Addis Ababa University
School of Graduates Studies
School of Information Science and School of Public Health

Master of Science in Health Informatics

Application of data mining technique to develop chronic disease distribution map using drug distribution data in Ethiopia

By

Marara Zerihun

April 10, 2013

Addis Ababa

Addis Ababa University
School of Graduates Studies
School of Information Science and School of Public Health

Master of Science in Health Informatics

Application of Data mining technique to develop chronic disease distribution map using drug distribution data in Ethiopia

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

By

Marara Zerihun

April 10, 2013

Addis Ababa

Addis Ababa University

School of Graduates Studies
School of Information Science and School of Public Health

Master of Science in Health Informatics

Application of Data mining technique to develop chronic disease distribution map using drug distribution data in Ethiopia

By

Marara Zerihun

Name and Signature of members of the examining board

Name	Title	Signature	Date
Dr. Rahel Bekele	Chair Person	_____	_____
Dr. Alemayehu Worku	Advisor,	_____	_____
Ato Getachewu Jemaneh	Advisor,	_____	_____
Dr. Alemu Sime	Examiner,	_____	_____

Declaration

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

Name: Marara Zerihun

Signature: _____

Date: _____

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

Name: Dr. Alemayehu Worku

Signature: _____

Date: _____

Name: Ato Getachew Jemaneh

Signature: _____

Date: _____

Acknowledgements

Apart from the efforts of myself, the success of this thesis depends largely on the encouragement and guidance of many others. I take this opportunity to express my gratitude to the people who have been instrumental in the successful completion of this thesis. I would like to thank all those who have put something towards the accomplishment of this thesis. First, I gratefully thank almighty God from the depth of my heart for his support to reach this spot, without his guidance and support surely I wouldn't be here today. Next, I would like to extend my deepest thanks to my advisors Dr. Alemayehu Worku and Ato Getachew Jemaneh for their guidance, encouragement and good advice, in spite of having practically no spare time, still managed to find time to provide help and advice. Thirdly, I have special thanks to PFSA officials and workers for providing me with necessary data and information.

In addition I thank Billy Moon and Chris for your feedbacks and comments to enrich the work I have done. Very special thanks go to my former boss Etefa Tola, who encouraged me and supported me morally all the time. The guidance and support received from all members of classmates, office mates and others who contributed and who are contributing to this thesis, was vital for the success of the thesis.

I would like to thank my lovely family (Mom, Dad, brothers and Sisters) and friends, who have supported me throughout entire process, both by keeping me harmonious and helping me putting pieces together. I will be grateful forever for your love and support. I can't say thank you enough for your tremendous support and help.

Contents

Declaration ..... d

Acknowledgements i

List of figures v

List of tables v

List of acronyms vi

Abstract vii

Chapter One 1

1.1 Introduction 1

1.2 Statement of the problem 3

Research questions 3

1.3 Objective 4

1.3.1 General objective 4

1.3.2 Specific objectives 4

1.4 Methodology 5

1.4.1 Study area 5

1.4.2 Data collection procedure 5

1.4.3 Data analysis procedures 5

1.4.4 Data quality management 6

1.5 Ethical consideration 6

1.6 Scope of the study and expected outcome of the study 6

1.7 Dissemination of results 7

1.8 Thesis organization 7

Chapter Two 8

2. Literature review 8

2.1 Overview of data mining 8

2.2 Data mining 10
2.3 Methodologies of data mining research 11
2.3.1 KDD 11
2.3.2 SEMMA 13
2.3.3 CRISP-DM 14
2.3.4 Hybrid model 16
2.4 Assessment of the methodologies 19
2.5 Data mining tasks 20
2.5.1 Classification 21
2.5.2 Association rules 24
2.5.3 Clustering 27
2.6 Overview of chronic disease 29
2.7 Related works 34
Chapter Three 38
3. Research design and methodology 38
3.1 Data understanding and collection 39
3.2 Data preprocessing 42
3.2.1 Data cleaning 42
3.2.2 Data integration and data transformation 43
3.2.3 Attribute selection and statistical summary of attributes 45
3.4 Model building 47
3.4.1 Selection of modeling technique 47
3.4.1.1 Descriptive modeling 47
3.4.1.2 Predictive modeling 51
3.4.1.2.1 Decision tree 51

3.4.1.2.1 Naïve bayesian 53
Chapter Four 55
4. Experimentation and result analysis 55
4.1 Model building 55
4.1.1 Descriptive study of the data 55
4.1.2 Predictive study of the data 58
4.1.1 Predictive model using J48 algorithm 60
Confusion matrix for J48 decision tree classifier 63
4.1.2 Predictive model using Naïve Bayias algorithm 65
Confusion matrix for Naïve Bayes classifier 66
4.3 Experimental results and discussion 68
Chapter Five 72
5. Summary, conclusions and recommendations 72
5.1 Summary 72
5.2 Conclusion 73
5.3 Recommendations 75
References 77
Annex I 81
Diabetes 81
Cardiac Failure 81
Hypertension 82
Anti Parkinson 84
Asthma 84
Annex II 86

List of figures

Figure 2.1 Hybrid model diagram 20

Figure 2.2 Mortality of Ethiopia 36

Figure 3.1 Hybrid process model diagram 41

Figure 4.1 Consumption of each chronic disease by hub 58

Figure 4.2 Amount of chronic disease drugs issued to hubs organized by type 59

Figure 4.3 Quantity of chronic disease drug type issued to hubs organized by hubs 60

Figure 4.4 Tree view of experiment 4 65

List of tables

Table 2.1 KDD, SEMMA and CRISP-DM Methodologies 21

Table 3.1 List of attributes and their description 48

Table 4.1 Running time test for algorithms 61

Table 4.2 Summary of J48 decision tree experiment results 63

Table 4.3 Confusion matrix for J48 Decision tree model 66

Table 4.4 Summary of naïve bayes experiment results 67

Table 4.4 Confusion matrix of naïve bayes algorithm 69

Table 4.6 classification summary of both experimental models 72

List of acronyms

AIDS	**Acquired Immunity Deficiency Syndrome**
CART	**Classification and Regression Trees**
ARFF	**Attribute Relation File Format**
CVD	**Cardio Vascular Disease**
CSV	**Comma Separated Values**
CRISP	**Cross Industry Standard Process for Data Mining**
DM	**Data Mining**
ECSA	**Ethiopian Central Statistical Agency**
EPFSA	**Ethiopian Pharmaceuticals Fund Supply Agency**
FMOH	**Federal Ministry of Health**
GIS	**Geographic Information Systems**
HIV	**Human Infections Virus**
HMIS	**Health Commodity Management Information Systems**
KDD	**Knowledge Discovery in Database**
NCD	**Non Communicable disease**
SEMMA	**Sample, Explore, Modify, Model and Assess**
SPSS	**Statistical Package for the Social Sciences**
WEKA	**Waikato Environmental for Knowledge Analysis**
WHO	**World Health Organization**

Abstract

Background: Due to the great difference in population structure, geographic environment, food composition, ethnicity and lifestyle, it could be predicted that there may be significant differences of chronic disease forms and distribution in the various administrative areas. The amount of data getting generated in any sector these days is enormous. There are many data mining tools and technique to uncover hidden knowledge in the data. At the same time Ethiopian PFSA has huge and useful drug distribution data in their data base to investigate chronic disease distribution.

Objective: The purpose of this study is to investigate distribution of chronic diseases in various administrative areas of the country based on chronic disease drug distribution data applying data mining techniques.

Methods: Drug distribution data was collected from EPFSA. Data that are retrieved from the organization is from 2003 up to 2005 EC. Since annual data follow is high and distribution density is the same, two and half year's data is enough to produce distribution map and identify increase in demand applying data mining technology. Any data beyond these years are redundant and over saturate the models. In order to optimize the desired outcome the researcher has followed Hybrid data mining process model. The model is selected because it is appropriate for academic research; it combines the best features of KDD and CRISP; and starts with problem domain understanding.

Results: The study revealed that some drugs are more important at one hub than the other. Gullele hub received the hieghtest persontage of Athma (17.3%), Cardiac (38.5%), Diatetes (45.6%) and Hypertenion (28.99%). While Parkinson drugs are issued mostly to Mekele (15.5%) hub. The mining software revealed that some drugs are more important at one hub than the other in specified time.

Conclusions: Issue date, issue number and expiry dates are selected as best attribute by the mining tool. Based on discussion with domain experts issue date is important for drug distribution while issue number and expiry date are not relevant to the drug distribution.

Chapter One

1.1 Introduction

Ethiopia is a county with different administrative region with different ethnic groups residing in each region. Ethiopia is categorized in countries that have fast growing population. According to ECSA 2012 projected report the country's population is estimated to be over 80 Million (CSA, 2012). These different administrative regions also have diversified geographical environment, food composition, life style and climate. For example, some residents are living in the highland areas while others live in lowland areas, etc.

This study has used data mining techniques to analyze drug distribution data from Ethiopian Pharmaceuticals Fund and Supply Agency (PFSA). The use of data mining tools could efficiently analyze and investigate the distribution of chronic diseases forms in the various administrative areas of the Country, as well as understand the major chronic disease forms and the clustering effects in various administrative areas, and to further establish a chronic disease map for each administrative region. The information obtained could be used to assist local health authorities and community-based medical institutions in promoting and focusing on the most prevalent chronic diseases that are related to the locality.

As for all national drug supply chains, the drugs are sensitivity to heat poses a risk of waste, especially for the health facilities. The current system is semi automated and includes the use of stock cards which is predominant in most health centers. This system is inefficient and prone to errors (including transcription errors), is not smooth enough to be useful, and cannot forecast demand of inventory and makes it tedious to extract useful information for decision making at any level of the supply chain. Beside this fact the organization is using push system to distribute drugs to the hubs and facilities (Brandeis University, 2011; World Bank, 2009).

According to the study conducted by WHO joint with Ministry of Health (2003) There is no proper stock management in health facilities as revealed by absence of stock control tools such as stock card in 60 % of the surveyed health facilities. The accecibility national average for availability of key essential drugs in health facilities was 70%, 85% and 91 % for public health facilities, regional drug stores and private drug retail out lets, respectively (WHO Assessment of the Pharmaceutical Sector in Ethiopia, 2003).

The study is also very important for health authorities to distribute medical equipment that is more appropriate to community residents in the localities more effectively, as well as provide appropriate medical resources based on the local consumption, promote beneficial health improvement projects and monitor the results of health improvement projects and the effectiveness of various health improvement activities.

The study is limited to designing chronic disease distribution map based on major chronic disease drug distribution. Clustering techniques was used by the investigator to classify the dataset based on their destination and types of drugs.

1.2 Statement of the problem

The chronic disease drugs are bulky and more expensive both to purchase and store as the prevalence of chronic disease is increasing in the country. Thus, it is essential that an improved and efficient supply system, which is based on research, is put in place to reduce wasteges, stock outs, overstocking, and expired stock, and to improve on decision and information flow between the PFSA storage, the hubs, the regional health facilities, and every level of the drug supply chain (Martin, 2012; Brandeis University, 2011).

If the information on chronic disease forms and the major health problems in each administrative area were fully obtained, it would be possible to effectively allocate resources to develop public health policies for various areas from the perspective of public health, which can enable the appropriate allocation of medical resources to meet the needs of each administrative area.

As to the researcher knowledge no one investigated and designed chronic disease distribution in Ethiopia using drug distribution that matched with the different administrative regions of the country. This research has used data mining techniques on chronic disease drug distribution of Ethiopian Pharmaceuticals Supply and Fund Agency. In which region what kind of chronic disease drug is consumed more than the other region?

Research questions

1. What chronic disease drugs are more important at which hub?
2. In order to optimally use chronic disease drugs what important factors should be considered in distribution of chronic disease drugs in the country?

1.3 Objective

1.3.1 General objective

The general purpose of this study is to investigate distribution of chronic diseases in various administrative areas of the country based on chronic disease drug distribution data applying data mining techniques.

1.3.2 Specific objectives

In order to achieve the general objective, the researcher has carryied out the following specific objectives:

- To prepare good quality data by eliminating outliers, missing values, and noisy data which help for analysis by applying preprocessing tasks such as data cleaning, data transformation and attribute selection.
- To build a descriptive and predictive model using data mining tools and techniques on cleaned PFSA dataset.
- To test the performance of the model using test set and validate mining results with domain experts.
- To predict distributions of chronic disease among the regions in the country.

1.4 Methodology

1.4.1 Study area

Ethiopian Pharmaceutical Fund and Supply Agency is governmental organization which is responsible to distribute drugs and pharmaceutical equipments for the country. The main office is located in Addis Ababa in Gullele sub city in front of St. Paul hospital and there are 11 hubs in different parts of the Country. PFSA was established in 2007 by proclamation No. 553/2007. The storage and distribution directorate is one of the major sections dealing with distribution and storage of different types of drugs procured by the organization in general.

1.4.2 Data collection procedure

Since the research is conducted over the available data warehouse, the drug distribution data was used as base for data collection. Therefore, there was no primary data collection mechanisms has been employed. Hence secondary data were used.

1.4.3 Data analysis procedures

Data cleaning and data preparation is the second step of any data mining tasks. In Particular, data collected from different sources has to be manipulated into a form that has allowed the data mining tools to be used properly. This process of data cleaning and preprocessing is highly dependent on the techniques (methods) to be employed. For this research clustering techniques of mining were used. Data preparation has been done by using SPSS and MS Excel. After preparing the data in a form which is suitable for data mining techniques then it was analyzed using Weka 3.6.

1.4.4 Data quality management

In order to handle the problem of variables with missing values, there are different approaches recommended for data mining techniques. Among these methods: filling manually the missing values from hard copy records; filling the missing values with mean values; filling missing values with the most probable values using different techniques were used according to their necessity.

1.5 Ethical consideration

The study was conducted after getting approval from the ethical clearance committee of Addis Ababa University, School of Information Science and School of Public health. Then secondary data was collected from the organization after getting written consent from Health Informatics program of the institute and the confidentiality of data has been maintained. The retrieved data has been electronically locked with password and only the researcher and advisors have got access to the dataset.

1.6 Scope of the study and expected outcome of the study

The scope of the study is limited to develop chronic disease distribution map from drug distribution using descriptive data mining techniques. After the study is conducted successfully the National chronic disease distribution map was produced, which shows which part of the regions are facing what kinds of chronic diseases burdens.

1.7 Dissemination of results

The findings of the study will be presented and submitted to School of Information Science and School of Public health, Addis Ababa University. After that findings are evaluated and examined by scholars from both schools the result will be sent to health organizations and local as well as international journal publishers for publication.

1.8 Thesis organization

The organization of this study is presented as follows. In chapter one details of background of the study, statements of the problem, general and specific objectives of the study, methodology that is followed throughout the study, scope of the study, and significance of the study.

The second chapter deals with different conceptual and theoretical study of other similar studies. Other previous studies conducted in the country or other countries are going to be assessed by the researcher. Different literatures that are pertinent to the study have been reviewed thoroughly by the researcher.

The third chapter deals with methodologies that are going to be implemented to achieve the objectives of the study. Details of data cleaning, data preparation, and data preprocessing are presented in this chapter.

The fourth chapter deals with model building for experimentation and the results of each experiment and techniques implemented to achieve the objectives are going to be described thoroughly.

In chapter five the conclusions of the findings are going to be presented and the future work recommendations are going to be discussed.

Chapter Two

2. Literature review

Data mining technology is widely used in many fields (be it in health, agriculture, education, business, etc.) for discovering hidden knowledge which is crucial for competitive advantage and sustainable growth. In this chapter, it is be attempted to search and include a range of standards and steps used in data mining methodologies, various concepts, theories and practices of data mining in relation with drug distribution. Related research works, which reveal the applicability of data mining technology in the health sector, particularly chronic disease drug distribution are considered.

2.1 Overview of data mining

Due to advancement in technology these days we have loads of way to store transactional data and operational data in different format. Because the technology to store, search and retrieve information is advancing dramatically. This brought up the phenomenon of "*data rich information poor*" (W. J. Frawley, 1991). How to dig the valuable knowledge and the information from the mass of data is the main challenge information management faces these days. The challenge of extracting knowledge from data is of common interest to several fields, including statistics, database, pattern recognition, machine learning, data visualization, optimization, and high performance computing.

Although the capabilities to collect and store data in large computer databases has increased significantly, the relational database technology of today offers little functionality to process and explore data and establish a relationship or pattern among data elements that are hidden or previously unknown (Raghavan, et. al., 1998). Prather et. al., (2002) on his part wrote that, although health care databases have accumulated large quantities of information about patients and their medical conditions, there are only a

few tools to evaluate and analyze this clinical data after it has been captured and stored. The authors further stated that evaluation of stored clinical data might lead to the discovery of trends and patterns hidden within the data that could significantly enhance the understanding of disease progression and management.

Due to advancement in technology it is easier to capture and store large quantities of data. The amount of data being generated in organizations seems to be increasing and there's no end in sight (Witten and Frank, 2005). It is estimated that the amount of information in the world doubles every 20 months; that is, many scientific, government and corporate information systems are being overwhelmed by a flood of data that are generated and stored routinely, which grow into large databases amounting to giga (and even tera) bytes of data (Deogun, et. al., 2001). Rapid advances in data collection and storage technology have enabled organizations to accumulate vast amounts of data. We are overwhelmed by data - medical data, demographic data, financial data and marketing data (Han and Kamber, 2006).

The authors further argued that given certain data analysis goal, it has been a common practice to either design a database for application online data or use a statistical (or analytical) package on off-line data along with a domain expert to interpret the results. Even if one does not count the problems related with the use of standard statistical packages (such as its limited power for knowledge discovery, the need for trained statisticians and domain experts to apply statistical methods, and to refine and interpret results, etc.), one is required to state the goal and gather relevant data to arrive at that goal. Consequently, there is still strong possibility that some significant and meaningful patterns in the database, waiting to be discovered, are missed (Deogun, et. al., 2001).

The same logic holds true in PFSA, large repository of data are being stored in the database and data warehouse of the organization. This large repository of data has different knowledge, patterns and information is hidden. We need to use data mining technology to discover what kinds of meaningful patterns are stored in large repository of data.

Finding meaningful patterns in data has been given different kinds of names such as data mining, knowledge discovery, information discovery, information harvesting, data archaeology, and data pattern processing (Fayyad et al., 1996). Most researchers use data mining and KDD interchangeably and think they are synonymous. Others view data mining as an essential step in the process of KDD. However, the term data mining is becoming more popular than the term KDD (Han and Kamber, 2006).

2.2 Data mining

Data mining is about solving problems by analyzing data already present in databases. There are wide ranges of availability of large amounts of data across different kinds of organizations and there is technological push to turn this data into valuable information and knowledge. Many data mining tools and techniques are available to analyze and uncover important data patterns that improve productivity and competency in business organizations. This will highly improve the way organizations choose business strategy (Han et al., 2001).

Data can be found in different forms in an organization it can be as simple as numerical figures or complex as spatial data, multimedia, and hypertext documents. To take complete advantage of these data; data retrieval is simply not enough, it requires a tool for automatic summarization of data, extraction of the essence of information stored, and the discovery of patterns in raw data. With these vast amount of data stored in files, databases, and other repositories, it is incredibly important, to develop powerful tools (both hardware and software) for analysis and interpretation of such data. This extracted knowledge could be used in decision-making. The only way to achieve all this is through application of 'Data Mining' (Deshpande and Thakare, 2010).

2.3 Methodologies of data mining research

Data mining is a dynamic research and development area reaching maturity. As a research, it requires stable and well defined foundations which are well understood and popularized throughout the community (Kurgan and Musilek, 2006). The primary goal of data mining methodologies is building stable models following some logical steps (Berry and Linoff, 2004). Hence, different methodologies of data mining research attempt to shape the activities the researcher performs in a typical data mining process (Refaat, 2007).

As Azevedo and Santos (2008) stated, popular methodologies applied in data mining research include KDD (Knowledge Discovery in Database), SEMMA (Sample, Explore, Modify, Model, and Assess), Hybrid Model and CRISP-DM (CRoss Industry Standard Process for Data Mining).

