to the sector on timely bases	Numeric	652	13.83
9.	NumSitesReportedLate	The number of sites which are reporting to the sector lately	Numeric	649	13.76
10.	MalL5yOutPtCases	Malaria outpatient cases to less than 5 years of age	Numeric	13	0.28
11.	MalL5yInptCases	Malaria inpatient cases to less than 5 years of age	Numeric	34	0.72
12.	MalL5yInptDeaths	Malaria inpatient deaths to less than 5 years of age	Numeric	37	0.78
13.	MalGE5yOutPtCases	Malaria outpatient cases to greater than 5 years of age	Numeric	18	0.38
14.	MalGE5yInPtCases	Malaria inpatient cases to greater than 5 years of age	Numeric	34	0.72
15.	MalGE5yInPtDeaths	Malaria inpatient deaths to greater than 5 years of age	Numeric	39	0.83
16.	MalL5ySevAnemiaInPtCases	Malaria inpatient cases to less than 5 years of age with severe anemia	Numeric	43	0.91
17.	MalL5ySevAnemiaInptDeaths	Malaria inpatient deaths to less than 5 years of age with severe anemia	Numeric	43	0.91
18.	MalG5yWithSevAnemiaInptDeaths	Malaria inpatient deaths to greater than 5 years of age with severe anemia	Numeric	47	1.00
19.	MalG5SevAnemiaInptCases	Malaria inpatient cases to greater than 5 years of age with severe	Numeric	45	0.95

		anemia			
20.	MalariaInPregnancyOutPtCases	Malaria outpatient cases to pregnancy women	Numeric	21	0.45
21.	MalariaInPregnancyInptCases	Malaria inpatient cases to pregnancy women	Numeric	44	0.93
22.	MalariaInPregnancyInptDeaths	Malaria inpatient deaths to pregnancy women	Numeric	47	1.00
23.	MalL5yLabOutPtCases	Malaria outpatient cases which are confirmed by laboratory	Numeric	16	0.34
24.	UncomplicatedMalPVivax	Plasmodium vivax uncomplicated malaria cases	Numeric	19	0.40
25.	UncomplicatedLabMalOutptcases	uncomplicated malaria outpatient cases which are uncomplicated and laboratory confirmed	Numeric	4714	99.98
26.	MalGE5yLabOutPtCases	Malaria outpatient cases for greater than 5 years of age with laboratory confirmation	Numeric	15	0.32
27.	UncomplicatedMalG5LabConfPVivax	uncomplicated malaria laboratory confirmed cases for greater than 5 years of age with plasmodium vivax	Numeric	23	0.49
28.	TyphoidFeverOutptCases	Typhoid fever outpatient cases	Numeric	14	0.30
29.	TyphoidFeverInptCases	Typhoid fever inpatient cases	Numeric	43	0.91
30.	TyphoidFeverInptdeaths	Typhoid fever inpatient deaths	Numeric	51	1.08
31.	RelapsingFeverOutptcases	Relapsing fever outpatient cases	Numeric	14	0.30
32.	RelapsingFeverInptcases	Relapsing fever inpatient cases	Numeric	42	0.89
33.	RelapsingFeverinptdeaths	Relapsing fever inpatient deaths	Numeric	48	1.02
34.	EpidemicTyphusOutptcases	Epidemic Typhus Outpatient cases	Numeric	17	0.36
35.	EpidemicTyphusinptcases	Epidemic Typhus inpatient cases	Numeric	47	1.00
36.	EpidemicTyphusinptdeaths	Epidemic Typhus inpatient deaths	Numeric	49	1.04
37.	MeningitisOutpt	Meningitis Outpatient cases	Numeric	16	0.34
38.	MeningitisInpt	Meningitis inpatient cases	Numeric	4709	99.87
39.	Meningitisdeaths	Meningitis inpatient deaths	Numeric	4709	99.87
40.	DateSentReport	Date of the Report the sector	Numeric	4653	98.69
41.	DateRecReport	The date of the report received	Numeric	4682	99.30
42.	Month	Months of the report on the outbreak occurrences	Text	0	0.00

Appendix 2:

Surveillance Datasets between the years 2009-2012G.C for attribute description, variable type, missing data elements.

