ADDIS ABABA UNIVERSITY

SCHOOL OF GRADUATE STUDIES

SCHOOL OF INFORMATION SCIENCE

AND

SCHOOL OF PUBLIC HEALTH

MINING EMERGENCY MEDICAL DATA: THE CASE OF TIKUR ANBESSA SPECIALIZED HOSPITAL

By

EMAMU ABDELA AWEL

JUNE, 2011

ADDIS ABABA UNIVERSITY

SCHOOL OF GRADUATE STUDIES

SCHOOL OF INFORMATION SCIENCE
AND

SCHOOL OF PUBLIC HEALTH

MINING EMERGENCY MEDICAL DATA: THE CASE OF TIKUR ANBESSA SPECIALIZED HOSPITAL

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
EMAMU ABDELA AWEL

JUNE, 2011

ADDIS ABABA UNIVERSITY

SCHOOL OF GRADUATE STUDIES

SCHOOL OF INFORMATION SCIENCE

AND

SCHOOL OF PUBLIC HEALTH

MINING EMERGENCY MEDICAL DATA: THE CASE OF TIKUR ANBESSA SPECIALIZED HOSPITAL

By

EMAMU ABDELA AWEL

Name and signature of Members of the Examining Board

Name	**Title**	**Signature**	**Date**
Mahder Alemayehu	Chairperson	_____	_____
Prof.Mesganaw Fantahun	Advisor(s),	_____	_____
Workshet Lamenew	Advisor(s),	_____	_____
Dr.Million Meshesha	Examiner,	_____	_____

Declaration

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

_____

Emamu Abdela

June, 2011

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

_____

Workshet Lamenew

_____

Prof. Mesganaw Fantahun

Dedication

I would like to dedicate this paper to my baby Ikram Emamu and my beloved wife Amina Somi, who is such a patient, strong, and supportive woman for me all the way.

ACKNOWLEDGEMENT

I would like to thank my advisors Ato Workshet Lamenew and Prof. Mesganaw Fantahun for their constructive comments and guidance while carrying out this research work.

I would like to acknowledge Dr.Aklilu the Head of Tikur Anbessa Specialized Hospital Emergency Medical Service Unit and Dr. Assefu for their cooperative during data collection.

I would like to thank Yoseph the Emergency Medical Service Unit database administrator of the Hospital for his cooperation during data collection .My warm appreciation goes to the data clerks, especially Papi,Yusuf and the emergency unit nurses for their technical support related to the domain area and the dataset.

I would like to thank Ato Tilahun, for his constrictive comment and encouragement.

My deepest gratitude goes to my family, especially my love Amina Somi for her encouragement, patience, support and love she gave me in doing this research. Special thanks also go to my father Abdela Awel and my sister Semira for their encouragement and support. I am also thankful to my baby Ikram Emamu who is the source of my pleasure but didn't get required care from me due to this research.

My special and deepest gratitude goes to my classmates, specially Habtom, Geletaw, Ellias, Biste Awel and Tadesse for their unreserved help through all the processes of the research and supply of materials used in the research.

TABLE OF CONTENTS

ACKNOWLEDGEMENT VI
LIST OF ABBREVIATIONS X
LIST OF TABLES XI
LIST OF FIGURES XII
ABSTRACT XIII
CHAPTER ONE 1
INTRODUCTION 1
1.1 Background 1
1.2 Statement of the problem 5
1.3 Objectives of the study 7
1.3.1 General objective 7
1.3.2 Specific objectives 7
1.4 Research methodology 8
1.4.1 Business or problem understanding 8
1.4.2 Data collection and understanding 9
1.4.3 Data preparation and preprocessing 9
1.4.4 Model building 10
1.4.5 Evaluation of the models 11
1.4.6 Report generation or Deployment 11
1.5 Scope and limitation of the study 12
1.6 Significance of the study 12
1.7 Ethical consideration 13
1.8 Organization of the Thesis 13
CHAPTER TWO 14
LITERATURE REVIEW 14
2.1 Data Mining and Knowledge Discovery in Database(KDD) 14
2.2 Data mining and Data warehousing 16
2.3 Data mining, Artificial Intelligence (AI) and statistics 18
2.4 The Process of data mining 19
2.4.1 Identifying the target data set (Selection) 19
2.4.2 Preparing the data for analysis (Preprocessing and transformation) 19