2.3.1 KDD

Data mining is often set in the broader context of KDD. The KDD process basically involves selecting the target data, preprocessing the data, transforming them if necessary, performing data mining to extract patterns and relationships, and then interpreting and assessing the discovered structures (Hand et al, 2001). Fayyad et al (1996) broadly outlined the basic steps of KDD. Several steps from selecting data to providing understandable knowledge to the user are involved in the KDD process. These are further explained in the following steps.

Step One Selection: Before directly going to selection one needs to understand the domain and identify the goal of applying the process. Then select the necessary data to solve the problem. The main part of the KDD process is the selection of raw data that is necessary for the discovery. There might be unnecessary data attributes provided, but only few data attributes are needed in the process. Selecting the necessary data attributes for the discovery places an important part which yields target dataset.

Step Two Preprocessing: Handling missing values, eliminating noise and duplicate records in the data sets is the main part of this process. Missing values in the data lead to loss of useful information, which might not result in discovering useful knowledge. Noise in data and duplicate records mislead the process in obtaining accurate knowledge. Therefore, data cleaning and the preprocessing are necessary to produce better quality results. There are various tools such as Microsoft Excel, SPSS and WEKA 3.6 that can be applied for these tasks.

Step Three Transformation: In this step data is transformed or consolidated into forms appropriate for data mining. Data transformation involves the following. First, smoothing, where the noise from data is removed. Second, aggregation, where aggregation operations are applied to data for the analysis of the data. Third, generalization, where raw data is replaced with high level concepts. Lastly, normalization, where the attribute data are scaled to fall within a specified range; and feature Selection, where new attributes are constructed and added from the given set of attributes.

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

Step Five Interpretation / Evaluation: The step where the mining results are interpreted and even the process can start again from step one if there are any errors or for providing further accurate results. It involves the visualization of extracted patterns and results. The knowledge discovery process then takes the raw results from data mining and transforms them into useful and understandable information for the users. Using the knowledge directly, incorporating the knowledge into another system for further action, or simply documenting it and reporting it to interested parties, checking for and resolving potential conflicts with previously believed or extracted knowledge can be part of this step.

2.3.2 SEMMA

As explained in SAS (2011), SEMMA focuses on the model development aspects of data mining and involves the following logical steps:

Step One Sampling: extracting a portion of large dataset such that the sample taken is big enough to contain the significant information, or else small enough to manipulate quickly.

Step Two Exploring: searching for unanticipated trends and anomalies in order to gain understanding and ideas. Exploration helps refine the discovery process. If visual exploration does not reveal clear trends, it is possible to explore the data through statistical techniques.

Step Three Modifying: creating, selecting, and transforming the variables to focus the model selection process. Based on the discoveries in the exploration phase, one may need to manipulate the data to include information such as the grouping of clients and significant subgroups, or to introduce new variables. One may also need to look for outliers and reduce the number of variables, to narrow them down to the most significant ones. One may also need to modify data when the mined data change. Because data mining is a dynamic, iterative process, one can update data mining methods or models when new information is available.

Step Four Modeling: allowing the software to search automatically for a combination of data that reliably predicts a desired outcome.

Step Five Assessing: evaluating the usefulness and reliability of the findings from the data mining process and estimate how well it performs. A common means of assessing a model is to apply it to a portion of data set aside during the sampling stage. If the model is valid, it should work for this reserved sample as well as for the sample used to construct the model. Similarly, one can test the model against known data.

By assessing the results gained from each stage of the SEMMA process, one can determine how to model new questions raised by the previous results, and thus proceed back to the exploration phase for additional refinement of the data.

2.3.3 CRISP-DM

Business (Problem) understanding phase

The first phase in the CRISP–DM standard process may also be termed the research understanding phase. According to (Shearer, 2000) the most important phase of any data mining project is the initial business understanding phase which focuses on understanding the project objectives from a business perspective, converting this knowledge into a data mining problem definition, and then developing a preliminary plan designed to achieve the objectives. In order to understand which data should later be analyzed and how, it is vital for data mining practitioners to fully understand the business for which they are finding a solution.

The business understanding phase involves several key steps; including determining business objectives, assessing the situation, determining the data mining goals, and producing the project plan (Chapman et al, 2000). The two crows' corporation (2005) also stated that to make the best use of data mining one must make a clear statement of the objective. Without clear understanding of the problem to be solved, it is difficult to identify, select, and prepare the data for mining or to correctly interpret the results.

Data understanding

According to (Chapman et al., 2000) the data understanding phase starts with an initial data collection. The analyst then proceeds to increase familiarity with the data, to identify data quality problems, to discover initial insights into the data, or to detect interesting subsets to form hypotheses about hidden information. The data understanding phase involves four steps, including the collection of initial data, the description of data, the exploration of data, and the verification of data quality.

Data Preparation Phase

The data preparation and preprocessing phase covers all activities to construct the final data set or the data that will be fed into the modeling tool(s) from the initial raw data. Tasks include

table, record, and attribute selection, as well as transformation and cleaning of data for modeling tools. The five steps in data preparation are the selection of data, the cleansing of data, and the construction of data, the integration of data, and the formatting of data (Chapman et al, 2000; Larose, 2005).

This phase is often the most time consuming task of KDD processes, especially if data is drawn directly from the company's operational databases rather than from the data warehouse. As mentioned by (Han and Kamber, 2006) the data stored in databases may reflect noise, exceptional cases, or incomplete data objects. When mining data regularities, these may confuse the process, causing the knowledge model constructed to over fit the data. As a result the accuracy of the discovered patterns can decrease.

Building Model Phase

In this phase, various modeling techniques are selected and applied and their parameters are adjusted to optimal values. Typically, several techniques exist for the same data mining problem type. Two crows' corporation (2005) describes that the most important thing to remember about model building is that it is an iterative process. You will need to explore alternative models to find the one that is most useful in solving your business problem. What you learn in searching for a good model may lead you to go back and make some changes to the data you are using or even modify your problem statement. Some techniques have specific requirements on the form of data. Therefore, stepping back to the data preparation phase may be necessary. Modeling steps include the selection of the modeling technique, the generation of test design, the creation of models, and the assessment of models (Chapman et al, 2000; Shearer, 2000).

Evaluation Phase

Before proceeding to the final deployment of the model, it is important to thoroughly evaluate the model and review the model's construction to be certain it properly achieves the business

objectives. Here it is critical to determine if some important business issue has not been sufficiently considered. At the end of this phase, the project leader then should decide exactly how to use the data mining results. The key steps here are the evaluation of results, the process review, and the determination of next steps (Shearer, 2000).

Deployment Phase

Model creation is generally not the end of the project. The knowledge gained must be organized and presented in a way that the customer can use it, which often involves applying "live" models within an organization's decision-making processes, such as the real-time personalization of Web pages or repeated scoring of marketing databases (Shearer, 2000).

According to (Chapman et al, 2000), depending on the requirements, the deployment phase can be as simple as generating a report or as complex as implementing a repeatable data mining process across the enterprise. Even though it is often the customer, not the data analyst, who carries out the deployment steps, it is important for the customer to understand up front what actions must be taken in order to actually make use of the created models.

2.3.4 Hybrid model

The development of academic and industrial models has led to the development of hybrid models that combine aspects of both. One such model is a six-step KDP model developed by Casio (Krzysztof J et al., 2007). It was developed based on the CRISP-DM model by adopting it to academic research. The main differences and extensions include providing more general, research-oriented description of the steps and introducing a data mining step instead of the modeling step (Krzysztof J et al., 2007).

Further description of the six steps of the model follows,

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

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

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

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

5. Evaluation of the discovered knowledge. Evaluation includes understanding the results, checking whether the discovered knowledge is novel and interesting, interpretation of the results by domain experts, and checking the impact of the discovered knowledge. Only approved models are retained, and the entire process is revisited to identify which alternative

actions could have been taken to improve the results. A list of errors made in the process is prepared.

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

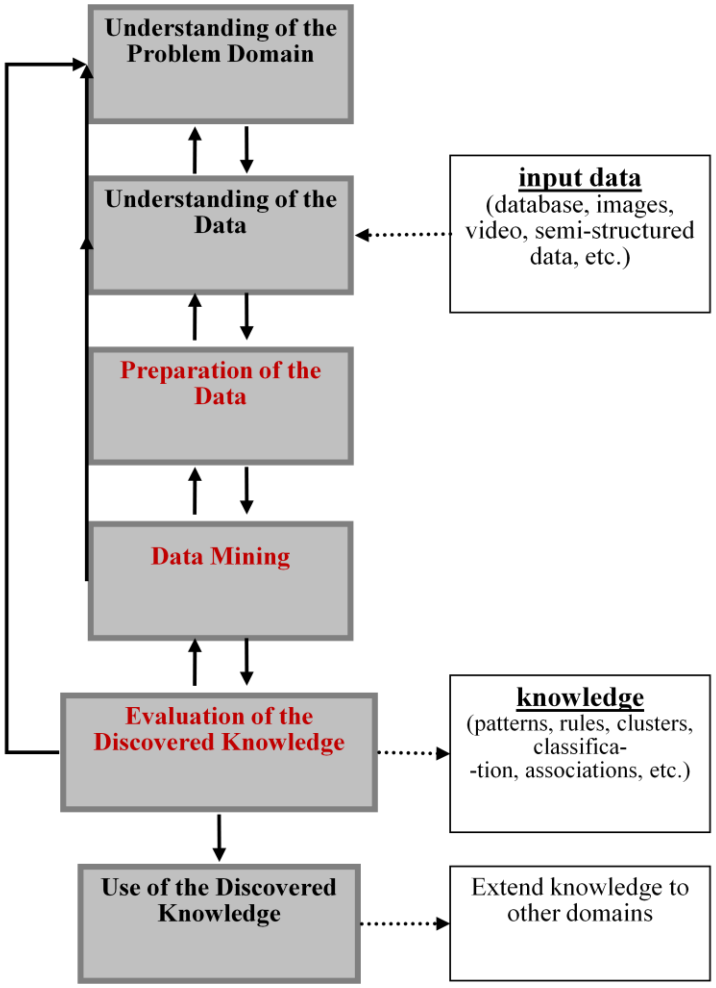


Figure 2.1 Hybrid model diagram

2.4 Assessment of the methodologies

Recently, the research efforts have been focused on proposing new models, rather than improving design of a single model or proposing a generic unifying model. Despite the fact that most models have been developed in isolation, a significant progress has been made these days. Most of the models provide generic and appropriate descriptions that are not tied specifically to academic or industrial needs, but rather provide a model that is independent of a particular tool, vendor, or application (Kurgan and Musilek , 2006).

Azevedo and Santos (2008) have summarized the association of the three popular methodologies as shown in table 2.1.

KDD	SEMMA	CRISP-DM
Pre KDD	-----	Business understanding
Selection	Sample	Data understanding
Preprocessing	Explore	
Transformation	Modify	Data preparation
Data mining	Model	Modeling
Interpretation/Evaluation	Assessment	Evaluation
Post KDD	----	Deployment

Table 2.1 KDD, SEMMA and CRISP-DM Methodologies

As shown in table 2.1, most of the steps to be followed in all methods seem somehow similar. Though KDD and SEMMA do not have a step before selection and sample steps respectively, obviously understanding the domain is often a requirement to perform selection or sampling.

According to (Refaat, 2007), all methodologies contain the following set of main tasks in one form or another. Typically these steps are performed iteratively and not necessarily in the presented linear order.

First, relevant data elements are extracted from the database into one table containing all the variables needed for modeling. In the case when the size of the data cannot be handled efficiently by available data modeling tools, which is frequently the case, sampling is used.

Second, exploratory data analysis (EDA) are performed to gain some insight about the relationships among the data and to create a summary of the properties.

Third, based on the results of EDA, transformation procedures are invoked to highlight and take advantage of the relationships among the variables in the planned models.

Fourth, a set of data mining models are then developed using different techniques, depending on the objective of the research and the types of variables involved. A data reduction procedure is often invoked to select the most useful set of variables.

Fifth, the data mining models are evaluated and the best performing model is selected according to performance criteria.

Finally, the selected model is used to achieve the required business objectives.

2.5 Data mining tasks

Data mining functionalities are used to specify the kind of patterns to be found in data. Data mining tasks can be classified into two categories: **descriptive** and **predictive**.

Descriptive mining tasks characterize the general properties of the data in the database (Han and Kamber, 2006). The goal of descriptive data mining is to gain an understanding of the analyzed system by discovering patterns and relationships in large datasets (Kantardzic, 2003). Predictive mining tasks, on the other hand, perform inference on the current data in order to make predictions (Han and Kamber, 2006).Classification is the commonly used data mining technique for prediction, whereas association and clustering are considered the descriptive data mining techniques (Kantardzic, 2003).

As Witten and Frank (2005) have stated, WEKA 3.6 data mining tool includes methods for all the standard data mining problems including classification, clustering, association rule mining, and attribute selection. The workbench is a collection of state-of-the-art machine learning algorithms and data pre-processing tools.

2.5.1 Classification

Classification is learning a function that maps (classifies) a data item into one of several predefined classes (weiss and kulikwski, 1991). Classification is a supervised learning which attempts to categorize objects through constructing a model from training datasets with predetermined targets, fine-tuning it with test datasets, and using it to classify other datasets of interest (Chen et al., 2006).

Basically classification involves dividing up objects so that each is assigned to one of a number of mutually exhaustive and exclusive categories of classes. The term 'mutually exhaustive and exclusive' implies that each object must be assigned to precisely one class that is never to more than one, and never to any class at all. For instance, a hospital system may want to classify medical patients into those who are at high, medium or low risk of acquiring a certain illness. So in this case, a patient should be classified as high, medium or low class (Bramer, 2007).

According to (Han and Kamber, 2006), classification is a two-step process: learning and classification. During learning, a classifier is built describing a predetermined set of classes. The classification algorithm builds the classifier by analyzing a training set and their associated class labels. The learned model or classifier is represented in the form of classification rules. The accuracy of the classification rules is estimated using test data.

If the accuracy is considered acceptable, the rules can be applied to the classification of new data tuples. Decision trees, classification rules, and neural networks are some of the techniques for creating models.

As (Berry and Linoff, 2004) have stated, any of the techniques used for classification can be adapted for use in prediction by using training sets where the value of the variable to be predicted is already known. The historical data is used to build a model that explains the current observed behavior. When this model is applied to current inputs, the result is a prediction of future behavior.

2.5.1.1 Decision tree induction

Tree models have been around for a very long time, although formal methods of building them are a relatively recent innovation. Before the development of such methods they were constructed on the basis of prior human understanding of the underlying processes and phenomena generating the data (Hand et al., 2001)

Decision trees are supervised learning methods that construct decision trees from a set of input-output samples. A typical decision tree learning system adopts a top-down strategy that searches for a solution in a part of the search space (Kantardzic, 2003). Decision tree induction is the learning of decision trees from class-labeled training dataset (Han and Kamber, 2006).

There is a large number of decision-tree induction algorithms described primarily in the machine-learning and applied-statistics literature (Kantardzic, 2003). ID3, C4.5, J48 and Classification and Regression Trees (CART) are the algorithms used to construct decision trees in a top-down recursive divide-and-conquer manner (Han and Kamber, 2006). As stated by Kantardzic (2003), a well-known tree-growing algorithm for generating decision trees is Quinlan's ID3 with an extended version called C4.5.

According to (Larose, 2005) there are requirements to be met before implementation of decision tree algorithms. First, since decision tree algorithms represent supervised learning they require predetermined target variables and a training dataset which provides the algorithm with the values of the target variable. Second, the training dataset should be rich and

varied, providing the algorithm with a healthy cross section of the types of records for which classification may be needed in the future. Decision trees learn by example, and if training set are systematically lacking for a definable subset of records, classification and prediction for this subset will be problematic or impossible. Third, the target attribute classes must be discrete values that are clearly demarcated as either belonging to a particular class or not belonging.

Decision tree classifiers have good accuracy, in general. However, successful use may depend on the data at hand. During tree construction, attribute selection measures are used to select the attribute that best partitions the tuples into distinct classes (Han and Kamber, 2006).

2.5.1.2 Rule-based classification

Even though the pruned trees are more compact than the originals, they can still be very complex. Hence, to make a decision tree model more readable, it can be transformed into an IF-THEN decision rule (Kantardzic, 2003). As stated by (Larose, 2005), decision rules can be generated from a decision tree by traversing any given path from the root node to any leaf. The complete set of decision rules generated by a decision tree is equivalent to the decision tree itself.

A rule-based or decision rule classifier such as C4.5 uses a set of IF-THEN rules for classification. An IF-THEN rule is an expression of the form IF condition THEN conclusion. If the condition in a rule antecedent holds true for a given tuple, we say that the rule antecedent is satisfied and that the rule covers the tuples (Han and Kamber, 2006).

2.5.1.3 Neural network

Neural network is a complex modeling technique based on a model of a human neuron. A neural net is given a set of inputs and weights to predict one or more outputs depending on the selected architecture (Bramer, 2007). One of the most common neural network architectures

has three layers. The first layer is the input layer, the layer exposed to external signals. The input layer transmits signals to the neurons in the next layer, which is called a hidden layer. The hidden layer extracts relevant features or patterns from the received signals. Those features or patterns that are considered important are then directed to the output layer (Bernander, 2009).

Neural networks are powerful mathematical models suitable for almost all data mining tasks, with special emphasis on classification and estimation problems (Refaat, 2007). (Larose, 2005) stated that, neural networks are quite robust with respect to noisy data. (Han and Kamber, 2006) also stated the success of neural networks on a wide area of application including handwritten character recognition and medicine due to their high tolerance of noisy data and their ability to classify patterns on which they have not been trained. However, (Han and Kamber, 2006) argue that, neural networks require long training times and have been criticized for their poor interpretability. Therefore, neural networks are more suitable for applications where this is feasible.

2.5.2 Association rules

Association rules are one of the major data mining techniques. It is perhaps the most common form of local-pattern discovery in unsupervised learning systems (Kantardzic, 2003). However (Larose, 2005) has stated that, association rule mining can be applied in either a supervised or an unsupervised manner. As (Hand, 2001) stated pattern discovery has been the focus of much attention in data mining and has been addressed using algorithmic techniques based on association rules. (Witten and Frank, 2005) argue that, unlike classification, association rules are able to predict any combinations of attributes and these different rules express different regularities that underlie the datasets.

Agrawal et al. (1993) have defined association rule problem as:

Given a set of items I={I_1,I_2,…,I_m} and a database of transactions D={t_1,t_2, …, t_n} where ti={I_{i1},I_{i2}, …, I_{ik}} and I_{ij} I, find all association rules X ➔ Y with a minimum support and confidence. Where, Support of AR (s) X ➔Y is the percentage of transactions that contain $X \cup Y$, and Confidence of AR (c) X ➔ Y is the ratio of number of transactions that contain $X \cup Y$ to the number that contain X.

As it has been pointed out by (Hand et al., 2001), association rules consist of a left -hand side proposition X (the antecedent or condition) and a right hand side Y (the consequent). The rule states that if the left-hand side is true, then the right-hand side is also true with conditional probability.

2.5.2.1 Apriori approach to pattern mining

Association Rules Mining (ARM) problem is usually decomposed in to finding frequent itemsets whose occurrences exceed a predefined threshold in the database; and generate association rules from those large itemsets with the constraints of minimum confidence (Kotsiantis and Kanellopoulos , 2006).

Agrawal and Srikant (1994) have observed an interesting downward closure property, called Apriori, among frequent k-itemsets. A k-itemset is frequent only if all of its sub-itemsets are frequent. As they stated, the downward closure property implies that frequent itemsets can be mined by first scanning the database to find the frequent 1-itemsets, then using the frequent 1-itemsets to generate candidate frequent 2-itemsets, and check against the database to obtain the frequent 2-itemsets. This process iterates until no more frequent k-itemsets can be generated.

2.5.2.2 FP-Growth approach to pattern mining

Frequent pattern growth (FP-growth) is another important ARM algorithm that mines frequent itemsets in large databases in an efficient way. The algorithm mines frequent itemsets without the time consuming generation of candidates (Kantardzic, 2003). As stated by Han and Kamber (2006), first start from each frequent length-1 pattern and construct its conditional pattern base which consists of the set of prefix paths in the FP-tree co-occurring with the suffix pattern, then construct its conditional FP-tree, and perform mining recursively on such a tree. The pattern growth is achieved by the concatenation of the suffix pattern with the frequent patterns generated from a conditional FP-tree.