No.	Attribute Types	Attributes description	Variable Name	Missing values	% of missing values
1.	RegionName	Name of the reporting region	Text	0	0.00
2.	ZoneName	Name of the reporting zone	Text	180	1.24
3.	WoredaName	Name of the reporting Woredas	Text	3	0.02
4.	Health Facilities	Name of the reporting health facilities	Text	12649	87.18
5.	Epidemic week	Number of the reporting week	Numeric	441	3.04
6.	month	Name of the months epidemic year	numeric	3	0.02
7.	Year	The epidemic year	Numeric	0	0.00
8.	Total Malaria Confirmed and Clinical	Total numbers of malaria cases which are confirmed and clinical	Numeric	60	0.41
9.	Malaria Out_P Confirmed Cases	The number of malaria outpatient confirmed cases	Numeric	60	0.41
10.	Total Mala_InP_Cases	The number of malaria inpatient cases	Numeric	60	0.41
11.	Total Mal Deaths	Total number of malaria deaths	Numeric	60	0.41
12.	TMSuspected Fever Examined	Total number of malaria cases which are suspected and with fever examined	Numeric	829	5.71
13.	P_Falciparum	Total number of plasmodium falcifaroum cases	Numeric	60	0.41
14.	P_Vivax	Total number of plasmodium Vivax cases	Numeric	60	0.41
15.	Meningits total Cases	Total number of meningitis cases	Numeric	90	0.62
16.	Meningitis_OutP_Cases	Total number of outpatient meningitis cases	Numeric	1604	11.06
17.	Meningitis_InP_Cases	Total number of inpatient meningitis cases	Numeric	1133	7.81
18.	Meningitis_InP_Deaths	Total number of inpatient meningitis deaths	Numeric	1136	7.83
19.	Typhoid fever Total cases	Total number of typhoid fever cases	Numeric	87	0.60
20.	Typhoid fever_OutP_Cases	Total number of outpatient typhoid fever cases	Numeric	177	1.22

21.	Typhoid fever_InP_Cases	Total number of inpatient typhoid fever cases	Numeric	623	4.29
22.	Typhoid fever_InP_Deaths	Total number of inpatient typhoid fever deaths	Numeric	653	4.50
23.	Relapsing fever_total Cases	Total number of Relapsing fever Cases	Numeric	87	0.60
24.	Relapsing fever_OutP_Cases	Total number of Relapsing fever outpatient Cases	Numeric	651	4.49
25.	Relapsing fever_InP_Cases	Total number of Relapsing fever inpatient Cases	Numeric	651	4.49
26.	Relapsing fever_InP_Deaths	Total number of Relapsing fever inpatient deaths	Numeric	654	4.51
27.	Epidemic Typhus Total Cases	Total number of epidemic typhus cases	Numeric	87	0.60
28.	Epidemic Typhus_OutP_Cases	Total number of epidemic typhus outpatient cases	Numeric	634	4.37
29.	Epidemic Typhus_InP_Cases	Total number of epidemic typhus inpatient cases	Numeric	652	4.49
30.	Epidemic Typhus_InP_Deaths	Total number of epidemic typhus inpatient deaths	Numeric	653	4.50
31.	No-of-Gov_HPs-expected-by-RHB	Number of governmental heath posts which are expected to report to regional health bureau	Numeric	3225	22.23
32.	No-of-Gov_HCs-expected-by-RHB	Number of governmental health centers which are expected to report to regional health bureau	Numeric	2952	20.35
33.	No-of-Gov_Hosps-expected-by-RHB	Number of governmental hospitals which are expected to report to regional health bureau	Numeric	3513	24.21
34.	No-of-NGOHF-expected-by-RHB	Number of nongovernmental health facilities which are expected to report to regional health bureau	Numeric	3695	25.47
35.	No-of-OtherHFs-expected-by-RHB	Number of other health facilities which are expected to report to regional health bureau	Numeric	5403	37.24
36.	No-of-Gov_HPs-Reported	Number of governmental health posts which actually reported	Numeric	3061	21.10
37.	No-of-Gov_HCs-Reported	Number of governmental health centers which actually reported	Numeric	2694	18.57
38.	No-of-Gov_Hosps-Reported	Number of governmental hospitals which actually reported	Numeric	3497	24.10
39.	No-of-NGOHFs-Reported	Number of nongovernmental health facilities which actually reported	Numeric	3624	24.98
40.	No-of-OthersHF-Reported	Number of other health facilities which actually reported	Numeric	3596	24.78
41.	All-Total-sites-Reported	Number of all sites which actually reported	Numeric	2003	13.81
42.	All-Total-sites-Expected-by-RHB	Number of all total sites expected to report to the regional	Numeric	2385	16.44