2.4.3 Building and testing the model 20
2.4.4 Evaluating the model 21
2.5 Data mining techniques and algorithms 22
2. 5.1 Clustering technique and algorithms 23
2. 5.2 Classification technique and algorithms 24
2.5.2.1 Decision tree 25
2.5.2.2 Rule induction 28
2.5.3 Association rules 29
2.6 Data mining methodologies 32
2.6.1 Knowledge Discovery in Databases (KDD) methodology 32
2.6.2 SEMMA methodology 33
2.6.3 CRISP (CRoss-Industry Standard Process) methodology 34
2.6.4 Comparison among the above data mining methodologies 36
2.7 The Application of data mining technology 37
2.7.1 Application of data mining in Healthcare 38
2.8 Issues and challenges of Health care data mining 39
2.9 Related works 40
CHAPTER THREE 44
BUSINESS UNDERSTANING AND DATA PREPROCESSING 44
3.1 Business understanding 44
3.1.1 Work flow in the Hospital emergency unit 45
3.2 Data Understanding 47
3.2.1 Data collection 47
3.2.2 Data source description 48
3.3 Data Preprocessing 50
3.3.1 Descriptive statistical summary of attributes 51
3.3.2 Data cleaning 58
3.3.3 Handling missing values 58
3.3.4 Handling outlier values 59
3.3.5 Handling noisy values 60
3.3.6 Data transformation and reduction 61
3.3.7 Summary of original and target datasets 62

CHAPTER FOUR....63
EXPERIMENTATION AND ANALYSIS....63
4.1 Model Building....63
4.1.1 Attribute ordering....63
4.1.2 Building classification models....64
4.1.2.1 Validation method selection for decision tree models....64
4.1.2.2 Building pruned decision tree....65
4.1.2.3 Modeling unpruned decision tree....66
4.1.3 Building association rule model....68
4.2. Experimentation and analysis of classification models....69
4.2.1 Experimentation and analysis using decision tree....69
4.2.2 Experimentation and analysis using rule induction....77
4.2.3Analysis and interpretation of classification model rules....78
4.2.4 Discussion on classification models generated rules....80
4.3 Experimentation and analysis of association models....82
4.2.3 Analysis and interpretation of the association rule model results....93
CHAPTER FIVE....96
CONCLUSIONS AND RECOMMENDATIONS....96
5.1 Conclusions....96
5.2 Recommendations....99
References....101
Annex....105

LIST OF ABBREVIATIONS

AAU	Addis Ababa University
AI	Artificial Intelligence
AIDS	Acquired Immuno Deficiency Syndrome
BRHP	Butajira Rural Health Project
CDC	Center for Disease Control and Prevention
CRISP_DM	Cross-Industry Standard Process for Data Mining
HIV	Human Immuno Virus
KDD	Knowledge Discovery in Databases
RDBMS	Relational Database Management System
RIPPER	Repeated Incremental Pruning to Produce Error Reduction
ROC	Receiver Operating Characteristics
RTA	Road Traffic Accident
SAS	Statistical Analysis System
SEMMA	Sample Explore Modify Model and Assess
SPSSS	Statistical Packages for Social Sciences
VCT	Voluntary Counseling and Testing
WEKA	Waikato Environment for Knowledge Analysis
WHO	World Health Organization

LIST OF TABLES

Table2.1: Summary of KDD, SEMMA and CRISP methodologies phases.....36

Table3.1: Description of the attributes.....49

Table 3.2: Statistical summary of sex attribute 51

Table 3.3: Statistical summary of age attribute 52

Table 3.4: Statistical summary of marital status attribute 53

Table 3.5: Statistical summary of referred from attribute 53

Table 3.6: Statistical summary of region attribute 54

Table 3.7: Statistical summary of intent attribute 54

Table 3.8: Statistical summary of cause of visit attribute 55

Table 3.9: Statistical summary of activity attribute 55

Table 3.10: Statistical summary of sub city attribute 56

Table 3.11: Statistical summary of triage assessment attribute 57

Table 3.12: Statistical summary of transferred to attribute 57

Table 3.13: Statistical summary of past medical illness attribute 58

Table 3.14: Discretized result of age attribute.....62

Table 3.15: Dataset summary for the original and target data.....62

Table 4.1: Description of J48 classifier parameter options in weka.....67

Table 4.2: Meanings of the Apriori parameters for association rule.....69

Table 4.3: Confusion matrix.....71

Table 4.4: Performance summary of experiment 1.....73

Table 4.5: Performance summary of experiment 2.....74

Table 4.6: Performance summary of experiment 3.....75

Table 4.7: Performance summary of experiment 4.....76

Table 4.8: Summary all measures of performance for all models.....77

Table 4.9: Association rule mining experiments.....82

LIST OF FIGURES

Fig. 2.1: Data mining processes.....21
Fig. 2.2: Decision tree.....22
Fig. 2.3: KDD process steps.....33
Fig. 2.4: Phases of CRISP-DM.....34
Fig. 4.1: Attribute ordering using information gain.....63
Fig. 4.2: Weka explorer window.....66
Fig. 4.3: Weka generic object editor window for J48 classifier.....67
Fig. 4.4: Weka generic object editor for Apriori parameters setting.....69
Fig. 4.5: ROC Area curve for experiment 3.....75

ABSTRACT

Today's world is encountered a lot of social, economic and political challenges. Among these health challenges take major part, three health related Millennium Development Goals (MDGs) were set and have been implemented to tackle these problems.