2.5.2.3 Measures of pattern interestingness

A pattern is interesting if it is easily understood by humans, valid on new or test data with some degree of certainty, potentially useful, and novel. It is also interesting if it validates a hypothesis that the user sought to confirm (Han and Kamber, 2007).

In the case of classification rules we are generally interested in the quality of the set of rules working in combination to determine the effectiveness of a classifier. However, in association rule mining the emphasis is on the quality of each individual rule. To distinguish between one rule and another some measures of rule quality are needed (Bramer, 2007).

Interestingness can be quantified in various ways using the frequency and accuracy of the pattern as well as background knowledge (Hand et al., 2001). Based on the structure of the discovered patterns and the statistics underlying them, there exists several objective measures of pattern interestingness including the most popular support and confidence (Han and Kamber, 2006).The rule frequency or support tells us how often a rule is applicable. In many cases, rules with low frequency are not interesting, and this assumption is indeed built into the definition of the association rule finding problem. The accuracy of an association rule is not necessarily a very good indication of its interestingness. For example, consider a medical

application where the rule is learned that pregnancy implies that the patient is female with accuracy 1 is not interesting since it is already known (Hand et al., 2001).

Subjective interestingness measures are based on user beliefs in the data. These measures find patterns interesting if they are unexpected or offer strategic information on which the user can act and such patterns are referred to as actionable (Han and Kamber, 2006).

Generation of too much rules satisfying the minimum threshold is a major challenge during pattern discovery (Han and Kamber, 2006; Kantardzic, 2003). Generating only “interesting” rules, generating only “non redundant” rules, or generating only those rules satisfying certain criteria such as coverage, leverage, lift or strength are among the strategies to alleviate the problem(Kotsiantis and Kanellopoulos ,2006).

According to (Hand et al., 2001), there exist many variants of the basic association rule algorithms. The methods typically strive toward minimizing the number of passes through the data, minimizing the number of candidates that have to be inspected, and minimizing the time needed for computing the frequency of individual candidates.

To effectively extract information from a huge amount of data in databases, data mining algorithms must be efficient and scalable. In other words, the running time of a data mining algorithm must be predictable and acceptable in large databases. From a database perspective on knowledge discovery, efficiency and scalability are key issues in the implementation of data mining systems (Han and Kamber, 2006).

2.5.3 Clustering

Clustering is concerned with grouping together objects that are similar to each other. The objects are grouped on the basis of maximizing the intraclass similarity and minimizing the interclass similarity. For instance, in a medical application we might wish to find clusters of patients with similar symptoms (Bramer, 2007; Han and Kamber, 2006).

Clustering techniques are applied when there is no class to be predicted but rather when the instances are to be divided into natural groups (Chakrabarti et al., 2009).

Clustering algorithms examine data to find groups of items that are similar (Bramer, 2007). As Kantardzic (2003) has stated, most clustering algorithms work based on partitioning and hierarchical methods.

Partitioning algorithm organizes the objects into k partitions, where each partition represents a cluster. The k-means algorithm takes the input parameter, k, and partitions a set of n objects into k clusters. Cluster similarity is measured in regard to the mean value of the objects in a cluster, which can be viewed as the center of the cluster (Han and Kamber, 2006; Refaat, 2007). The objects or instances assigned to their closest cluster center according to the ordinary Euclidean distance metric (Chakrabarti et al., 2009).

The most apparent drawback of this k means of clustering is that there is no principled way to know what the value of k ought to be (Bramer, 2007).

Hierarchical clustering techniques organize data in a nested sequence of groups, which can be represented in the form of dendrogram or tree structure (Kantardzic, 2003). As stated by Han and Kamber (2006), hierarchical clustering methods can be further classified as either agglomerative or divisive, depending on whether the hierarchical decomposition is formed in a bottom-up or top-down approach.

Clustering method has been applied in areas of medicine, such as psychiatry, to try to identify whether there are different subtypes of diseases lumped together under a single diagnosis (Hand et al., 2001).

2.6 Overview of chronic disease

Over the last 15 years Jimma University Hospital in southwest and, University of Gondar Hospital in northwest Ethiopia have been pioneering primary care treatment for chronic disease in rural communities (Mamo Y et al., 2007). Patients with chronic diseases must attend regularly for monitoring of treatment over long periods of time and they find it difficult to attend distant hospitals. Adherence is likely to be better and the costs to the patient reduced if treatment is made available closer to their homes. It has been demonstrated that, with support and training, nurses and health officers can manage chronic diseases effectively in health centres and there is now a strong argument for this model to be adopted more widely.

According to Martin P (2012) there is large hidden burden of chronic disease identified in the population studied by the team from Jimma University and in other studies indicates that, as well as improving access to treatment, there is a need for community education and improved case detection. It is good to use experience gained through educational campaigns on HIV, tuberculosis and malaria, schools, churches, mosques and local mass media could be used to improve awareness about chronic diseases. Health extension workers also have the potential to play an important part in education and improving case detection.

Martin P (2012) further recommended chronic disease seek due attention by policy makers. Although the studies from Jimma University were performed in a single, relatively small population, the findings are consistent with previous reports from both Jimma and Gondar call for a greater prioritization of chronic disease in healthcare policy.

The research team from Jimma conducted a cross sectional study of chronic disease and risk factors for chronic disease around Gilgel Gibe Field Research Centre in southwest Ethiopia using the World Health Organization's STEPS protocol. They found an overall prevalence of chronic disease of 8.9% (diabetes 0.5%, cardiac disease 3%, hypertension 2.6%, asthma 1.5%, epilepsy 0.5%, depression 1.7%), and 80% of the subjects studied had at least one risk factor for chronic disease. The data on prevalence of chronic diseases were dependent on subjects reporting that

they had been given a diagnosis by a health professional. When a sample was screened for hypertension and diabetes, the prevalence of hypertension was found to be 3.5 times higher than that reported by the subjects and the prevalence of diabetes six times higher, indicating a large hidden burden of disease (Prevett M, 2012)

Diabetes mellitus is one of the most serious threats to American health. Recent studies indicate that the prevalence of diabetes is increasing among the general population in the U.S. and worldwide. More than 135 million people are affected by diabetes worldwide including more than 8.5 million Americans. In many developed countries, 10 to 20% of people over 45 years are affected by non-insulin dependent diabetes mellitus. About 15% of people over age 70 have type II diabetes. The World Health Organization has concluded that diabetes is becoming a global epidemic and predicts that the disease will be one of the world's major contributors to morbidity by 2025 (WHO, 2003). Blacks and Hispanics have a twofold to threefold increased risk of developing type II diabetes; poverty is also an important risk factor. Diabetes increases the risk of blindness, loss of limbs, heart disease, stroke, kidney failure, and depression (Talbot F, 2000) and generally reduces life expectancy. The costs of diabetes arise from the treatment of the disease, its complications, lost workdays, and disability. Of the three diseases under consideration, diabetes clearly has the greatest public health impact.

According to (WHO, 2005) Chronic disease risk factors are Smoking, Heavy drinking of Alcohol, Dietary, Environmental Condition and Life style. These problems are also common in Ethiopia.

There are different risk factors contributing to chronic diseases in Ethiopia. According to Fessahaye A et al., (2012) the prevalence of these risk factors varies from place to place. For example the prevalence of smokers in Jimma (9.3%) was higher than findings of a national study in Ethiopia (9.0%) (WHO Ethiopia Report, 2003), findings of community-based studies in Butajira, Ethiopia (Tesfaye, 2008; Schoenmaker et al., 2005), and findings among higher education communities in Ethiopia and elsewhere (Eshetu E and Gedif T, 2006).

According to Fessahaye A et al. (2012) the prevalence of alcohol consumption in Jimma population is 7.3% which is much lower than the prevalence found in Butajira town, Ethiopia. The lower prevalence in our study population might be due to religion difference. Alcohol consumption prevalence was higher among men which is 8.5% compared to women 5.7% in the area. This shows that the matter to risk factors of chronic disease distribution. The prevalence of alcohol consumption was higher for urban population which is 19.6% than rural population 2.9%.

Environmental risk factors include lifestyle factors, like diet and exercise habits, are even more important than environmental exposures. Obesity is a risk factor for type II diabetes; 80 to 90% of the people with this disease are obese. The incidence of non-insulin dependent diabetes mellitus has also been found to increase, up to a point, with the number of cigarettes smoked per day. Rates of diabetes have risen with those of obesity and poor physical fitness.

According to the study in Jimma Concerning dietary pattern about 27.0% of the population reported to consume fruits and vegetables below adequate level (below five servings per day). The proportion of the population who consumed fruits and vegetables below adequate level in this study is by far lower than the findings in all 52 countries taking part in the 2002-2003 world health survey including Ethiopia. The prevalence of low fruit and vegetable intake in pooled sample from all the 52 countries was 78.0%.

According to WHO report 2005, the life's of far too many people in the world both in developing and developed countries in the world are being blighted and cut short by chronic diseases such as heart disease, stroke, cancer, chronic respiratory diseases and diabetes (WHO report, 2005). It is also major cause of morbidity and mortality in Ethiopia.

Joint with different nongovernmental organizations Ethiopian government is implementing different programs to improve the health of chronic disease patients. For example, a community health approach to palliative care for HIV and Cancer Patients in Africa is a joint

project with Ministry of Health and World Health Organization (WHO report, 2005). The main goal of this project is to improve the quality of life of HIV/AIDS and Cancer patients in the country by developing comprehensive palliative care programmes with a community health approach.

Chronic diseases are increasing in global prevalence and seriously threaten developing nations' ability to improve the health of their population. Although often associated with developed nations, the presence of chronic disease has become the dominant health burden in many developing countries. Chronic diseases were responsible for 50% of the disease burden in 23 high burden developing countries in 2005 and will cost those countries $ 84 billion by 2015 if nothing is done to slow their growth (Rachel Nugent, 2008).

Numerous studies have shown that differences in geographical environments affect health, and that the occurrence of certain diseases is relevant to specific ethnic groups. Our country Ethiopia has diversified geographical environment and ethnicity that can contribute to different health problems in different areas and among different ethnic groups. For example, among the high risk groups defined by the American Diabetes Association (Diab. Care, 2002) Latinos, African Americans, and the residents living on Pacific islands are all ethnic groups who are more likely to suffer from diabetes. The place of birth, living style, air to be breathed, environment, etc. all has an effect on the public health policies to be established (Dummer, 2008).

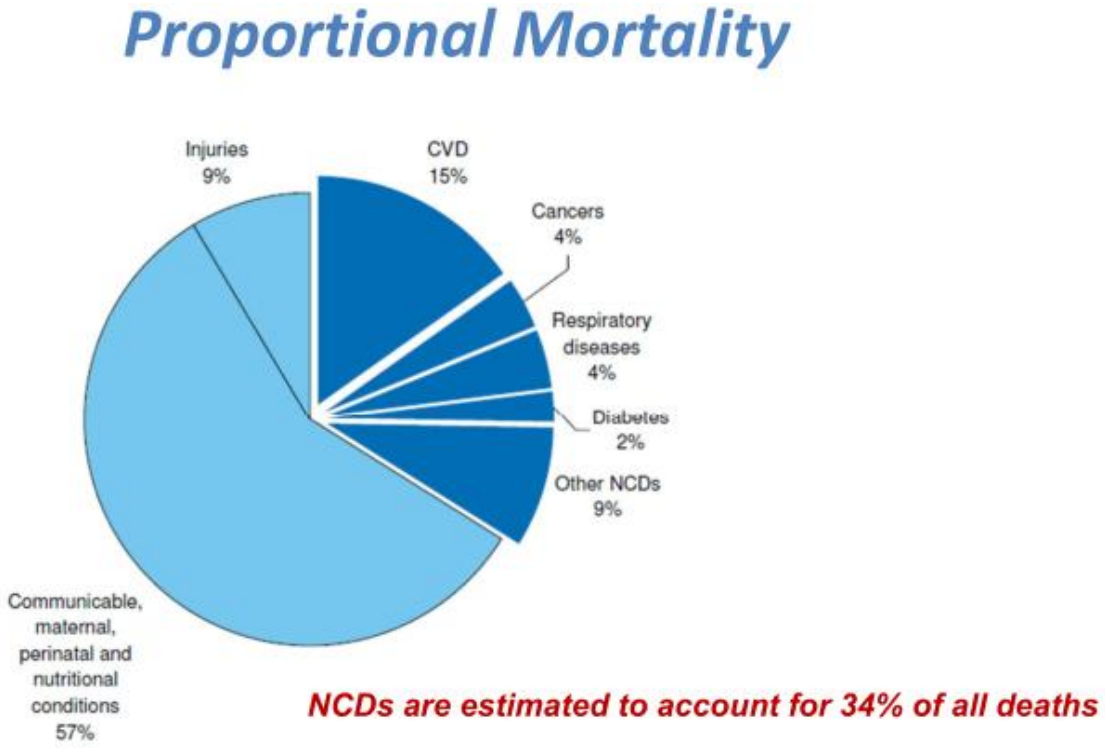
Cutchin described that over the past decade, social and environmental epidemiological studies have made important contributions by indicating the close relationship between geographical environment and health. The study gave an introduction to health geography and investigated how it revealed information on social epidemiology and health, using streets as the unit to divide areas. It found that exposure to large-scale petrochemical plants in Texas would directly and indirectly affect the environment and environmental exposure (Cutchin, 2007).

Harvard university joint with Addis Continental Institute of public health conducted a cohort study among bank workers and teachers in Addis Ababa to identify the burden of NCD (Non Communicable Disease) in Ethiopia. The overall prevalence of Diabetes in the county is 6.5%. Among these female are 6.6% and male are 6.4%. On the other hand the prevalence of Hypertension is 19.1% overall. The distribution by sex in the study showed huge difference among male and female. Male (22%) are more hypertensive than women (14.9%). Age differences also show huge variation according to the study. Interesting pattern was revealed on the study that the prevalence of hypertension increases as the age increases. The age categorized into less than or equal to 24 (5.6%) which is less than the age categorized into 25-34 (9%) and again this is a lot less than the next group which is categorized into 35-44 (18.5%). Again the next age group 45-54 (37.7%) and age category greater than or equal to 55 (53.5%) which is much higher than the other categories (Michelle, 2012).

According to the study by Harvard University risk factors contributing to the high prevalence of NCDs burdens in the country such as smoking, chewing khat, physical inactivity, reduced fruits and vegetables consumptions and obesity were found to be major problems. According to a study conducted in 1995 only 0.7% of males and 6% of females were obese and in 2008 a community based study revealed 20% of males and 38% of females are overweighed among these females 10.8% were obese. This can clearly show that the prevalence of obesity is increasing in the country. On the other hand, alcohol consumption of the study group is categorized into two groups moderate (1-21 units) and heavy (>21 units). And 71% of males and 68.6% of females are found to be moderate consumers of alcohol. On the other group 11.8% of male and 0.4% of female were found to be heavy consumers of alcohol (Michelle, 2012).

NCDs are major burdens of developing countries and developed countries and it contributes for 27% of disease burdens in Ethiopia. A recent report of health and health related indicator revealed that hypertension and related diseases are 7^{th} causes of mortality in patient care.

There is also steady increase in proportion of hospital admissions due to stroke and diabetes mellitus (Michelle, 2012).



Source: WHO NCD Country Profiles, 2011

Figure 2.2 Mortality of Ethiopia

2.7 Related works

Courtney indicated that GIS is an important tool for health policy assessment, and the success of policies relies on the analysis on geographical location. The study investigated the distribution of nursing staff, using the GIS-based presentation of data, space and classical statistical data to establish the policies for recruiting and retaining nurses (Courtney, 2005). Gregorio et al. indicated that a cancer map can provide essential clues, including the risk factor for different geographic area and the utilization of disease conditions and clinical services. Such geographic variations may put emphasis on on-site environmental hazards. In terms of the

methods for disease testing, variations in geographic location can reflect the known risk factors of different prevalence rates and the changes in signals. In addition, the study suggested that geographic variations would increase the incidence of breast cancer based on the assessment of visualized representatives, disease incidence, and the potential population. Increases in the incidence rates in two or more neighboring areas may be neglected; and thus it requires more attention (Gregorio et al, 2003).

The allocation of medical resources to different regions and facilities is a becoming an important and complex issue for health authorities in many countries. In order to reasonably meet the needs for community health services and to develop methods for assessing the need for community health services in the field of public medical service, Kaneko et al. used GIS to develop strategies. The study investigated how to allocate resources properly, promote interdepartmental cooperation, increase transparency, and found the need for community health services at the places where services were planned and provided (Kaneko et al., 2007).

Lee et al. used descriptive statistics to divide Taiwan into remote areas and non-remote areas. The analysis results indicated that subjects living in non-remote areas possess higher incomes, receive more education, obtain better medical health services and experience relatively fewer chronic diseases. On the contrary, those living in remote areas experience more chronic diseases. In addition, age and lifestyle are also important factors, which result in regional differences and income-related inequalities. Therefore, regional differences will affect the differences in socioeconomic status and psychological health status (Lee et al., 2007).

Soares et al. proposed a new approach for classification in the intending to provide information on the formulation of policies for regional development. The major indicators of the classification are the population distribution, the income, health status, literacy, employment opportunity and culture. It divided Portugal into four areas with different levels of development, which reflects the finding about asymmetry, coastal area, and interior area (Soares et al., 2003). Lavrac et al. proposed an innovative method where data mining and

visualization technology were used to support decision-making concerning public health and care in Slovenia. The purpose of using data mining and statistical methods to analyze the databases of public health institutions is to confirm the availability and accessibility of public health services in atypical areas. The results can be used to develop health authorities' health care policies (Lavrac et al. 2005).

The Arkansas Data Network Group Health Cooperative stratifies its patient populations by demographic characteristics and medical conditions to determine which groups use the most resources, enabling it to develop programs to help educate these populations and prevent or manage their conditions (Kincade, 1998).

Therefore all the literature reviewed showed that geographic environment, air to be breathed, a kind of food eaten, life style and other factors are determinant for chronic disease distribution and prevalence. At the same time we are going to use drug distribution data of two to three years to mine and to identify the patterns that is hidden in the database of Ethiopian Pharmaceutical Supply Fund Agency.

To aid healthcare management, data mining applications can be developed to better identify and track chronic disease states and high-risk patients, design appropriate interventions, and reduce the number of hospital admissions and claims (Hian Chye Koh et al, 2005).

For example, to develop better diagnosis and treatment protocols, the Arkansas Data Network looks at readmission and resource utilization and compares its data with current scientific literature to determine the best treatment options, thus using evidence to support medical care (Kincade, 1998). Also, the Group Health Cooperative stratifies its patient populations by demographic characteristics and medical conditions to determine which groups use the most resources, enabling it to develop programs to help educate these populations and prevent or manage their conditions. Group Health Cooperative has been involved in several data mining efforts to deliver better quality healthcare at lower costs. In the Seton Medical Center, data

mining is used to decrease patient length of stay, avoid clinical complications, develop best practices, improve patient outcomes, and provide information to physicians to maintain and improve the quality of healthcare (Dakins, 2001).

Chapter Three

3. Research design and methodology

In order to achieve the desired results, appropriate research design and methodology is mandatory. There are various data mining techniques presented in literature review, detailing their suitability for application in the health sector. To develop a chronic disease distribution map of the country, based on drug distribution data using data mining tools and techniques, the Hybrid Process Model has been implemented in the research. The Hybrid Process Model is appropriate because as it is mentioned in literature review it is appropriate for academic research, as it has been identified in academic papers as being useful for has been chosen for its suitability for this research, starting with problem understanding. WEKA 3.6 data mining tools and techniques are used to address the problem.

The methodology used in this study which is Hybrid Process Model has three major phases. Phase one is Data understanding and collection. Phase two is Data preprocessing and Phase three is Modeling and data analyzing and strategies development. The methodology is described in the following figure.