		health bureau			
43.	No-of-Gov_HPs-expected-by-PHEM	Number of government health posts expected to report to PHEM	Numeric	6300	43.42
44.	No-of-Gov_HCs-expected-by-PHEM	Number of government health centers expected to report to PHEM	Numeric	5276	36.36
45.	No-of-Gov_Hosps-expected-by-PHEM	Number of government hospitals expected to report to PHEM	Numeric	7946	54.77
46.	No-of-NGOHF-expected-by-PHEM	Number of none government health facilities expected to report to PHEM	Numeric	9667	66.63
47.	No-of-OtherHFs-expected-by-PHEM	Number of other facilities expected to report to PHEM	Numeric	9618	66.29
48.	Total-Gov-sites-Reported	Total number of government site which actually reported	Numeric	2385	16.44
49.	Total-Gov-Expected-byPHEM	Total number of government site expected to report to PHEM	Numeric	2385	16.44
50.	Date Sent	The time in which the date of the expected sites sent the report	Date/time	14424	99.41
51.	Date Received	The time in which the date of the expected organization receives the report	Date/time	7099	48.93
52.	Timeliness	The manner in which the reports are arrived to expected sites	Text	7099	48.93
53.	EthiopianMonth	Epidemic months in Ethiopian calendar	Text	2385	16.44
54.	EthMonNum	The number of the month in Ethiopian calendar	Numeric	2385	16.44
55.	EthYear	The name of the year in Ethiopian calendar	Numeric	2378	16.39

Appendix 3:

=== Run information ===

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

Relation: Model development for classification algorithms on PHEM datasets-weka.filters.unsupervised.attribute.Remove-R6-9-weka.filters.unsupervised.attribute.NumericToNominal-Rfirst-last

Instances: 4093

Attributes: 5

RegionName

ZoneName

WoredaName

Month

EpidemicTyphusOutbreak

Test mode: 10-fold cross-validation

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

J48 pruned tree

RegionName = Addis Ababa

| WoredaName = Kolfe Keraniyo: Yes (33.0)

| WoredaName = Chirkos: Yes (34.0)

| WoredaName = Gulele: Yes (34.0/1.0)

| WoredaName = Addis Ketema: Yes (33.0)

| WoredaName = Akaki Kaliti: Yes (31.0/1.0)

| WoredaName = Bole: Yes (32.0/1.0)

| WoredaName = Menilik Hospital: No (19.0/2.0)

| WoredaName = Nefas Silk Lafto: Yes (32.0)

| WoredaName = Ras Desta Damitew Hospital:No (20.0/1.0)

| WoredaName = Tikur Anbessa Hospital: No (18.0/1.0)

| WoredaName = Yeka: Yes (32.0)

| WoredaName = Yekatit 12 Hospital: No (20.0/4.0)

| WoredaName = Zewditu Hospital: No (20.0)

| WoredaName = Lideta: Yes (32.0/2.0)

| WoredaName = Alert Hospital: No (3.0)

| WoredaName = Arada: Yes (31.0/1.0)

| WoredaName = Alert Hospital: No (15.0/2.0)

RegionName = Oromia

| ZoneName = East Shewa: Yes (33.0/2.0)

| ZoneName = Adama Special Zone: No (29.0/6.0)

| ZoneName = Bishoftu Town: No (29.0)

| ZoneName = Finfine Zuria: Yes (27.0/12.0)

| ZoneName = Jimma Spe Town: No (30.0/2.0)

| ZoneName = East Wellega: Yes (31.0/15.0)

| ZoneName = Arsi: Yes (35.0)

| ZoneName = Bale: Yes (32.0/8.0)

| ZoneName = Borena: Yes (33.0/13.0)

| ZoneName = Burayu: No (4.0)

| ZoneName = Dukem Town: Yes (4.0/1.0)

| ZoneName = East Hararge: No (33.0/6.0)

| ZoneName = North Shewa: Yes (22.0/1.0)

| ZoneName = Guji: Yes (31.0/15.0)

| ZoneName = Horo Gudru Wellega: No (33.0/13.0)

| ZoneName = Ilu Aba Bora:Yes (28.0/12.0)

| ZoneName = Jimma: No (30.0/8.0)

| ZoneName = Nekemte Town: No (32.0/11.0)

| ZoneName = Qeleme Wellega: No (31.0/8.0)

| ZoneName = Sebeta Town: No (2.0)

| ZoneName = Shashamane Town: Yes (32.0/8.0)

| ZoneName = South West Shewa: No (30.0/15.0)

| ZoneName = West Arsi: Yes (34.0/7.0)

| ZoneName = West Hararge: No (31.0/10.0)

| ZoneName = West Shewa: Yes (28.0/11.0)

| ZoneName = West Wellega: No (31.0/10.0)