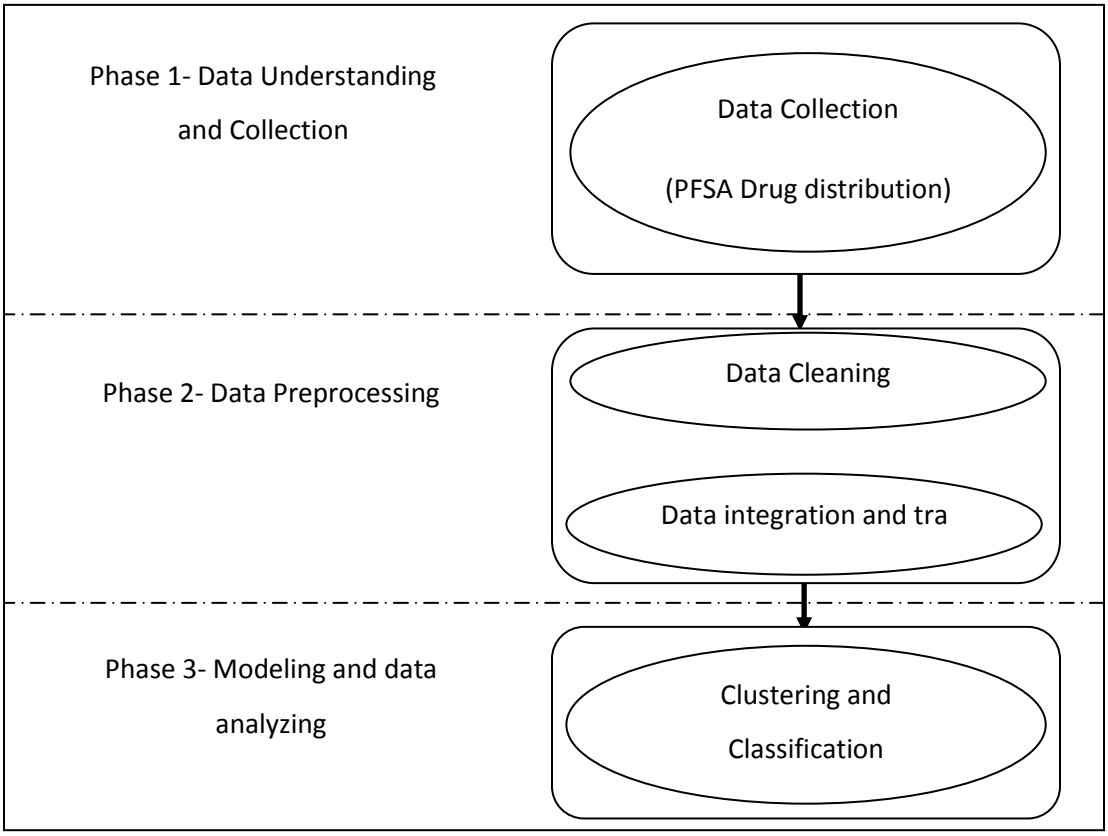


Figure 3.1 Hybrid process model diagram

3.1 Data understanding and collection

After conducting a critical survey on how PFSA works by reviewing related documents and interviewing, it is attempted to thoroughly understand how supply chain management systems work in the organization to supply drugs for various administrative regions. Various data mining concepts, theories, tools and techniques are also explored to use them as a means of solving the problem. The main objectives of the research are identified and try to map to data mining problem. In order to fully understand the way drugs are distributed in the organization document analysis, observation and group discussion with domain experts has been undertaken. Based on the understanding gained, the research problem has been formulated.

Based on this, the research objectives in relation with how the data mining techniques can solve the problem are designed.

Day to day health care activities are highly dependent on medical instruments and drugs. Health facilities ranging from rural health clinic to large specialized hospitals are in need of medical instruments and drugs from a simple to a more complex and advanced medical devices drugs. PFSA is structured to ensure supply of medical equipment and essential drugs to the country. Preparedness to logistics and fund related to medical equipments and drugs supply are administrated by this organization. The structure of PFSA is: there is Central PFSA which is responsible to procure and distribute drugs and medical equipments. There are eleven PFSA Hubs located in different parts of the country. The central PFSA distributes medical equipments and drugs to these HUBS when they request for drugs and medical equipments. Then these HUBS again distribute drugs and medical equipments to facilities and Health Centers which found under them.

As it has been discussed PFSA is responsible to distribute all kinds of pharmaceuticals and health commodity to the country. However, for the study the researcher is looking for specific types of drugs distributed to the HUBs. In order to fully understand what type of drug is prescribed for selected chronic disease the researcher reviewed list of Medicines for Ethiopia recent edition by Food, Medicine and Healthcare Administration and Control Authority of Ethiopia. As shown in the annex I of this document list of categorized drugs that are prescribed for each selected chronic disease are prepared.

There are some drugs that can be prescribed for more than one chronic disease in the list given by Food, Medicine and Healthcare Administration and Control Authority of Ethiopia. For example, Captopril can be prescribed for Cardiac failure and Hypertension and there are similar cases for other drugs also. Based on the advice gained from the physicians and by closely working with people working in the area, these kinds of drugs are categorized to the area where they are prescribed most often.

Since the establishment PFSA has used two different types of computerized databases. In this study, the researcher used the CDS and HCMIS as one of the sources for the data mining process. All medical equipment and drugs issued to Hubs are recorded in one central (PFSA central), primarily for administrative and financial control.

Information regarding the distribution of selected drugs was extracted from PFSA computerized drug issue records for selected six chronic disease of interest. PFSA is managing its distribution records in a system known as CDS and HCMIS (Health Commodity Management Information System). The CDS files contain drug distribution data and the following attributes: Item ID, Item Name, Item unit, Transferred amount, Unit cost, Expiry date, Transfer number, Transferred form, Transferred date, Transferred to branch and Item source. ID codes of each drug were used to identify selected drug. On the other hand HCMIS files contain the following attributes: Item ID, Item Reference, Item Manufacturer, transferred date, Item batch number, Expiry date, Amount received, Amount issued, Balance, and transferred to branch.

Joining the two databases (CDS and HCMIS) on the unique drugs ID number (attribute) that allowed the researcher to create a data set of selected drugs with selected chronic disease along with the type and amount of drugs distributed to different PFSA hubs that they have received during the study period. The drug issue that was extracted from the CDS database initially included 29,623 records of all drugs supplied for 11 PFSA hubs of the country and HCMIS database contains 1,495 records of selected drugs supplied for the hubs. These datasets were first arranged and organized independently, and later they were integrated or merged together for the sake of developing model on a single datasets.

The researcher used PFSA drug issue data for selected diseases from drug issue database. As per the research design methodology, datasets from 2003-2005 E.C were included in the study. All together the dataset has 31,118 numbers of record and 11 numbers of attributes before applying data preprocessing techniques. Whereas the data which is used for mining technique has a collection of 4,894 numbers of records and 7attributes collected from 2003-2005 E.C.

3.2 Data preprocessing

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 is carried out in order to get quality data. Data quality is critically important in the application of data mining technology as the reports can only be as good as the data they are based on. Therefore to get the best out of data mining algorithms high quality data is essential. Prior to data modeling, the required data should be prepared in a better quality in the required format.

Data preprocessing activities are the most important parts of the process for data quality. In general there are four ways of data preprocessing techniques to change the data and improve its quality so that it suits the requirements of the data mining tools and algorithms. The first and major step is data cleansing. It is applied to the dataset by filling in missing values, removing or smoothing noisy data, identifying and removing outliers and inconsistent data in the dataset. The second part is data integration: It is used to merge the data from multiple data sources (different databases) into coherent datasets. The third is data transformation: It is used to transform and combine data into formats that are appropriate for mining activities. The last but not the least is data reduction. It is used to reduce the data size and dimensionality by aggregation, elimination of redundant features, or clusters.

3.2.1 Data cleaning

As described above after the data have been collected from both databases the merging process took place, the process of data cleaning and preparation for allowing to apply data mining algorithms where followed. The data cleaning task is highly dependent on what type data mining algorithm and technique is chosen i.e. clustering technique require certain type of data quality while classification require another type of data quality. But for better prediction and description all data mining techniques require well organized preprocessing and cleaning.

In the data cleaning stage, duplicate records were removed and missing attributes were completed by using appropriate data (e.g. with the help of staff working in the organization missing drug destination data were filled). Records with missing drug name, destination and/or date attributes were completely removed from the dataset. There are some noisy data which were found in the dataset such as texts inputs in the field of number inputs. These kinds of data were completely removed from the dataset to keep the quality of the dataset at required standard. There is no outlier data found in the dataset. The data cleaning task is undertaken so that the dataset quality meets the requirements of selected data mining software, Weka 3.6 and selected data mining algorithm such as simple-K-Means algorithm for data clustering and J48 and naivebayes of classification algorithms.

Hence, in the selected processing application there are major data mining functionalities incorporated for preprocessing, classification, association, clustering, attribute selection and visualizing. To use the drug distribution dataset for data mining applications the researcher checked the quality of data. The selected data mining tool, Weka, does not allow any inconsistency among attribute values. To mitigate this problem considerable amount of time was spent in checking this consistency.

3.2.2 Data integration and data transformation

According to Han and Kamber (2006), data mining techniques sometimes require data integration or the merging of data from multiple data sources and transformation. The data may also need to be transformed into forms appropriate for mining. In this study there are two data sources which the organization used namely CDS and HCMIS (Health Commodity Management Information Systems) both systems sources use the same standard output forms which is .xls. Therefore there is no problem faced during data integration. The next issue is transforming some of the values to a manageable range.

However, the output format i.e. .xls format is not suitable for weka. Therefore the data need to be transformed into formats that are appropriate for weka 3.6. 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. The process of data transformation includes smoothing (e.g. using bin means to replace data errors), Normalization, where the attribute data are scaled so as to fall within a small specified range (scaling the data inside a fixed range), and Attribute construction, where new attributes are constructed and added from the given set of attributes to help the mining process (Chackrabarti et al, 2009).

Data integration involves combining data from CDS and HCMIS sources, which are stored using completely different technologies and they provide a unified view of the data. Data integration is important in case of merging the two systems to proceed with analysis. Consolidating these two separate datasets within one dataset to provide a unified and consistence dataset is vital task of this stage. The integration is one of the most important stages of data. Data is fed from both sources into a single dataset. As the data is fed it is converted, reformatted, summarized, and so forth. The result is that dataset that produced finally for farther analysis has a single physical corporate format.

There are some inconsistencies in some attributes and in their values for example In ID attribute values were written in small and capital letter for the same data element. And some values were written in abbreviations and others were written in full format. There were also cases like Gullele PFSA Hub and GULL GF HAPCO for indicating the same hub. So the researcher discussing with the domain experts made all attribute values to have similar formatting styles and spelling in both datasets to correct the inconsistency among data elements throughout the datasets. Similar procedures were followed for all attributes that are selected by the researcher.

The data from both databases are collected using the same output format. Initially the data were in .xls file format. The data in .xls file format are then exported to Microsoft Excel Comma delimited format (.CSV) for further preparation and to make it ready for WEKA 3.6 data mining tool. This initial dataset is described and visualized using Microsoft excel and SPSS to examine the properties of the dataset in gross. Simple statistical analysis frequency is performed to verify the quality of the dataset such as missing values, error values and to obtain high level information regarding the data mining questions.

In order to get the preferred software i.e. Weka to understand the dataset the researcher needs to prepare the dataset in some Weka understandable formats. The researcher first saved the Excel file into Weka acceptable comma separated values (CSV) or comma delimited file format. Weka native data format is known as the ARFF (Attribute Relation File Format). It is basically a CSV (comma separated value) format with some extra headers to specify what type each attribute is (numerical, binary, nominal). Then after, the CSV file format is converted into ARFF file format by using Weka mining software, to take advantage of easier data manipulation and also compatible interaction with Weka software. During scan of the preprocessed data some basic statistics summary has been produced for each attributes. For categorical attributes, the frequency for each attribute value is shown. Moreover, the data that was converted into ARFF format is passed through important steps of data preprocessing mentioned in the previous section.

3.2.3 Attribute selection and statistical summary of attributes

There are 11 attributes to that are extracted from the first database i.e. CDS. These attributes are directly related with drug are issue to Hubs. These attributes are: “Item ID”, “Item Name”, “Item Unit”, “Amount Transferred”, “Unit Cost”, “Expiry Date”, “Transfer Number”, “Transferred From”, “Transferred Date”, “Transferred To” and “Item Source”.

There are 11 attributes to that are extracted from the second database i.e. HCMIS. These attributes are directly related with drug are issue to Hubs. These attributes are: “Type”, “Reference”, “Manufacturer”, “Date of issue”, “Batch number”, “Expiry date”, “Received”, “Issued”, “Balance”, and “Hubs”.

Attribute “Type” is used only to identify for what kind of chronic disease the drug can be ordered. Because one drug can be ordered for more than one type of chronic disease. “Reference”, is used to identify the transaction for example is it beginning balance or new balance. “Manufacturer” is used to identify who manufactured the drug which means the brand of the drug. “Date of Issue“ is used to identify when the drug is issued to the Hubs, the date when the transaction took place. “Batch number” is used to identify the batch number of the drugs that are issued to the Hubs. “Expiry date” is used to identify the expiry date of the drug that is issued to the Hubs. “Received” identifies the amount of drug that is received initially. “Issued” identifies the amount of drug issued to Hubs in that day. “Balance” identifies the amount of drug on stock after the transaction take place i.e. is drug on hand or on stock. “Hubs” PFSA hubs which are located at 11 different parts of the country, where the drug is sent with that transaction.

No	Attribute	Data type	Description
1	Type	Text	Drug category when it is classified
2	ItemID	Number	Id number to identify what type of drug
3	ItemUnit	Text	Unit of measurement of the drug
4	DateOfIssue	Date	Date when the drug is issued to the hub
5	UnitCost	Currency	The cost of individual drugs
6	Expiry date	Date	Date when the drug is going to expire
7	TransferredQuantity	Number	Amount specific drug transferred to hub
8	TransferNumber	Number	The number given for each transaction
9	TransferredToBranch	Text	Hub where the drug is issued to

Table 3.1 List of attributes and their description selected for mining tool

3.4 Model building

According to Hybrid methodology the next step after data preparation is model building. The major activities that are carried out in building model include selection of modeling technique, generation of test design, model building and model evaluation. In this section, all major aspects of the choice of a suitable methodology for drug distribution data stratification with the PFSA dataset are discussed. Many of these aspects are relevant for other clustering analysis tasks as well.

Following the research procedures different statistical tools designed and used to analyze. So in this research we use a descriptive study that helps to identify the data such as clustering techniques that describes data in terms of numbers and percentages. The next phase helps to deal with the predictive aspects of the study which tries to answer why and when questions of the study.

3.4.1 Selection of modeling technique

3.4.1.1 Descriptive modeling

Selecting specific model for the study is the initial and important step in model building. The selection of modeling technique is based on the objective formulated initially. Since the purpose of the study is to develop a descriptive model to prepare chronic disease distribution map based on drug distribution data, clustering algorithms has been used for building the model. The analysis is performed using SPSS 15.0 environment. SPSS has many clustering algorithms such as Simple K means algorithm, K medoid, Hierarchical clustering and etc.

In this section main principles of the clustering philosophy that are going to be used for the study are discussed as follows. According to Hennig and T. F. Liao (2013), there is no unique objective 'true' or 'best' cluster in a dataset. Clustering requires that the researchers define what kind of clusters they are looking for. The kind of cluster depends on the subject matter

and the aim of the study. Selecting a suitable clustering method requires matching the data analytic features of the resulting clustering with the context-dependent requirements that are desired by the researcher.

The important and difficult task to choose a suitable clustering methodology according to the philosophy that is discussed in the above section is the translation of the desired 'interpretative meaning' of the clustering into data analytic characteristics of the clustering method. This includes various kinds of decisions that cannot be made 'objectively' from the data alone, including the choice of a clustering method, variable transformations and definition of the required concept of 'similarity' and 'dissimilarity'.

The widely used of clustering analysis is classification. That is, subjects or classes are differentiated into groups or clusters such that each subject is more similar to other subjects in its group than to subjects outside the group. Cluster analysis is thus concerned ultimately with classification and represents a set of techniques that are part of the field of numerical taxonomy (Frank and Green, (1968); Punj and Stewart (1983); Aldenderfer and Blashfield (1984)).

In this study the researcher has initially focused on clustering procedures that result in the assignment of each subject to one and only one class and each subject in the class are more similar. Subjects within a cluster are usually assumed to be inseparable from one another. Thus, we assume that the underlying structure of the data involves an unordered set of discrete classes. In some cases we may also view these classes as hierarchical in nature, with some classes divided into subclasses.

The choice of proximity, similarity, association, or resemblance measure of the individual data (all four terms will be used synonymously here) is an interesting problem in cluster analysis. The concept of similarity always connotes the question: similarity with respect to what? Proximity measures are usually viewed in relative terms–two objects are similar, relative to the group, if

their profiles across variables are “close” or they share “many” aspects in common, relative to those which other pairs share in common. Most clustering procedures use pair wise measures of proximity. The choice of which subjects and variables to use in the first place is largely a matter for the researcher’s judgment. While these (prior) choices are important ones, they are beyond our scope of coverage. Even assuming that such choices have been made, however, the possible measures of pair wise proximity are many.

Generally speaking, these measures fall into two classes:

(a) distance-type measures (including correlation coefficients); and

(b) Matching-type measures.

The characteristics of each class are discussed in turn.

3.3.1.1.1 Distance-type measures

A surprisingly large number of proximity measures–including correlation measures–can be viewed as distances in some type of metric space. The notion of Euclidean distance between two points in a space of r dimensions. The formula is:

$$d_{ik} = \sqrt{\sum_{k=1}^{r}(X_{ik} - X_{jk}^2)} \quad \text{..... (3.1)}$$

Where: X_{ik} , X_{jk} are the projections of points i and j on dimension k; (k = 1, 2, …..r). In as much as the variables are often measured in different units, the above formula is usually applied after each variable has been standardized to mean zero and unit standard deviation. Our subsequent discussion will assume that this preliminary step has been taken. The Euclidean distance measure assumes that the space of (standardized) variables is orthogonal, i.e., that the

variables are uncorrelated. While the Euclidean measure can still be used with correlated variables, it is useful to point out that (implicit) weighting of the components underlying the associated variables occurs with the use of the Euclidean measure:

1. Squared Euclidean distance in the original variable space has the effect of weighting each underlying principal component by that component's Eigen value.
2. Squared Euclidean distance in the component space (where all components are first standardized to unit variance) has the effect of assigning equal weights to all components.
3. In terms of the geometry of the configuration, in the first case all points are rotated to orthogonal axes with no change in squared inter-point distance. The general effect is to portray the original configuration as a hyper ellipsoid with principal components serving as axes of that figure. Equating all axes to equal length has the effect of transforming the hyper-ellipsoid into a hyper-sphere where all "axes" are of equal length.

The above considerations can be represented in terms of the following squared distance model:

$$d_{jk}^2=\sum_{k=1}^{r}\left(y_{ik}-y_{jk}\right)^2 \quad \text{..... (3.2)}$$

Where: y_{ik}, y_{jk} denote unit variance components of profiles i and j on component axis k (k = 1,2,…,r). If one weights the component scores according to the variances of the components (before standardization) the expression is:

$$d_{jk}^2=\sum_{k=1}^{r}\lambda_k\left(y_{ik}-y_{jk}\right)^2 \quad \text{..... (3.3)}$$

Where the k^{th} component variance, or Eigen value. This expression is equivalent to express in original variable space. The above relationships assume that all principal components are extracted. As described earlier, if such is not the case, squared inter-point distances will be affected by the fact that they are computed in a component space of lower dimensionality than the original variable space. In summary, both the Euclidean distance measure in original variable space and the Euclidean distance in component space (assuming all components have been extracted) preserve all of the information in the original data matrix. Finally it should be pointed out that if (in addition to being standardized to mean zero and unit variance) the original variables are uncorrelated, both and *will be equivalent.

3.4.1.2 Predictive modeling

The initial step in model building by using predictive modeling technique is the selection of the best algorithm that best solve the problem. The selection of modeling technique is based on the objective formulated. Since the objective of this phase is to build predictive model for drug distribution dataset the researcher decided to use classification algorithms and naïve Bayesian of classification techniques. The analyses were performed using WEKA environment.

3.4.1.2.1 Decision tree

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

The decision tree algorithm used in this research is J48 algorithm, which is an implementation of the C4.5 decision tree learner. This implementation produces decision tree models. The algorithm uses the greedy technique to induce decision trees for classification. A decision-tree

model is built by analyzing training data and the model is used to classify unseen data. J48 generates decision trees, the nodes of which evaluate the existence or significance of individual features, C4.5 algorithm uses the concept of information gain or entropy reduction to select the optimal split. Suppose that we have a variable X whose k possible values have probabilities p1, p2, . . . , pk . What is the smallest number of bits, on average per symbol, needed to transmit a stream of symbols representing the values of X observed? The answer is called the entropy of X and is defined as

$H(x) = -\sum_j^p Pjlog2^{(Pj)}$ ..... (3.4)

Decision trees and decision rules are data-mining methodologies applied in many real-world applications as a powerful solution to classification problems (Han et al., 2001). 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.

3.4.1.2.1 Naïve bayesian

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 c1, c2, c3,. . . ck in which they have prior probabilities p(c1), p(c2), p(c3). . . p(ck) respectively, and the datasets with "m" number of attributes a1, a2, a3, . . . am and each of the attributes with "n" number of values are v1, v2, v3, . . . vn respectively. Therefore the posterior probability of the class ci occurring for the specified attribute can be shown to be proportional to:

P (ci)*p (a1=v1 and a2=v2 and a3 = v3, . . . am=vm / ci)…… (3.5)

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

P (ci)*p (a1=v1| ci)* p(a2=v2| ci)* p(a3=v3| ci)*…*p(am=vm / ci) …….(3.6)

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 ci 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:

$$P(ci)\prod_{i=1}^{m} P(aj = uj|class = ci) \quad \text{..... (3.7)}$$

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.

Chapter Four

4. Experimentation and result analysis

4.1 Model building

Because of the complex nature of drug distribution and disease distribution data different methods are used to best understand these complexities. In order to fully solve the problem stated in the statements of the problem the researcher has decided to use the mix of statistical analysis and data mining tools for specific problems. So the methodology part has two phases. The first phase uses SPSS statistical tool to classify the dataset into different clusters according to their destination where they are distributed i.e PFSA hubs to which the drugs are issued and according to the type of drug that are issued to this hubs. The second phase of the study uses Weka data mining tool to reveal hidden knowledge in the dataset. This phase is used to answer When, where, and what questions of statements of problems based on the previous phase findings.

4.1.1 Descriptive study of the data

This study used the cluster analysis to analyze the drug distribution data and the clustering effect in the eleven PFSA hubs that are located in different parts of the County. The statistical analyses were performed using the SPSS 15.0 package. For example, the drug distribution for each forms of disease selected for study were divided into eleven clusters in the first experiment and it were divided into five clusters in the second experiment. In the first experiment the clusters where chosen so that all type of chronic disease drugs distribution is evaluated against the hubs they were issued to. This is used to help visualize were what type of drug is sent mostly. The second experiment was conducted to visualize which drug type is more issued to which hub. The pie chart output of the dataset is presented below.

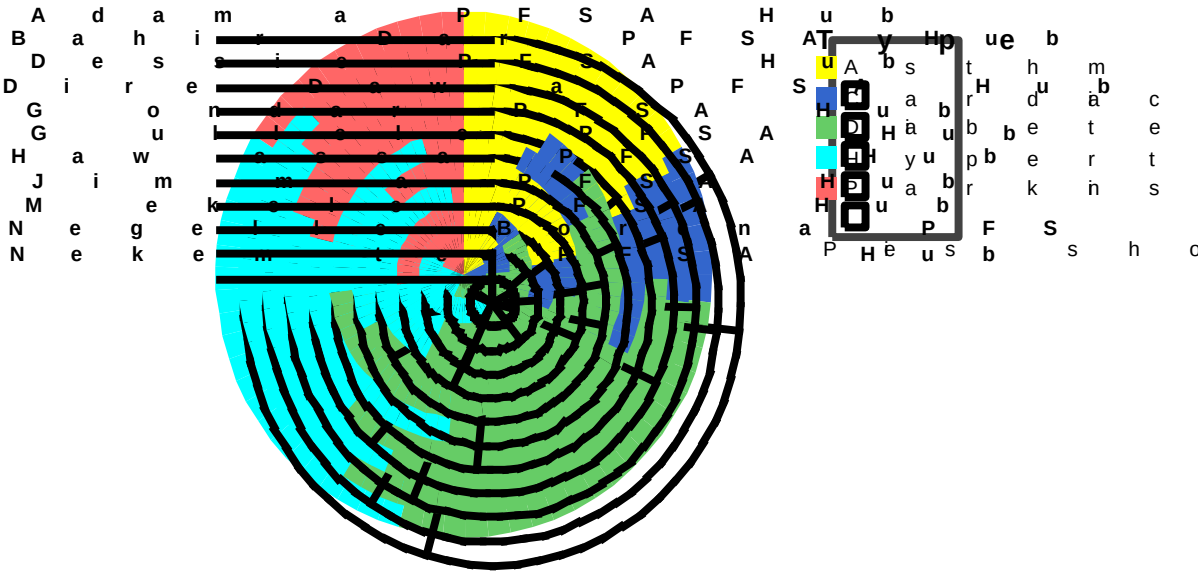


Figure 4.1 Consumption of each chronic disease by hub

In the above chart each hub need against specified types of drug is presented.

The researcher decided to use the K-means clustering algorithm to identify groups of PFSA hubs having similar utilization patterns of each type drugs over the period of three years. The features used by the researcher to characterize each hub were the quantity of drugs issued by central PFSA throughout the study period. All records had three years values for drug utilization by each hub; hence, the researcher clustered vectors having three dimensions i.e. Quantity sent to hubs, Branch (Hub), and Type of drug.

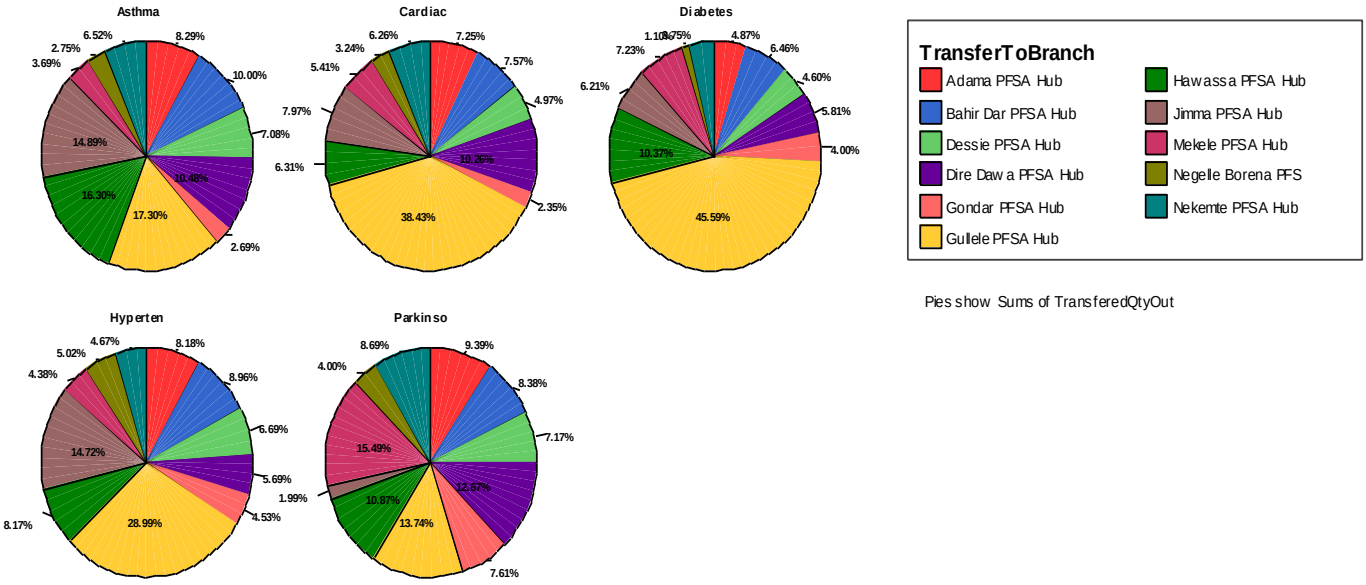


Figure 4.2 Amount of chronic disease drugs issued to hubs organized by type

The results from this clustering have shown that the distribution pattern (representing all drugs of interest) should be partitioned according to the hubs: Based on the above output Asthma, Cardiac, Diabetes and Hypertension drugs are distributed to Gulelle PFSA hub. 17.3 % of Asthma drug were issued to Gullele PFSA Hub next to Hawasa 16.3 % of Asthma drugs was issued to Hawassa PFSA hub and 14.89% were issued to Jimma PFSA hub. When we consider Cardiac again Gullele takes a lead by 38.43%, Diredawa by 10.26% and Jimma with 7.97% took the second and third place. 45.99% of Diabetes drug issued during the study period was issued to Gulele and 10.37 % were issued to Hawasssa. Relatively the other nine hubs received slightly equal percentage of diabetes drugs. Again Gullele received 26.99% of hypertension drugs issued during the study period. Jimmaa with 14.72% and Bahirdar with 8.9% took the second and third position by hypertension. Among the drugs issued for Parkinson Mekele received the height share with 15.49% and Gulele 13.74% and Dire Dawa 12.67%. To visualize the drug distribution according to their destination and type more clearly the following chart gives brief explanation. In the next sub-section, the researcher has be applied classification algorithms to

these clusters with the purpose of discovering the most influential features that affect the drug distribution for the issue dataset.

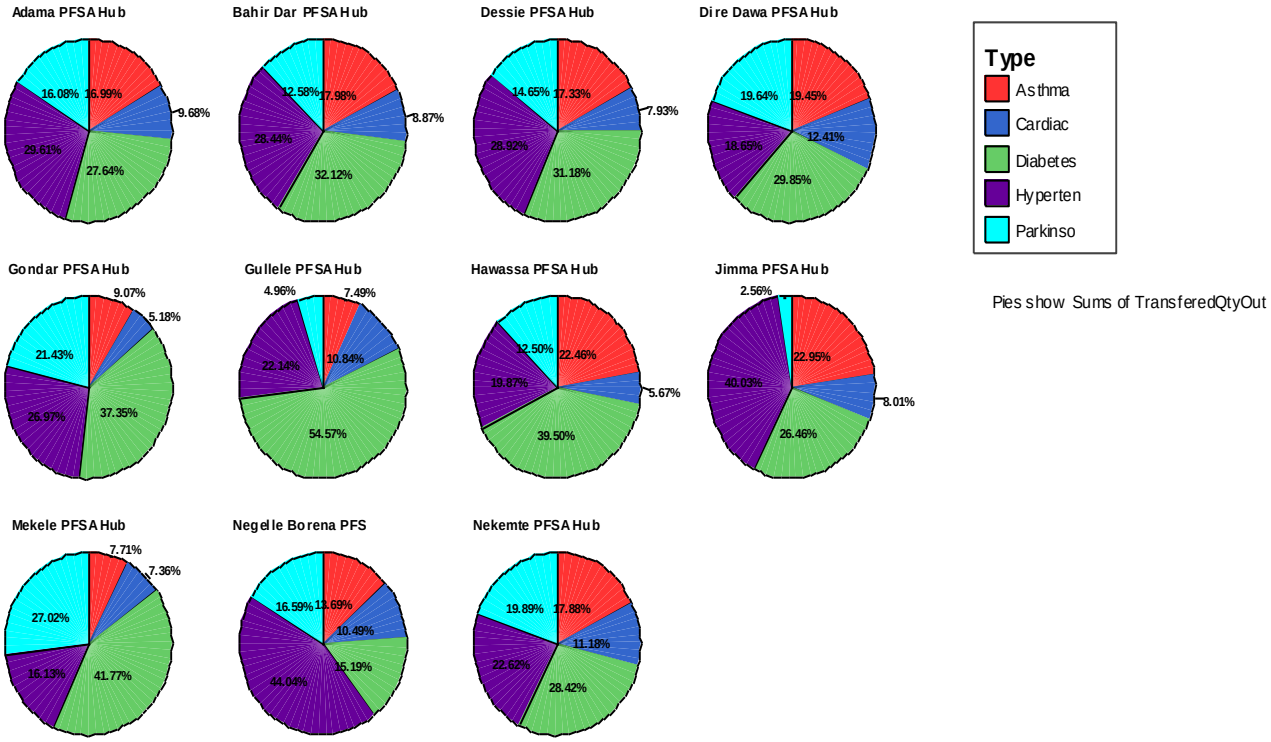


Figure 4.3 Quantity of chronic disease drug type issued to hubs organized by hubs

4.1.2 Predictive study of the data

In most of data mining methodology, model building is an iterative process to discover the knowledge hidden in the dataset. Therefore it is highly recommended to conduct different experiments to identify the best model for tackling the problem in a suitable manner. As it has been discussed in the previous section clustering and classification algorithms of data mining techniques were selected to solve the problem.

Two different classification techniques were selected to build the model namely Decision tree and Naïve Bayesian. And both techniques have different algorithms under them in Weka data mining software. Taking the running time as criteria to select the best performing algorithm,

the researcher tested the performance of each algorithm that is found under both categories. Based on the output shown in the table 4.1 below the researcher selected to use J48 and NaiveBayes for Decision tree and Naïve Bayesian respectively. To get the best out of the model created the researcher has conducted each experiment by using 10-fold cross validation option and 2/3 (66%) of the data for training data set and the rest 34% data as test dataset.

Decision Tree	Running Time
J48	0.33 Sec
BFTree	79.56 Sec
LADTree	113.59 Sec

Decision Tree	Running Time
BayesNet	0.09 Sec
NaiveBayes	0.03 Sec
NaiveBayesSimple	0.03 Sec

Table 4.1 Running time test for algorithms

In the second stage, after the researcher has found the best clustering and assigned each hub is assigned against drug type in the dataset to appropriate cluster, the researcher has applied supervised classification algorithms to the data and thus identifies the features and the models that can identify the best attribute, when the drugs are issued and the drug utilization of each hubs. The available attributes are date of issue, type, expiry date; quantity issued, unit cost, unit of measure, drug ID, transferred to branch and transfer number. In the dataset, we have 4,894 records after cleaning during the study period and nine attributes of interest. The value in the column of every drug type is the number of issue of the specific drug type out of all drug types issued to the hubs.

After the data has been prepared, the data mining method used to build the model is classification technique to classify similar classes into their own classes. The data analysis is processed using WEKA 3.6 data mining tool for exploratory data analysis, machine learning and

statistical learning algorithms. The researcher conducted the experiments using different tools for analyzing statistical aspects of the dataset. J48 and naivebayes classification algorithm was used to classify the dataset into its position for each of these experiments. The whole dataset is first divided into training dataset and test dataset. And two options were selected to conduct the experiment namely 10 fold cross validation and 66 percent split options the dataset list is then divided into two parts that is 66% of the data are used for training and 34% are used for testing. Accordingly the training data set consists of 3,230 instances with 9 different attributes for 66% split option. The instances in the dataset are representing the results of different types of testing to predict the accuracy of the assignment of each class into correct clusters. The results of each experiment are discussed in the following sections. For each experiment, accuracy was evaluated using 10-folds cross validation, and 66% method. The accuracy percentages for each of these techniques were evaluated using table.

A dataset is said to be imbalanced if the classification categories are not approximately represented equally. And performance of machine learning algorithms is typically evaluated based on how accurately predicted (Nitesh et al., 2007). However, this is appropriate when the dataset is imbalanced and/or the cost differences of error are large. The final dataset after cleaning is imbalanced if the classification categories are not approximately equally and there is high imbalance. Therefore, the researcher used SMOTE Automatic operations by filter where minority classes are oversampled by generating synthetic examples of minority class and adding them to the dataset. This way, the class distribution in the dataset changes and probability of correctly classifying minority class increases.

4.1.1 Predictive model using J48 algorithm

Decision tree J48 classification algorithm has been applied on the dataset on hand to build classification model. The classifier has options that are highly interactive and can benefit the

researcher for powerful visualization features. The experiment on the decision tree predictive model building was based on the dataset from PFSA drug issue dataset, which has been preprocessed and introduced to the Weka software.

To build the predictive model, J48 and Naïve bayes decision tree algorithms are trained and evaluated. In order to identify the best classifier parameters using J48 algorithms a number of experiments has been undertaken. Conducting different experiment by changing the parameters and their values are important to produced optimal and accurate decision tree with minimized incorrectly classified instances.

After building four classification experiments, it has been observed that the performances of the models are not the same. Thus, as indicated in the methodology part, based on the measures of performance and interpretability an evaluation is made by comparing these models.

Experiment	No Attr.	True +ve rate		False +ve rate		Accuracy of the model	
		10 fold	66% Split	10 fold	66% Split	10 fold	66% Split
E1	9	97.3%	95.9%	0.4%	0.5%	97.28%	95.92%
E2	9	94.5%	94.2%	1.7%	1.8%	94.52%	94.77%
E3	7	96.4%	94.7%	0.4%	0.6%	96.34%	94.7%
E4	7	98.3%	94.7%	0.2%	0.6%	98.32%	94.7%

Table 4.2 Summary of J48 decision tree experiment results

The result of experiments conducted using J48 algorithms of decision tree in the above model shows that all attributes included in the dataset i.e. 9 attributes including dependent variable, the decision tree algorithm selected all of the nine variables as important.

The model has accuracy of 97.28% using 10 fold cross validation and 95.92% accuracy using 66% split test options. Moreover the model has a true positive rate of 97.3% and false positive

rate of 0.4% for 10 fold cross validation and 94.3% true positive rate and 0.5% false positive rate for 66% split test. After changing the class from transferred branch to type of drug the second experiment was conducted using all the same attributes as the first experiment. In the second experiment the researcher also used J48 classification algorithm of decision tree. The second experiment also has high performance though the output is slightly different. The accuracy of the second experiment is 94.52% using 10-fold cross validation and 94.77% using 66% split option. The true positive rate of 10-fold cross validation is 94.5% and 1.7% false positive rate. While using 66% split option the true positive output 94.2% and 1.8 false positive, the output of these two algorithms are similar. The third experiment was conducted using 7 attributes by reducing two attributes from the original dataset that are considered by the researcher, namely Drug ID and Transfer number was reduced. The 10 fold cross validation experiment scored 96.34% accuracy, 96.4% true positive rate and 0.4% false positive rate. Whereas the 66% split test scored 94.7% accuracy, 94.7% true positive rate and 0.6% false positive rate. The fourth experiment is conducted using the same reduced attributes because the third output has shown improvement. The class is then changed back to the branch transferred to because the first experiment has shown better performance than the second and the third. As expected the performance of the fourth experiment is better than all other outputs. The accuracy of 10 fold cross validation was 98.54% and the records that are classified as true positive is also 98.5% false positive records also decreased to 0.2%. Whereas using 66% split option the accuracy was 94.7%, true positive 94.7% and 0.6% false positive was recorded for this experiment.

The best J48 decision tree classification model with high accuracy and correctly classified classes was produced on the fourth experiment. The model shows a better performance evaluation than other three models. The 66% split test model also scored the lowest performance than 10 fold cross validation. Therefore, the 7 attributes, used to build the decision tree for experiment 4 with 10 fold cross validation option, was taken to be statistically significant in splitting the decision tree. Furthermore, opinions collected from the domain

experts from PFSA indicated that selected attributes have a crucial role in the distribution of drugs to the hubs.

Figure 4.4 shows the tree view of the predictive model built using J48 algorithm using 9 attributes. For clear understanding of the tree, the run information for the model is annexed at annex II.