RegionName = Tigray: No (1180.0/26.0)

RegionName = Gambella: No (26.0/2.0)

RegionName = SNNPR

| | ZoneName = Alaba: No (32.0/10.0)

| | ZoneName = Awassa Town: Yes (0.0)

| | ZoneName = Basketo: Yes (0.0)

| | ZoneName = Bench Maji: Yes (31.0/7.0)

| | ZoneName = Dawuro: No (32.0/12.0)

| | ZoneName = Gamo Gofa: Yes (32.0/2.0)

| | ZoneName = Gedeo: Yes (32.0/11.0)

| | ZoneName = Gurage: Yes (33.0/2.0)

| | ZoneName = Hadiya: No (32.0/8.0)

| | ZoneName = Keffa: Yes (31.0/4.0)

| | ZoneName = Kembata Timbaro: Yes (31.0/11.0)

| | ZoneName = Konta: Yes (0.0)

| | ZoneName = South Omo: Yes (32.0/3.0)

| | ZoneName = Sheka: No (32.0/13.0)

| | ZoneName = Sidama: Yes (33.0/6.0)

| | ZoneName = Siliti: Yes (32.0)

| | ZoneName = Wolayita: No (32.0/2.0)

| | ZoneName = Yem: Yes (0.0)

| WoredaName = Konso: Yes (21.0/5.0)

| WoredaName = Amaro: No (22.0)

| WoredaName = Awassa Town: Yes (33.0/3.0)

| WoredaName = Basketo: No (26.0/6.0)

| WoredaName = Burgi: No (22.0)

| WoredaName = Derashe: No (22.0/2.0)

| WoredaName = Konta Special: Yes (31.0/15.0)

| WoredaName = Yeme Special: No (22.0/7.0)

RegionName = Afar: No (436.0/20.0)

RegionName = B-Gumuz

| ZoneName = Assossa: No (25.0/3.0)

| ZoneName = Kemashi: No (22.0/2.0)

| ZoneName = Tongo: No (3.0)

| ZoneName = Maokomo Special: No (16.0/1.0)

RegionName = Dire Dawa: No (63.0/2.0)

RegionName = Harari: No (161.0/1.0)

RegionName = Amhara

| ZoneName = North Gonder: No (24.0/12.0)

| ZoneName = North Wollo: No (26.0/13.0)

| ZoneName = South Gonder: No (26.0/5.0)

| ZoneName = South Wollo: Yes (28.0/3.0)

| ZoneName = West Gojjam: Yes (25.0/5.0)

| ZoneName = Wag Himra: No (22.0/1.0)

| ZoneName = Bahir Dar Liyu: No (23.0/6.0)

| ZoneName = Oromiya: Yes (25.0/10.0)

| ZoneName = East Gojjam: Yes (25.0)

| ZoneName = North Shewa: Yes (13.0)

| ZoneName = Gonder Town: No (7.0)

| ZoneName = Dessie Town: Yes (5.0)

RegionName = Somali: No (20.0)

Number of Leaves: 91

Size of the tree: 97

Time taken to build model: 0.08 seconds

=== Stratified cross-validation ===

=== Summary Results of J48 classifier ===

Correctly Classified Instances	3588	87.6619 %
Incorrectly Classified Instances	505	12.3381 %
Kappa statistic	0.7143	
Mean absolute error	0.166	
Root mean squared error	0.294	
Relative absolute error	38.1998 %	
Root relative squared error	63.0666 %	
Total Number of Instances	4093	

=== Detailed Accuracy by Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
	0.795	0.085	0.814	0.795	0.804	0.935	Yes
	0.915	0.205	0.905	0.915	0.91	0.935	No
Weighted Avg.	0.877	0.167	0.876	0.877	0.876	0.935	

=== Confusion Matrix ===

a	b	<-- classified as
1038	268	a = Yes
237	2550	b = No