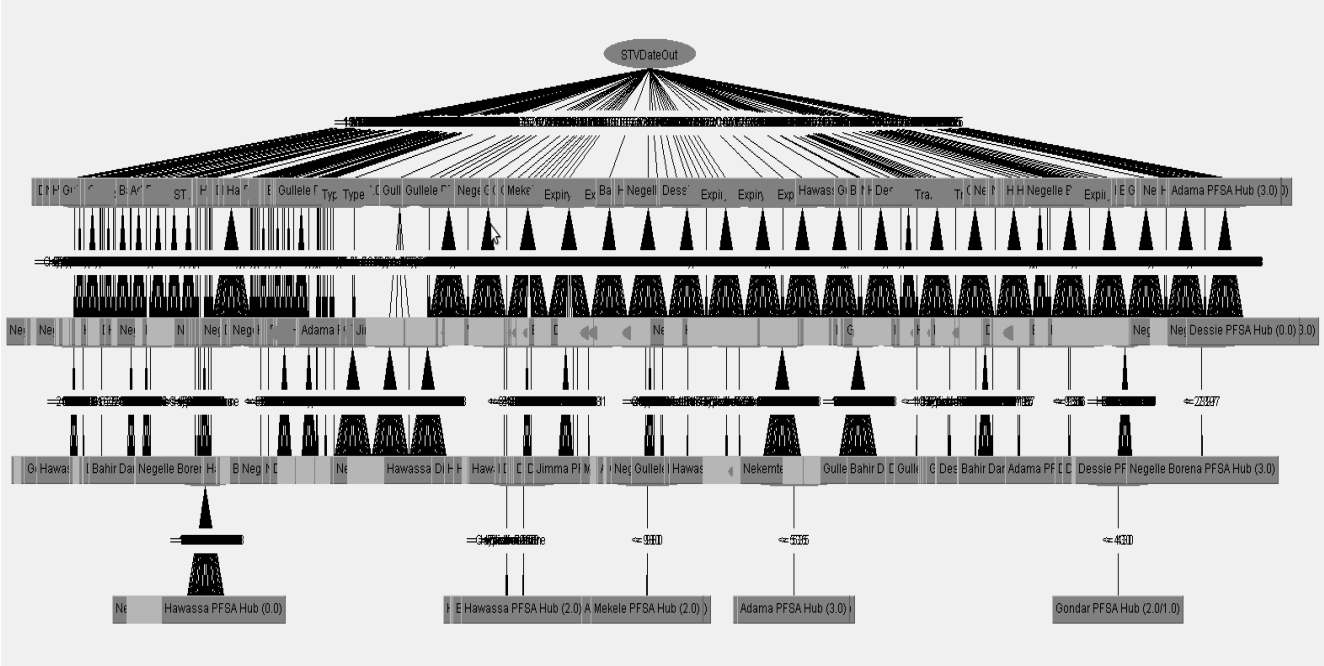


Figure 4.4 Tree view of experiment 4

Confusion matrix for J48 decision tree classifier

The confusion matrix is useful for analyzing how well the classifier recognizes records of different classes. A confusion matrix displays the number of correct and incorrect assignment made by the model compared with the actual classifications in the test dataset. The matrix is n-by-n, where n is the number of classes. The confusion matrix for the decision tree shown in table 4.3 illustrate that out of the total 4,894 records 4,823 records were correctly classified into their correct class i.e to correct hub. Whereas 71 records where classified to different class than class they deserve to be classified in. In general the classifier correctly classified 4,823

records and incorrectly classified 71 records out of a possible 4,894. The accuracy of the classifier to correctly predict the class value into their own class is 98.54%.

Confusion Matrix											
a	b	c	d	e	f	g	h	i	j	k	← Classified as
876	2	0	0	0	2	5	0	0	0	2	a
6	424	0	0	1	0	0	0	3	0	0	b
2	0	358	0	0	0	2	0	0	0	0	c
3	3	0	239	0	0	0	3	0	0	0	d
0	3	0	0	360	1	3	1	0	2	0	e
2	0	3	0	0	623	0	0	3	0	0	f
3	0	0	0	3	0	401	0	6	2	0	g
0	0	0	0	2	3	0	378	0	0	0	h
0	1	0	0	0	4	1	0	433	3	0	i
2	0	0	0	0	0	0	0	2	409	0	j
1	0	0	1	0	0	0	0	0	0	311	k

Table 4.3 Confusion matrix for J48 Decision tree model

Key: a → Gullele PFSA Hub

b → Dessie PFSA Hub

c → Mekele PFSA Hub

d → Gonder PFSA Hub

e → Nekemte PFSA Hub

f → Hawassa PFSA Hub

g → Adama PFSA Hub

h → Bahirdar PFSA Hub

i → Dire Dawa PFSA Hub

j → Jimma PFSA Hub

k → Negelle Borena PFSA Hub

In order to understand how well the algorithm has categorized each record into their right full place the sensitivity and specificity measures of the model are very useful. Sensitivity is also referred to as the true positive (recognition) rate, while specificity is the true negative rate.

4.1.2 Predictive model using Naïve Bayias algorithm

Naïve bayes classification modeling technique was selected to build the second experimental model. This classifier assumes that the presence or absence of a particular feature is unrelated to the presence or absence of any other feature, given the class variable. Four experiments were conducted with the naïvebayes algorithm using different combination and numbers of attributes. As the previous experiment 10 fold cross validation and 66% split to train a model was used to test the performance of the selected algorithm and model. Table 4.4 summarizes the result obtained from these experiments with their respective accuracies, true positive rate and true negative rate.

Experiment	No Attr.	True +ve rate		False –ve rate		Accuracy of the model	
		10 fold	66% split	10 fold	66% split	10 fold	66% split
E1	9	68.3%	69.4%	3.4%	3.2%	68.34%	69.35%
E2	7	73.1%	72.6%	2.9%	2.9%	73.1%	72.69%
E3	7	72.2%	70.9%	3.4%	3.5%	72.19%	70.85%
E4	6	69.5%	70.3%	3.7%	3.6%	69.53%	70.31%

Table 4.4 Summary of naïve bayes experiment results

From table 5.3 the first experiment using Naïve Bayes used all 13 attributes and the model scored 69.34% accuracy, 698.3.1% true positive rate and 3.4% false negative rate using 10 fold cross validation and 69.35% accuracy, 69.4% true positive rate and 3.2% false negative rate with 66% split option. The second experiment the model was built using 9 attributes by reducing

issue number and drug ID attributes from the first experiment to see the difference. The scored the best accuracy of 73.1% with 10 fold cross validation and 72.69% using 66% split test option. The third Experiment has also been carried out with 7 attributes by reducing expiry date and drug ID. This experiment has scored 72.19% accuracy by 10 fold cross validation option and 70.58% using 66% split test option. The last and fourth experiment has been under taken using 6 attributes by reducing unit of measurement from the previous experiment. This experiment has 69.53% accuracy by 10 fold cross validation option and 70.31% using 66% split option. Therefore from all the four experiments undertaken Experiment two with 7 attributes were chosen to have best performance evaluation for naïve bayes classifier. Because it has 73.1% accuracy using 10 fold cross validation option and 72.69% accuracy using 66% split option.

Confusion matrix for Naïve Bayes classifier

Table 5.4 below shows the confusion matrix of the best model built using naïve bayes experiment i.e. experiment two. The confusion matrix shows that the amount of records that is correctly classified in their true class and amount of records that are incorrectly classified by the classifier into wrong class. The output of 10 fold cross validation shows that 3,578 records of the whole dataset was classified into their correct class with percentage of 73.10% while 1,316 records were classified into wrong classes with percentage of 26.89%.

Confusion Matrix											
a	b	c	d	e	f	g	h	i	j	k	← Classified as
586	22	30	17	40	33	18	22	21	74	24	a
20	365	12	7	3	9	8	2	8	8	0	b
0	36	274	21	2	2	8	0	6	0	12	c
12	2	4	162	1	11	9	11	6	14	14	d
10	3	0	0	277	23	10	8	3	26	0	e
0	0	14	4	8	474	0	5	10	22	19	f
13	4	3	8	0	0	257	1	9	18	15	g
11	0	2	6	9	0	1	275	7	25	1	h
1	4	8	12	0	3	1	2	299	30	3	i
1	1	5	8	9	7	2	0	0	339	1	j
45	15	25	1	49	70	44	1	3	2	217	k

Table 4.5 Confusion matrix of naïve bayes algorithm

Key: a → Gullele PFSA Hub

b → Dessie PFSA Hub

c → Mekele PFSA Hub

d → Gonder PFSA Hub

e → Nekemte PFSA Hub

f → Hawassa PFSA Hub

g → Adama PFSA Hub

h → Bahirdar PFSA Hub

i → Dire Dawa PFSA Hub

j → Jimma PFSA Hub

k → Negelle Borena PFSA Hub

In order to identify how well the classifier has identified the records into their own class the sensitivity and specificity measures were used.

4.3 Experimental results and discussion

As the purpose of this research is to find out the chronic disease distribution map and their determining factors affecting the drug distribution of the country that enables to map these selected chronic disease were identified and promising results that demand action or further research are discussed in this section. The results are obtained using the application of simple statistical tool and cluster drug distribution data and data mining algorithms that enabled the researcher to identify the factors affecting the this distribution.

As shown in the Figure 4.1, Figure 4.2 and Figure 4.3, the clusters were formed by eleven hubs based on their consumption of selected five chronic disease drug types. The highest portion of Diabetes drug distributed during the study period which was issued by the central PFSA to the hubs in the County was received by Gullele (45.55%). This hub also received the highest portion of Cardiac (38.43%), Hypertension (26.99%), and Asthma (17.30%), suggesting that these chronic diseases are consumed the most in the County. The Asthma drugs were issued to Hawassa (16.30%) and Jimma (14.89%) next to Gullele revealing that Asthma is the most important factor affecting public health of these areas. It is also interesting to note that the distributions of these drugs in remaining areas are more even. Dire Dawa (10.25%) of Cardiac drug distributed during the study period. Highest portion of Parkinson drug was sent to Gonder (15.45%) hub followed by Gullele (13.74%) and Dire Dawa (12.67%). The rest of hubs received slightly similar amount of drug during the study period.

The occurrence of diseases is highly correlated with the race, age and lifestyle of the people. Ford et al. (2002), found in a survey that the adult population of the United States has a common metabolic syndrome, while race, age, obesity, and lifestyle, including smoking and physical activity, affect the prevalence of metabolic syndrome. The study of the geographic

environment and public health information can improve the disease prevention and health plans, and effectively control the increasing medical costs (Zhan, 2006).

It can be inferred from the analysis different hubs consume different type of chronic disease drugs than other drug. As discussed above different factors such as race, age, obesity, life style, geographic environment, smoking, physical activity and can affect the health of the society especially chronic diseases. This difference in consumption of chronic disease drug might be due to difference in race, geographic environment and habit of physical exercise of locations of hubs. For example the life style, living standard, and environment of people living around Addis Ababa and Dire Dawa has much difference and this obviously contributes to their health practice. In general, hypertension and Diabetes are major chronic disease drugs that are consumed by most hubs in the country's population in different parts. Therefore, in order to promote public health correct dietary and lifestyle habits should be advised through health education.

As mentioned in the previous section after discussing the descriptive part of the dataset the researcher tried to find out the reason behind the irregular distribution of these drugs using data mining technique, the tool that has power to learn the dataset and infer or predict some values.

The quality of the classifier algorithm is usually measured by the accuracy rate which means the rate of correctly classified instances. The following table 4.5 explains the results obtained during the classification of test dataset based on the model created using training dataset. The table contains number of correctly classified instances, incorrectly classified instances, Kappa statistics, which is a measure of the degree of nonrandom agreement between observers or measurements of the same categorical variable; Mean absolute error which is a quantity used to measure how close forecasts or predictions are to the eventual outcomes; Root mean squared error, which is a frequently used measure of the differences between values predicted by the model or an estimator and the values actually observed; Relative absolute error which is

the magnitude of the difference between the exact value and the approximation; Root relative squared error which is relative to what it would have been if a simple predictor had been used. Simply it is the average of the actual values.

Algorithm / Classifier	Correctly Classified Instances	Incorrectly Classified Instances	Kappa Statics	Mean Absolute Error	Root mean squared error	Relative absolute error	Root relative squared error
J48 Decision Tree	4,812	82	0.9813	0.0033	0.0502	2.0062%	17.6013%
Naïve Bayesian	1,157	507	0.662	0.0832	0.2056	51.04%	72.0%

Table 4.6 classification summary of both experimental models

The weka data mining software selected DatesentOut as a best attribute for the classification. And all classification rules is generated according to the dates the drugs are sent to the hubs. This shows that certain types of drugs are more important than the other type at specific hub. According to the rules generated by the classification techniques, which is annexed in annex II, the dates at which each drugs are more important at specific hubs are listed. This phenomenon was not noticed by domain experts before this study but the domain experts agreed that the dates have impact on distribution of drugs.

The other best attributes selected by the algorithms are Expiry dates of the drug and the transfer number of the transaction. But this attributes were not accepted by the domain experts and the researcher as well. Because these attributes have no connection with the

transfer of the drugs. These attributes are only used to identify the drugs among the other drugs that are distributed by the organization.

The researcher has also tried to find the reason why large amount of certain types of drugs are sent to one hub while the small amount of drugs of same type of drug is sent to different hub. But the attributes that were collected from the PFSA doesn't include necessary attributes to analyze the reason. The attributes that are useful to analyze the reason can be found at facility level. Given the time and resource the researcher couldn't get this data from facility level of the PFSA hubs.

Chapter Five

5. Summary, conclusions and recommendations

5.1 Summary

Drug distribution is the main issue determining the health of the public in the country. Unless there is fast and reliable provision of necessary medical equipments and drugs are ensured in the country, which is based on research and study, many plans that are designed by different organs both governmental and nongovernmental organization would be difficult to achieve. So this study tried to identify what type of chronic drug are more important at what part of the country when? The study can reduce the wastage of highly expensive drugs by identifying and sending them to the hubs.

The research can improve the supply chain management of the organization by enabling the organization to map the type of drugs that are consumed in different hubs of the organization. The study can further contribute to the public health of the country by investigating further causes why the these drugs are consumed in that area and why not in another area.

Different models were built using data mining technique to identify and solve the factors affecting the drug distribution of the country. Suitable data that helps to solve the problem was collect from PFSA databases. The methodology applied in the research was Hybrid methodology, which contains six major phases: understanding of the problem domain, understanding of the data, preparation of the data, data mining, evaluation of the discovered knowledge, and use of the discovered knowledge. A total of 4,894 records were used to conduct experiments for the study. Missing values are reduced from the dataset during preprocessing stages.

The data mining techniques used for predicting analysis was classification algorithms. J48 of decision tree and naïvebayes of naïve Bayesian algorithms were selected for experiments. These algorithms were selected by conducting experiments to evaluate their performance using

the dataset provided. And these algorithms can be validated by using their accuracy by the researcher much more easily.

Several models were built implementing the J48 decision tree classifier algorithm and naivebayes of naïve bayes algorithms. These experiments are done using pruning with all and reduced attributes, by giving J48 classifiers parameters of different values. The researcher compared the classification performance of the decision trees with tree prunning and without tree prunning, and found that tree prunning can significantly improve decision tree's classification performance.

5.2 Conclusion

The research has used PFSA drug issue data for the purpose of describing trends of utilization of selected chronic disease drugs that are distributed to the hubs of PFSA that are located in different parts of the country. The datasets contain relevant information about the transaction of drugs selected by the researcher with the help of domain experts to conduct predictive and descriptive analysis. The data was extracted from the PFSA databases namely CDS and HCMIS that the organization uses for the day today activity. First simple statistical analysis was performed on the dataset to describe consumption of each hub against the type of chronic disease selected. Then after, the same dataset was used to conduct experiment using decision tree and naïve Bayesian algorithms of classification techniques to discover hidden knowledge in the dataset.

After conducting serious discussion with domain experts from the organization the researcher made sure that business understanding stage is finished successfully to go for pre-processing steps such as cleaning of error data, handling missing values and balancing the dataset have been effectively undertaken so that it will help in achieving the final objectives.

After conducting a number of experiments the comparison of those experiments were made using different criteria. K fold cross validation testing was conducted by assigning k=10 as it is suggested by weka. Increasing minimum instances (minNumObj) from two in 10 fold cross validation test decreased the number of correctly classified instances and increased number of incorrectly classified instances and reduced the number of leaves and sizes of the tree. On the other hand, reducing the confidence in the training data (confidenceFactor) does not only reduce the tree size, but also helps in filtering out statistically irrelevant nodes that would otherwise lead to classification errors. So this can lead to the conclusion that several values of the confidence factor should be tested while generating decision trees to find the most suitable value for the training set that is being examined.

In general while conducting the experiments the overall best performance was scored by J48 decision tree classifier setting pruned to true, confidence factor to 0.25, minimum numbers of instance (minNumObj) to 2. This experiment scored highest accuracy with a recall (true positive rate) of 98.4%, a false positive rate of 0.3%, a precision (positive predictive value) of 98.3%, and an accuracy of 98.54%. The analysis of the experimental results shows that the model is quite effective and efficient in classifying the instances.

The second algorithm selected by performance was naïve Bayesian algorithms. The result show accuracy of 69.53% and correctly and Incorrectly Classified Instances are 3,367 and 1,527 respectively and with recall (true positive rate) of 68.8%, a false positive rate of 3.3%, a precision (positive predictive value) of 69.7%.

In general, the results from this study were interesting and encouraging; it can help different health authority to make wise decision. The extracted rules in both algorithms are very effective for the prediction of drug distribution. On both algorithms, it can be observed that the attributes such as Date sent, Transfer number, Expiry date, ItemID, Transferred Quantity, Unit cost, and Item unit are selected as best attributes. However, some of the selected attributes are not accepted by the researcher and the domain experts. For example attributes such as

Transfer number, Expiry date and unit costs are not accepted as a best attributes by the domain experts and the researcher.

5.3 Recommendations

The results of both descriptive and predictive analysis have revealed interesting patterns and knowledge hidden in the dataset of the organization. Especially the data mining technology algorithms implemented such as decision tree of J48 and naïve Bayesian of simpleaivebayes have revealed interesting rules that are very useful in decision making. The selected experimental setups that are conducted for the study was effective and efficient in detecting what attributes are very important in drug distribution and mapping them to selected chronic disease.

The researcher highly believes that the findings of this research can be used by health authorities to further explore the efficient way of drug distribution mechanisms that bases the needs of each hub in Ethiopia. The following recommendations are based on the result of this study and nature of the data set.

- Both experiments conducted using J48 decision tree and naivebayes of naïve Bayesian algorithms produced efficient and effective models and interpretable rules (patterns). Hence it is important for concerned government officials specially PFSA and Federal Ministry of health authorities jointly to deploy the model developed with these data mining technique in order to use as a decision support tool in the distribution of expensive drugs especially when there is scares of drugs.
- The researcher believes that further experimentation needs to be performed to have several occurrences of the model based on which improved, better and efficient models can be obtained with better accuracy measurement. As of this research the data mining algorithms implemented was classification and the class

attribute was supplied by the researcher to weka, but the researcher recommends the use of clustering algorithms so that the mining tool can determine the class by analyzing the relationship between attributes.

- Based on the descriptive study some of the chronic disease drugs are sent much more frequently to one hub than others. This phenomenon requires another further study why this is happening. The researcher highly believes that selected chronic diseases are dependent on the geographic environment and life styles of the society living in different area where the hubs are located. Therefore, to reveal the root causes of this diversity the researcher recommends further study to be undertaken by using appropriate research design.
- Due to time and budget constraints the researcher haven't included necessary attributes from the facility level that could help to answer why and what questions irregularity of selected chronic disease distribution in the research. Furthermore, it is very essential to assess the applicability of data mining techniques in identifying determinant factors by including other data sets from facility level (such as Hospitals, clinics, and health centers). The researcher recommends other researcher to investigate the causes by taking enough time and budget.

References

1. Agrawal , R., Imielinski, T., & Swami ,A. (1993). Mining association rules between sets of items in large databases. Proceedings of the 1993 ACM SIGMOD conference: 1-10.
2. American Diabetes Association, Standards of medical care for patients with diabetes mellitus. Diab. Care , 2002.
3. Azevedo, A. & Santos, M.F. (2008).KDD, SEMMA and CRISP-DM: a parallel overview. IADIS European conference data mining: 182-185.
4. Bernander ,O.(2009). Microsoft Encarta 2009: neural network. Microsoft Corporation.
5. Berry, M. & Linoff , G. (2004). Data mining techniques for marketing, sales, and customer relationship management. (2nd edition).Indiana: Wiley publishing.
6. Bramer , M. (2007).Principles of data mining. London: Springer.
7. Chakrabarti, S., Cox, E., Frank, E. ,Güting, R. H. , Han, J. , Jiang, X. , Kamber, M. , Lightstone, S.S., Nadeau, T. P. , Neapolitan, R. E. , Pyle, D. , Refaat, M. , Schneider, M. , Teorey, T. J. , & Witten, I. H. (2009). Data mining: know it all. Burlington: Morga Kaufmann Publishers.
8. Chapman P., Julian C., Randy K.,Thomas K., Thomas R.,Colin S., and Rüdiger W. (2000). CRISP-DM 1.0 - Step-by-step data mining guide. Available at www.spss.ch/file.php?file=/upload/1107356429_CrispDM1.0.pdf (Accessed on 23 February, 2013)
9. Chen, G ,Liu, H., Yu ,L., Wei, Q., Zhang, X. (2006) .A new approach to classification based on association rule mining. Decision Support Systems 42: 674– 689.
10. Cios Krzystof J., Pedrycz Wiltod., Swiniarrski Roman w. Kurgan Lukasz A. Data Mining: A knowledge discovery approach. New York: Springer-Verlag Science Business Media LLC, 2007
11. Colin Shearer. (2000). The CRISP-DM Model: The New Blueprint for Data Mining. Journal of Data Ware Housing Volume 5, Number 4.
12. Cutchin, M. P., The need for the new health geography in epidemiologic studies of environment and health. Health Place, 2007.

13. Dakins, D.R. (2001). Center takes data tracking to heart. Health Data Management, 9(1), 32-36.
14. Daniel T. Larose. (2005). Discovering Knowledge in Data: An Introduction to Data Mining. John Wiley & Sons, Hoboken, New Jersey.
15. Data Mining Applications in Healthcare, Journal of Healthcare Information Management- Vol.19, No.2
16. Deogun, Jitender S. 2001. Data Mining: research Trends, Challenges, and Applications.
17. Dummer, T. J., Health geography: supporting public health policy and planning. CMAJ, 2008.
18. Fayyad, U Piatetsky-Shapiro, G and Smyth, P, 1996b , "From data mining to knowledge discovery: an overview " In Fayyed, U, Piatestky-Shapiro, G, Smyth, P and Uthurusamy, R(eds) Advances in Knowledge Discovery and Data Mining, AAAI Press, pp. 1-4
19. Granados, R. M., Aguilera, J. D., and Martín, J. J., Estimation of an index of regional health needs in Spain using count regression models with filter. Health Policy 81(1):4–16, 2007.
20. Gregorio, D. I., and Samociuk, H., Breast cancer surveillance using gridded population units, Connecticut, 1992 to 1995. Ann. Epidemiol. 13(1):42–49, 2003.
21. Han, Jiawei and Kamber, Micheline. (2001). Data Mining: concepts and Techniques. San Fransisco: Morgan kufman.
22. Hand, D., Mannila, H. & Smyth, P.(2001). Principles of data mining. Massachusetts: Massachusetts Institute of Technology press.
23. Jiawie, Han and Micheline Kamber. (2006). Data mining Concept and Techniques. 2nd Edition. Morgan Kaufmann Publishers, San Francisco
24. Kantardzic, M. (2003).Data mining: concepts, models, methods, and algorithms. John Wiley & Sons
25. Kincade, K. (1998). Data mining: digging for healthcare gold. Insurance & Technology, 23(2), IM2-IM7.

26. Kotsiantis, S. & Kanellopoulos, D. (2006).Association rules mining: a recent overview. GESTS International Transactions on Computer Science and Engineering, 32(1), 71-82.
27. Krzysztof J., Witold P., Roman W., Lukasz A. (2007). Data Mining: A Knowledge Discovery Approach. New York: Springer-Verlag New York Inc.
28. Kurgan, L. A. & Musilek, P. (2006).A survey of knowledge discovery and data mining process models. The Knowledge Engineering Review, 21(1):1-24.
29. Larose, Daniel T. (2005) .Discovering knowledge in data: an introduction to data mining. New Jersey: John Wiley & Sons
30. Lee, M. C., and Jones, A. M., Understanding differences in income-related health inequality between geographic regions in Taiwan using the SF-36. Health Policy 83(2):186–195, 2007.
31. Michelle A. Williams (2012) Non Communicable Diseases among Working Adults in Addis Ababa.
32. PFSA BPR document, 2010
33. Prather, Jonathan C. et. al. 2001. Medical Data Mining: Knowledge Discovery in a clinical Data Worehouse.
34. Rachel Nugent. Chronic Disease in Developing Countries, Health and Economic burdens. Center for global development, Washington, DC, USA. 2008.
35. Raghavan, Vijav V., et al. 1998. A Perspective on Data Mining. Journal of the American Society for Information Science; 49(5): 397-402.
36. Refaat, M. (2007). Data preparation for data mining using SAS. San Francisco: Morgan Kaufmann Publishers.
37. S. P. Deshpande and V. M. Thakare (2010). Data Mining System and Applications: A Review. International Journal of Distributed and Parallel systems (IJDPS) Volume 1, Number 1: 445-463
38. Soares, J. O., Marques, M. M. L., and Monteiro, C. M. F., A multivariate methodology to uncover regional disparities: acontribution to improve European Union and governmental decision. Eur. J. Oper. Res. 145:121–135, 2003.

39. Summary and Statistical Report of the 2007 Population and Housing Census, Federal Democratic Republic of Ethiopia Population Census Commission, 2008
40. Two Crows Corporation. (2005). Introduction to Data Mining and Knowledge Discovery. 3rd Edition. Two Crows Corporation. 500 Falls Road, Potomac, USA
41. W. J. Frawley, and G. Piatetsky-Shapiro. 1991. Knowledge discovery in databases, AAAI Press and MIT Press, Cambridge, MA.
42. Witten, I. H. & Frank, E. (2005). Data mining: practical machine learning tools and techniques. (2nd ed.). San Francisco: Morgan Kaufmann Publishers.
43. World Health Organization Full report, 2005
44. WHO Assessment of the Pharmaceutical Sector in Ethiopia, 2005

Annex I

Diabetes

- Acarbose — Tablet, 50mg, 100mg
- Biphasic Insulin (BP)* — Injection (highly purified), 100 Units/ml
- Biphasic Isophane Insulin (Soluble/Isophane mixture) Injection, 50/50, 30/70, 100U/ml
- Chlorpropamide — Tablet, 100mg, 250mg
- Glibenclamide — Tablet, 5mg
- Gliclazide — Tablet, 30mg, 40mg, 80mg (sustained release)
- Glimperide — Tablet, 1mg, 2mg, 4mg
- Glipizide — Tablet, 2.5mg, 5mg, 10mg
- Insulin Lispro — Injection, 100u/ml (equivalent to 3.5mg Insulin lispro)
- Insulin lispro/Insulin protamine — (Equivalent to 3.5mg insulin lispro)Injection, 25/75, 50/50 100u/ml
- Insulin Soluble/Neutral (HPB)* — Injection, 100u/ml
- Insulin Zinc Suspension (Insulin Lente) (HPB)* — Injection, 100U/ml
- Isophane/NPH Insulin (HPB)* — Injection, 100u/ml
- Metformin — Tablet, 500mg, 850mg
- Pioglitazone — Tablet, 15mg, 30mg, 45mg
- Rosiglitazone Maleate — Tablet, 1mg, 2mg, 4mg, 8mg
- Tolbutamide — Tablet, 500mg

Cardiac Failure

- Amrinone Lactate — Injection, 5mg/ml in 20ml ampoule
- Captopril — Tablet, 12.5mg, 25mg, 50mg, 100mg
- Digoxin — Elixir, 0.05mg/ml
  Injection, 0.1 mg/ml in 1ml
  ampoule; 0.25mg/ml in 2ml
  ampoule Tablet, 0.25mg
- Enalapril Maleate — Tablet, 2.5mg, 5mg, 10mg, 20mg, 40mg
- Enalapril Maleate + Hydrochlorothiazide Tablet, 10mg +25mg

• Enalaprilat	Injection, 1.25mg/ml
• Hydralazine	Tablet, 10mg, 25mg, 50mg, 100mg
	Powder for injection, 20mg/ml in1ml ampoule
• Hydralazine +Isosorbide	Tablet, 37.5mg +20mg
• Isosorbide	Tablet, 5mg, 10mg, 20mg, 30mg, 40mg, 60mg
	Sublingual Tablet, 2.5mg, 5mg, 10mg
	Tablet/Capsule(s/r), 40mg
• Lisinopril	Tablet, 2.5 mg, 5mg, 10mg, 20mg

Hypertension

• Alfuzosine	Tablet, 2.5mg, 5mg, 10mg
• Amiloride +Hydrochlorothiazide	Oral Solution, 5mg +50mg/5ml
	Tablet, 2.5mg +25mg, 5mg +50mg
• Amlodipine	Tablet, 2.5 mg, 5 mg, 10mg
• Atenolol	Tablet, 50mg, 100mg
• Candesartan	Tablet, 4mg, 8mg, 16mg
• Candesartan +Hydrochlorthiazide	Tablet, 16mg+12.5mg
• Carvedilol	Tablet, 3.125mg, 6.25mg, 12.5mg, 20mg, 25mg
• Captopril	Tablet, 12.5mg, 25mg, 50mg, 100mg
• Captopril +Hydrochlorthiazide	Tablet, 50mg+25mg
• Clonidine	Injection, 0.15mg/ml in 1ml ampoule
	Tablet, 0.025 mg, 0.15mg
• Diazoxide	Injection, 15mg/ml in 20ml ampoule
• Enalapril Maleate	Tablet, 2.5mg, 5mg, 10mg, 20mg, 40mg
• Enalapril Maleate + 5. Candesartan	Tablet, 4mg, 8mg, 16mg
• Candesartan +Hydrochlorthiazide	Tablet, 16mg+12.5mg
• Carvedilol	Tablet, 3.125mg, 6.25mg, 12.5mg, 20mg, 25mg
• Captopril	Tablet, 12.5mg, 25mg,50mg, 100mg
• Captopril +Hydrochlorthiazide	Tablet, 50mg+25mg
• Clonidine	Injection, 0.15mg/ml in 1ml ampoule
	Tablet, 0.025 mg, 0.15mg

• Diazoxide	Injection, 15mg/ml in 20ml ampoule
• Enalapril Maleate	Tablet, 2.5mg, 5mg, 10mg, 20mg, 40mg
• Enalapril Maleate + Hydrochlorothiazide	Tablet, 10mg + 25mg
• Enalaprilat	Injection, 1.25mg/ml
• Felodipine	Tablet, 5mg, 10mg
• Fosinopril	Tablet, 10mg, 20mg
• Hydralazine	Injection, 20mg/ml in 1ml ampoule
	Tablet, 25mg, 50 mg
• Hydrochlorothiazide	Tablet, 25mg
• Hydrochlorothiazide	Tablet, 10mg + 25mg
• Enalaprilat	Injection, 1.25mg/ml
• Felodipine	Tablet, 5mg, 10mg
• Fosinopril	Tablet, 10mg, 20mg
• Hydralazine	Injection, 20mg/ml in 1ml ampoule
	Tablet, 25mg, 50 mg
• Hydrochlorothiazide	Tablet, 25mg
• Irbesartan	Tablet, 75mg, 150mg, 300mg
• Isradipine	Tablet, 2.5mg
• Labetalol Hydrochloride	Tablet, 50 mg, 100 mg, 200 mg,300mg, 400mg
	Injection, 5mg/ml, 1000mg/20ml
• Lisinopril	Tablet, 2.5 mg, 5mg, 10mg, 20mg
• Lisinopril +Hydrochlorthiazide	Tablet, 10mg +12.5mg
• Losartan	Tablet, 25mg, 50mg, 100mg
• Methyldopa	Tablet, 250mg, 500mg
• Metoprolol Tartrate	Injection, 1mg/ml, 5mg/ml
	Tablet, 50mg, 100mg, 200mg (s/r)
• Nicardipine	Injection, 0.1mg/ml
• Nifedipine	Capsule, 5mg, 10mg, 20mg, 30mg (m/r)
	Tablet, 10mg , 20mg, 30mg, 40mg,60mg, 90mg (m/r)
• Phenoxybenzamine HCl	Capsule, 10mg

- Phentolamine Mesilate — Injection, 10mg/ml in 1ml ampoule
- Prazosin Hydrochloride — Tablet, 1mg, 2mg, 5mg
- Propranolol — Injection, 1mg/ml in 1 ml ampoule
  Tablet, 10mg, 40mg
- Ramipril — Capsule, 1.25mg, 2.5mg, 5mg, 10mg

Anti Parkinson

- Amantadine Hydrochloride — Capsule, 100mg
- Benzhexol (Trihexyphenidyl Hydrochloride) — Tablet, 2mg, 5mg
- Bromocriptine — Tablet, 2.5mg, 5mg
- Diphenhydramine Hydrochloride — Tablet, 250mg, 500mg
- Benztropine (Mesylate) — Injection, 1mg/ml in 2ml ampoule
  Tablet, 2mg
- Levodopa — Tablet, 250mg, 500mg
- Levodopa + Benserazide — Capsule, 50mg +12.5mg, 100mg +25mg, 200mg + 50mg
  Tablet, 50mg+12.5mg,100mg+25mg,200mg+ 50mg
- Levodopa + Carbidopa — Tablet, 100mg + 10mg, 250mg + 25mg
- Orphenadrine Hydrochloride — Tablet, 50mg
- Procyclidine — Injection, 5mg/ml in 2ml ampoule

Asthma

- Adrenaline (Epinephrine) — Injection, 0.1% in 1ml ampoule; 1:1000 1mg/ml
- Aminophylline — Injection, 250mg/10ml, in 10ml and 20 ml
  Tablet, 100mg, 200mg
  Tablet (m/r), 100 mg, 225mg, 350mg
- Beclomethasone Dipropinate — Oral Inhalation (aerosol), 50mcg/dose, 100mcg/dose, 200mcg/dose
- Budesonide +FormoterolFumarate — Aerosol, 80mcg +4.5mcg/dose, 160mcg +4.5mcg/dose
- Ephedrine Sulphate — Injection, 50mg/ml
- Ephedrine + Theophylline — Elixir, 6mg + 30mg in each 5ml
  Syrup, 2.24% + 0.30%
  Tablet, 11mg + 120mg

• Fluticasone Furoate	Nasal spray, 27.5 mcg
• Fluticasone +Salmeterol	Aerosol, 100mcg +50mcg /dose, 250mcg +50mcg /dose, 500mcg +50mcg/dose

Annex II

=== Classifier model ===

J48 unpruned tree

STVDateOut = 12/7/2011

| Type = Diabetes

| | ItemUnit = 10ml: Gullele PFSA Hub (16.0)

| | ItemUnit = 100.0: Gullele PFSA Hub (0.0)

| | ItemUnit = 10x10: Adama PFSA Hub (8.0)

| Type = Hypertension: Bahir Dar PFSA Hub (40.0)

| Type = Asthma: Dessie PFSA Hub (8.0)

STVDateOut = 12/15/2011

| ItemUnit = 10ml: Dessie PFSA Hub (16.0)

| ItemUnit = 100.0: Gullele PFSA Hub (8.0)

| ItemUnit = 10x10: Mekele PFSA Hub (8.0)

STVDateOut = 8/25/2011

| Type = Diabetes

| | STVNumberOut <= 116: Mekele PFSA Hub (8.0)

| | STVNumberOut > 116: Gondar PFSA Hub (16.0)

| Type = Asthma: Negelle Borena PFSA Hub (16.0)

STVDateOut = 12/23/2011

| ItemID = ED.400.10: Nekemte PFSA Hub (8.0)

| ItemID = ED.400.12: Dessie PFSA Hub (8.0)

STVDateOut = 12/27/2011: Hawassa PFSA Hub (16.0)

STVDateOut = 1/2/2012

| STVNumberOut <= 1315: Hawassa PFSA Hub (8.0)

| STVNumberOut > 1315: Gullele PFSA Hub (24.0)

STVDateOut = 1/18/2012

| STVNumberOut <= 1459: Adama PFSA Hub (8.0)

| STVNumberOut > 1459: Mekele PFSA Hub (16.0)

STVDateOut = 9/2/2011: Bahir Dar PFSA Hub (16.0)

STVDateOut = 1/30/2012: Dire Dawa PFSA Hub (24.0)

STVDateOut = 2/1/2012

| Type = Diabetes

| | STVNumberOut <= 1628: Gullele PFSA Hub (24.0)

| | STVNumberOut > 1628: Hawassa PFSA Hub (16.0)

| Type = Parkinson: Bahir Dar PFSA Hub (8.0)

| Type = Asthma: Adama PFSA Hub (24.0)

STVDateOut = 9/5/2011

| STVNumberOut <= 11: Hawassa PFSA Hub (8.0)

| STVNumberOut > 11: Dire Dawa PFSA Hub (24.0)

STVDateOut = 2/9/2012: Nekemte PFSA Hub (16.0)

STVDateOut = 2/14/2012

| ItemUnit = 10ml: Gullele PFSA Hub (8.0)

| ItemUnit = 10x10: Negelle Borena PFSA Hub (32.0)

STVDateOut = 9/7/2011

| Type = Diabetes: Gondar PFSA Hub (8.0)

| Type = Asthma: Dessie PFSA Hub (8.0)

STVDateOut = 2/29/2012: Bahir Dar PFSA Hub (8.0)

STVDateOut = 3/1/2012: Hawassa PFSA Hub (16.0)

STVDateOut = 8/12/2011

| STVNumberOut <= 2: Gullele PFSA Hub (8.0)

| STVNumberOut > 2: Jimma PFSA Hub (16.0)

STVDateOut = 9/16/2011: Gullele PFSA Hub (16.0)

STVDateOut = 9/21/2011

| ItemID = CV.600.6: Nekemte PFSA Hub (24.0)

| ItemID = CV.400.1: Gullele PFSA Hub (8.0)

| ItemID = CV.500.29: Gullele PFSA Hub (8.0)

| ItemID = CV.200.10: Nekemte PFSA Hub (8.0)

| ItemID = RE.200.17: Nekemte PFSA Hub (8.0)

| ItemID = RE.200.40: Gondar PFSA Hub (8.0)

STVDateOut = 9/26/2011

| STVNumberOut <= 274: Mekele PFSA Hub (40.0)

| STVNumberOut > 274: Dessie PFSA Hub (8.0)

STVDateOut = 9/29/2011

| STVNumberOut <= 297: Dessie PFSA Hub (32.0)

| STVNumberOut > 297

| | ItemUnit = 10ml: Dessie PFSA Hub (8.0)

| | ItemUnit = 10x10: Adama PFSA Hub (32.0)

STVDateOut = 10/12/2011

| ItemID = ED.400.12: Mekele PFSA Hub (8.0)

| ItemID = ED.400.13: Gullele PFSA Hub (8.0)

STVDateOut = 10/13/2011: Nekemte PFSA Hub (8.0)

STVDateOut = 10/14/2011: Gullele PFSA Hub (32.0)

STVDateOut = 10/18/2011

| STVNumberOut <= 444: Hawassa PFSA Hub (40.0)

| STVNumberOut > 444

| | ItemUnit = 10ml: Nekemte PFSA Hub (16.0)

| | ItemUnit = 10x14: Adama PFSA Hub (8.0)

STVDateOut = 8/19/2011

| STVNumberOut <= 47: Adama PFSA Hub (80.0)

| STVNumberOut > 47

| | STVNumberOut <= 51: Dire Dawa PFSA Hub (16.0)

| | STVNumberOut > 51: Bahir Dar PFSA Hub (8.0)

STVDateOut = 10/24/2011

| ItemID = ED.400.12: Jimma PFSA Hub (8.0)

| ItemID = ED.400.6: Dire Dawa PFSA Hub (8.0)

| ItemID = RE.200.18: Dire Dawa PFSA Hub (8.0)

| ItemID = RE.200.40: Jimma PFSA Hub (8.0)

| ItemID = OP.700.15: Jimma PFSA Hub (8.0)

STVDateOut = 10/25/2011

| Type = Diabetes: Gullele PFSA Hub (8.0)

| Type = Parkinson: Adama PFSA Hub (8.0)

STVDateOut = 8/22/2011

| ItemID = ED.400.10: Gullele PFSA Hub (8.0)

| ItemID = ED.400.12: Hawassa PFSA Hub (8.0)

STVDateOut = 11/10/2011: Gondar PFSA Hub (8.0)

STVDateOut = 11/11/2011: Bahir Dar PFSA Hub (56.0)

STVDateOut = 11/18/2011: Adama PFSA Hub (8.0)

STVDateOut = 11/22/2011

| ItemID = ED.400.10: Gullele PFSA Hub (8.0)

| ItemID = CV.100.12: Gondar PFSA Hub (24.0)

STVDateOut = 11/24/2011

| STVNumberOut <= 869

| | STVNumberOut <= 861: Gondar PFSA Hub (8.0)

| | STVNumberOut > 861: Jimma PFSA Hub (16.0)

| STVNumberOut > 869: Adama PFSA Hub (16.0)

STVDateOut = 11/28/2011

| Type = Diabetes: Negelle Borena PFSA Hub (16.0)

| Type = Asthma

| | STVNumberOut <= 51: Mekele PFSA Hub (16.0)

| | STVNumberOut > 51: Gullele PFSA Hub (8.0)

STVDateOut = 12/2/2011

| STVNumberOut <= 971

| | Type = Diabetes: Dire Dawa PFSA Hub (8.0)

| | Type = Hypertension: Gondar PFSA Hub (8.0)

| STVNumberOut > 971: Mekele PFSA Hub (16.0)

STVDateOut = 1/24/2012

| STVNumberOut <= 1521

| | ItemUnit = 10x10: Hawassa PFSA Hub (32.0)

| | ItemUnit = 100x10: Hawassa PFSA Hub (48.0)

| | ItemUnit = 125ml

| | | ExpiryDate = 3/1/2013: Negelle Borena PFSA Hub (8.0)

| | | ExpiryDate = 10/1/2013: Hawassa PFSA Hub (8.0)

| | ItemUnit = 100ml: Gullele PFSA Hub (8.0)

| STVNumberOut > 1521: Jimma PFSA Hub (24.0)

STVDateOut = 2/21/2012

| Type = Diabetes: Hawassa PFSA Hub (8.0)

| Type = Hypertension: Dessie PFSA Hub (32.0)

STVDateOut = 2/24/2012: Bahir Dar PFSA Hub (16.0)

STVDateOut = 9/8/2011

| TransferedQtyOut <= 12665

| | STVNumberOut <= 193: Nekemte PFSA Hub (16.0)

| | STVNumberOut > 193: Hawassa PFSA Hub (8.0)

| TransferedQtyOut > 12665: Gullele PFSA Hub (56.0)

STVDateOut = 9/19/2011

| STVNumberOut <= 241: Gullele PFSA Hub (24.0)

| STVNumberOut > 241: Gondar PFSA Hub (40.0)

STVDateOut = 8/18/2011

| Type = Diabetes

| | STVNumberOut <= 33: Gondar PFSA Hub (8.0)

| | STVNumberOut > 33: Negelle Borena PFSA Hub (8.0)

| Type = Parkinson: Negelle Borena PFSA Hub (8.0)

STVDateOut = 11/1/2011

| ExpiryDate = 4/1/2015: Negelle Borena PFSA Hub (8.0)

| ExpiryDate = 3/1/2015: Adama PFSA Hub (8.0)

| ExpiryDate = 4/1/2013: Negelle Borena PFSA Hub (48.0)

| ExpiryDate = 6/1/2013: Adama PFSA Hub (8.0)

| ExpiryDate = 5/1/2013: Negelle Borena PFSA Hub (8.0)

| ExpiryDate = 8/1/2011: Adama PFSA Hub (8.0)

STVDateOut = 11/15/2011

| TransferedQtyOut <= 3516: Hawassa PFSA Hub (32.0)

| TransferedQtyOut > 3516: Bahir Dar PFSA Hub (8.0)

STVDateOut = 8/16/2011: Mekele PFSA Hub (56.0)

STVDateOut = 8/17/2011: Gondar PFSA Hub (16.0)

STVDateOut = 1/19/2012

| UnitCost <= 5.28: Mekele PFSA Hub (8.0)

| UnitCost > 5.28: Dessie PFSA Hub (16.0)

STVDateOut = 10/3/2011

| UnitCost <= 12.43: Gullele PFSA Hub (8.0)

| UnitCost > 12.43: Jimma PFSA Hub (40.0)

STVDateOut = 10/5/2011

| Type = Diabetes: Bahir Dar PFSA Hub (8.0)

| Type = Parkinson: Gullele PFSA Hub (16.0)

STVDateOut = 10/17/2011

| STVNumberOut <= 436: Nekemte PFSA Hub (8.0)

| STVNumberOut > 436: Hawassa PFSA Hub (8.0)

STVDateOut = 1/11/2012: Dire Dawa PFSA Hub (8.0)

STVDateOut = 8/24/2011

| Type = Diabetes: Gullele PFSA Hub (8.0)

| Type = Asthma: Dire Dawa PFSA Hub (8.0)

STVDateOut = 1/25/2012

| STVNumberOut <= 1531: Dire Dawa PFSA Hub (32.0)

| STVNumberOut > 1531: Nekemte PFSA Hub (72.0)

STVDateOut = 1/31/2012: Hawassa PFSA Hub (24.0)

STVDateOut = 2/6/2012

| ItemUnit = 10x10

| | UnitCost <= 39.3: Jimma PFSA Hub (80.0)

| | UnitCost > 39.3: Gullele PFSA Hub (16.0)

| ItemUnit = 500.0: Jimma PFSA Hub (8.0)

| ItemUnit = 100x10: Hawassa PFSA Hub (48.0)

| ItemUnit = 3x10: Jimma PFSA Hub (8.0)

STVDateOut = 2/16/2012: Jimma PFSA Hub (24.0)

STVDateOut = 12/12/2011

| Type = Cardiac Failure: Dessie PFSA Hub (8.0)

| Type = Hypertension

| | UnitCost <= 12.43: Gullele PFSA Hub (8.0)

| | UnitCost > 12.43: Hawassa PFSA Hub (32.0)

| Type = Parkinson: Dessie PFSA Hub (16.0)

STVDateOut = 12/16/2011

| TransferedQtyOut <= 4515: Hawassa PFSA Hub (24.0)

| TransferedQtyOut > 4515: Dire Dawa PFSA Hub (8.0)

STVDateOut = 12/26/2011

| Type = Cardiac Failure: Gondar PFSA Hub (8.0)

| Type = Hypertension: Nekemte PFSA Hub (8.0)

STVDateOut = 8/31/2011

| STVNumberOut <= 134: Hawassa PFSA Hub (40.0)

| STVNumberOut > 134: Bahir Dar PFSA Hub (8.0)

STVDateOut = 1/12/2012

| ItemUnit = 10x10: Dessie PFSA Hub (40.0)

| ItemUnit = 5.0: Negelle Borena PFSA Hub (8.0)

STVDateOut = 9/6/2011: Dire Dawa PFSA Hub (88.0)

STVDateOut = 3/5/2012

| Type = Cardiac Failure: Adama PFSA Hub (8.0)

| Type = Hypertension: Adama PFSA Hub (24.0)

| Type = Parkinson: Nekemte PFSA Hub (8.0)

STVDateOut = 9/30/2011: Bahir Dar PFSA Hub (24.0)

STVDateOut = 10/26/2011

| STVNumberOut <= 539: Jimma PFSA Hub (32.0)

| STVNumberOut > 539

| | ItemID = CV.100.17: Bahir Dar PFSA Hub (24.0)

STVDateOut = 12/6/2011: Hawassa PFSA Hub (8.0)

STVDateOut = 2/7/2012

| ItemUnit = 10ml: Dire Dawa PFSA Hub (8.0)

| ItemUnit = 10x10: Mekele PFSA Hub (40.0)

| Type = Cardiac Failure: Dessie PFSA Hub (8.0)

| Type = Hypertension: Dessie PFSA Hub (24.0)

| Type = Parkinson: Mekele PFSA Hub (16.0)

| Type = Asthma: Adama PFSA Hub (8.0)

STVDateOut = 11/2/2011

| ItemID = CV.500.17: Gullele PFSA Hub (24.0)

| ItemID = AL.100.14: Dire Dawa PFSA Hub (8.0)

STVDateOut = 1/23/2012: Bahir Dar PFSA Hub (8.0)

STVDateOut = 12/13/2011: Bahir Dar PFSA Hub (24.0)

STVDateOut = 12/19/2011: Jimma PFSA Hub (32.0)

STVDateOut = 12/21/2011

| TransferedQtyOut <= 2376: Negelle Borena PFSA Hub (72.0)

| TransferedQtyOut > 2376

| | ItemID = RE.200.17: Gullele PFSA Hub (8.0)

| | ItemID = RE.200.20: Nekemte PFSA Hub (8.0)

STVDateOut = 10/4/2011

| STVNumberOut <= 332

| | STVNumberOut <= 18: Adama PFSA Hub (8.0)

| | STVNumberOut > 18: Dire Dawa PFSA Hub (16.0)

| STVNumberOut > 332

| | STVNumberOut <= 339: Mekele PFSA Hub (32.0)

| | STVNumberOut > 339: Negelle Borena PFSA Hub (24.0)

STVDateOut = 11/29/2011

| STVNumberOut <= 911: Adama PFSA Hub (8.0)

| STVNumberOut > 911: Hawassa PFSA Hub (32.0)

STVDateOut = 11/30/2011: Dessie PFSA Hub (24.0)

STVDateOut = 12/20/2011

| STVNumberOut <= 623: Adama PFSA Hub (8.0)

| STVNumberOut > 623: Dessie PFSA Hub (24.0)

STVDateOut = 1/16/2012

| UnitCost <= 35.63: Gullele PFSA Hub (8.0)

| UnitCost > 35.63: Mekele PFSA Hub (16.0)

STVDateOut = 9/27/2011: Gullele PFSA Hub (24.0)

STVDateOut = 8/23/2011: Jimma PFSA Hub (16.0)

STVDateOut = 1/27/2012: Dessie PFSA Hub (16.0)

STVDateOut = 2/15/2012: Gondar PFSA Hub (8.0)

STVDateOut = 2/20/2012: Mekele PFSA Hub (8.0)

STVDateOut = 11/9/2011: Dessie PFSA Hub (8.0)

STVDateOut = 2/13/2012

| Type = Parkinson: Gullele PFSA Hub (8.0)

| Type = Asthma: Adama PFSA Hub (8.0)

STVDateOut = 11/23/2011

| STVNumberOut <= 41: Jimma PFSA Hub (16.0)

| STVNumberOut > 41

| | Type = Parkinson: Gullele PFSA Hub (8.0)

| | Type = Asthma: Nekemte PFSA Hub (16.0)

STVDateOut = 12/5/2011: Bahir Dar PFSA Hub (8.0)

STVDateOut = 2/8/2012: Hawassa PFSA Hub (16.0)

STVDateOut = 12/14/2011: Adama PFSA Hub (8.0)

STVDateOut = 12/22/2011

| STVNumberOut <= 1234: Hawassa PFSA Hub (8.0)

| STVNumberOut > 1234: Gondar PFSA Hub (8.0)

STVDateOut = 8/29/2011: Dessie PFSA Hub (8.0)

STVDateOut = 1/3/2012: Dire Dawa PFSA Hub (8.0)

STVDateOut = 9/13/2011: Jimma PFSA Hub (16.0)

STVDateOut = 9/20/2011: Bahir Dar PFSA Hub (16.0)

STVDateOut = 11/25/2011: Adama PFSA Hub (8.0)

STVDateOut = 12/1/2011: Jimma PFSA Hub (16.0)

STVDateOut = 11/16/2011: Dessie PFSA Hub (8.0)

STVDateOut = 12/22/2004

| Type = Diabetes: Gullele PFSA Hub (6.0)

| Type = Cardiac Failure: Gullele PFSA Hub (12.0)

| Type = Asthma: Bahir Dar PFSA Hub (12.0)

STVDateOut = 3/26/2005: Gullele PFSA Hub (6.0)

STVDateOut = 4/16/2005

| Type = Diabetes: Gullele PFSA Hub (17.0)

| Type = Asthma

| | TransferedQtyOut <= 1683: Dessie PFSA Hub (3.0)

| | TransferedQtyOut > 1683: Dire Dawa PFSA Hub (3.0)

STVDateOut = 4/17/2005

| Type = Diabetes

| | ExpiryDate = 5/13/2013: Gullele PFSA Hub (3.0)

| | ExpiryDate = 9/13/2013: Gullele PFSA Hub (2.0)

| | ExpiryDate = 5/14/2013: Gondar PFSA Hub (3.0)

| | ExpiryDate = 9/14/2013: Gondar PFSA Hub (6.0)

| | ExpiryDate = 12/12/2013: Gullele PFSA Hub (3.0)

| Type = Asthma: Adama PFSA Hub (3.0)

STVDateOut = 4/22/2005

| Type = Diabetes

| Type = Parkinson

| | TransferedQtyOut > 2000: Gullele PFSA Hub (6.0)

| Type = Asthma: Jimma PFSA Hub (6.0)

STVDateOut = 4/25/2005

| Type = Diabetes

| | ExpiryDate = 9/13/2013: Gullele PFSA Hub (6.0)

| Type = Parkinson: Mekele PFSA Hub (9.0)

| Type = Asthma: Mekele PFSA Hub (3.0)

STVDateOut = 5/17/2005: Gullele PFSA Hub (72.0)

STVDateOut = 2/21/2005

| ExpiryDate = 7/14/2013: Gullele PFSA Hub (3.0)

| ExpiryDate = 9/14/2013: Dire Dawa PFSA Hub (6.0)

| ExpiryDate = 12/15/2013: Gullele PFSA Hub (3.0)

| ExpiryDate = 8/14/2013: Gullele PFSA Hub (3.0)

STVDateOut = 5/29/2005: Gullele PFSA Hub (3.0)

STVDateOut = 12/18/2004

| Type = Diabetes: Mekele PFSA Hub (6.0)

| ExpiryDate = 3/14/2013: Dire Dawa PFSA Hub (6.0)

| ExpiryDate = 4/15/2013: Dire Dawa PFSA Hub (3.0)

| ExpiryDate = 11/15/2013: Dire Dawa PFSA Hub (6.0)

| ExpiryDate = 11/12/2013: Dessie PFSA Hub (3.0)

| ExpiryDate = 10/12/2013: Dire Dawa PFSA Hub (24.0)

STVDateOut = 12/30/2004

| TransferedQtyOut <= 815

| | TransferedQtyOut <= 435: Hawassa PFSA Hub (9.0)

| | TransferedQtyOut > 435

| | | Type = Diabetes: Hawassa PFSA Hub (3.0)

| | | Type = Cardiac Failure: Gullele PFSA Hub (6.0)

| TransferedQtyOut > 815: Gondar PFSA Hub (9.0)

STVDateOut = 1/5/2005: Gullele PFSA Hub (24.0)

STVDateOut = 1/30/2005

| | ItemUnit = 10ml

| | | STVNumberOut <= 258: Bahir Dar PFSA Hub (2.0/1.0)

| | | STVNumberOut > 258: Hawassa PFSA Hub (2.0)

| | ItemUnit = 10x10: Hawassa PFSA Hub (6.0/2.0)

| | ItemUnit = 100ml: Bahir Dar PFSA Hub (1.0)

| | ItemUnit = 200doses: Dire Dawa PFSA Hub (1.0)

| ExpiryDate = 9/14/2013

| | TransferedQtyOut <= 970: Dire Dawa PFSA Hub (3.0)

| | TransferedQtyOut > 970: Hawassa PFSA Hub (3.0)

| ExpiryDate = 12/15/2013: Bahir Dar PFSA Hub (3.0)

STVDateOut = 2/8/2005: Negelle Borena PFSA Hub (15.0)

STVDateOut = 2/9/2005: Gullele PFSA Hub (24.0)

STVDateOut = 2/10/2005: Gullele PFSA Hub (6.0)

STVDateOut = 2/14/2005: Jimma PFSA Hub (9.0)

STVDateOut = 2/19/2005: Gondar PFSA Hub (15.0)

STVDateOut = 2/27/2005: Mekele PFSA Hub (27.0)

STVDateOut = 3/10/2005

| ExpiryDate = 7/14/2013

| | ItemID = ED.400.10: Jimma PFSA Hub (1.0)

| | ItemID = ED.400.12: Dire Dawa PFSA Hub (1.0)

| ExpiryDate = 3/14/2013

| | TransferedQtyOut <= 648: Dire Dawa PFSA Hub (3.0)

| | TransferedQtyOut > 648: Bahir Dar PFSA Hub (6.0)

| ExpiryDate = 4/15/2013

| | TransferedQtyOut <= 915: Dire Dawa PFSA Hub (3.0)

| | TransferedQtyOut > 915: Jimma PFSA Hub (3.0)

| ExpiryDate = 12/15/2013: Jimma PFSA Hub (3.0)

STVDateOut = 5/8/2005: Bahir Dar PFSA Hub (3.0)

STVDateOut = 12/17/2004: Gullele PFSA Hub (9.0)

STVDateOut = 12/19/2004: Hawassa PFSA Hub (6.0)

STVDateOut = 4/24/2005: Dessie PFSA Hub (12.0)

STVDateOut = 5/28/2005: Hawassa PFSA Hub (24.0)

STVDateOut = 1/13/2005: Gullele PFSA Hub (20.0)

STVDateOut = 1/14/2005: Nekemte PFSA Hub (6.0)

| Type = Cardiac Failure: Gullele PFSA Hub (12.0)

| Type = Parkinson: Mekele PFSA Hub (6.0)

| Type = Asthma: Gullele PFSA Hub (6.0)

STVDateOut = 5/9/2005: Gullele PFSA Hub (15.0)

STVDateOut = 5/20/2005: Nekemte PFSA Hub (9.0)

STVDateOut = 5/23/2005: Nekemte PFSA Hub (9.0)

STVDateOut = 1/2/2005: Hawassa PFSA Hub (3.0)

STVDateOut = 3/11/2005

| ExpiryDate = 7/14/2013

| | TransferedQtyOut <= 930: Gullele PFSA Hub (3.0)

| | TransferedQtyOut > 930: Adama PFSA Hub (3.0)

| ExpiryDate = 9/13/2013

| | TransferedQtyOut <= 156: Gullele PFSA Hub (2.0)

| | TransferedQtyOut > 156: Adama PFSA Hub (4.0/1.0)

| ExpiryDate = 8/13/2013: Gullele PFSA Hub (9.0)

| ExpiryDate = 2/17/2013: Gullele PFSA Hub (3.0)

STVDateOut = 5/1/2005: Hawassa PFSA Hub (6.0)

STVDateOut = 1/28/2005: Gullele PFSA Hub (15.0)

STVDateOut = 2/3/2005

| Type = Cardiac Failure: Nekemte PFSA Hub (3.0)

| Type = Asthma: Bahir Dar PFSA Hub (9.0)

STVDateOut = 3/3/2005: Negelle Borena PFSA Hub (6.0)

STVDateOut = 4/23/2005

| ExpiryDate = 7/14/2013: Nekemte PFSA Hub (3.0)

| ExpiryDate = 11/17/2013: Dessie PFSA Hub (3.0)

| | ItemID = CV.100.3: Dessie PFSA Hub (2.0/1.0)

| | ItemID = CV.100.9: Dessie PFSA Hub (1.0)

| | ItemID = CV.100.8: Gondar PFSA Hub (1.0)

| | ItemID = AL.100.14: Dessie PFSA Hub (3.0/1.0)

| Type = Parkinson: Adama PFSA Hub (3.0)

| Type = Asthma: Adama PFSA Hub (3.0)

STVDateOut = 3/22/2005: Mekele PFSA Hub (6.0)

STVDateOut = 12/25/2004: Hawassa PFSA Hub (6.0)

STVDateOut = 5/27/2005: Bahir Dar PFSA Hub (3.0)

STVDateOut = 1/16/2005: Gullele PFSA Hub (6.0)