There are various health problems. Emergency medical health problem is one of them which is critical and affect many people in the world, especially in developing countries. As statistics shows, in Africa region in general and Ethiopia in particular, the emergency medical case situation is still worse and needs special attention particularly Road Traffic Accident (RTA).
A lot of demographic and clinical (related) data is recorded about patients who come and receive treatment in the emergency medical service unit of the hospital. As these data are getting larger and larger, there can be a probability in which hidden, implicit and non trivial knowledge exist within these data. So far it is recognized among scientific scholars traditional statistical methods might not be good enough to discover such hidden knowledge from large and complex volume of data. This is where data mining becomes very important to mine such hidden, complex, necessary data to generate vital knowledge.

There were various activities carried out throughout the research work based on CRISP (Cross-Industry Standard Process) data mining methodology. The source data for this research purpose was collected from Tikur Anbessa Specialized Hospital Emergency Medicine Registration Database which has 5708 instances. Important patterns and variables related to cause of visit were identified. Data preprocessing activities were made which include handling missing, inconsistent and noisy values that took most of the research time.

Appropriate data mining techniques or functionalities were selected and employed for the research work. These are classification and association rule mining. For classification purpose decision tree classification with J48 algorithm and rule induction with PART algorithm were employed. On the other hand, Apriori algorithm was used for association rule mining purpose.
Weka(Waikato Environment for Knowledge Analysis) 3.6.0 version data mining tool was used for model building and experimentations.

Four experiments for decision tree classification models, one experimental model for rule induction, and ten experiments for association rule mining were done by varying parameters.

Among PART and decision tree classification models experimentation made, pruned decision tree with default confidence factor (0.25) has slightly better performance, accuracy measures and generated rules than the other four models. On the other hand, many association rules with acceptable patterns by domain experts were obtained. Interesting patterns were generated by using classification and association rule discovery models related to the problem domain.
Over all the researcher tried to use techniques in discovering patients cause of visit in the emergency unit.

CHAPTER ONE

INTRODUCTION

1.1 Background

The progress which has been made for emergency and necessary surgical care components are not comparable with that of communicable and vaccine preventable diseases in low and middle income countries. Much emphasis has been given to reduce the load of communicable diseases. Despite, global health policies have emphasized on variety of measures aimed at reducing maternal mortalities and controlling communicable childhood diseases. These measures do not strengthen the health care delivery systems in these countries whose major deficiencies become apparent in the provision of emergency services. To provide and have adequate emergency care, countries require a positive governmental commitment in terms of funding and legislation, there also needs to be community response and education (Razzak & Kellerman, 2002).

In recent years, hospital's emergency service unit has been frustrated with a lot of problems, which include, large number of patients' (overcrowding), delayed patient treatment, long processing time, and high costs. There are a lot of internal and external causes for these problems such as, characteristics of patients', shortage of professionals in emergency unit, access to health care providers, and qualification of man power, patient arrival time, management practices and treatment strategies selected by emergency unit. Understanding these causes properly is the first step to solve them and hence it helps to improve the efficiency of patient care in emergency departments (Fromm et al., 1993).

On the other hand, healthcare practitioners face big problem to turn the huge amount of data which stored during patients' visit into useful information that would enable to make intelligent clinical decisions. In addition, the fast growth of data content, size, and diversity makes researchers to be focused on techniques which enable to find useful information from collections of data (Bellazzi & Zupan, 2008).

Huge amount of emergency data is collected from different emergency medical cases such as injury especially Road Traffic Accident (RTA), sudden collapse, poisoning, suicide, foreign body swallow and the like. Injury specially road traffic injury continue to account for a large number of clients data attending emergency department in Addis Ababa (Tsegaye, Abdella, Ahmed, Tadesse & Bartolomeos ,2010).

There are two big categories of emergency medical cases in Tikur Anbessa Specialzed Hospital which are Trauma (case due to serious physical injury mostly caused by violence or an accident) and Non-Trauma (illness other than trauma case such as respiratory and intestinal). Injuries which are the common case of trauma have been recognized as some of the most life frightening public health problems. They represent 12% of the global burden of diseases and the third cause of overall mortality. According to WHO(2001), estimates 5.06 million people die each year as a result of some form of injuries, comprising almost 9% of all deaths. This is almost equivalent to 14,000 injury deaths every single day (WHO , 2001; Peden, McGee & Sharma ,2002).

In many developing countries, there is an increasing incidence of various injuries in addition to the causative factors change from the historical patterns such as falling from trees to injuries due to occupational hazards, interpersonal violence and road traffic injuries. From these causative factors road traffic injury appear to be the leading cause of traumatic injuries (Nordberg, 1994).

Even though injuries are increasing as the trend suggested by available data, road traffic injuries in particular have not received the attention they deserve in most developing countries. Lack of empirical data and poor quality of the data that exist is probably part of the problem (Museru, 1999).

The situation here in Ethiopia is also similar to other developing countries that is, injuries are common but the attention being given to this problem is below the expected. It constitutes half of all surgical emergencies and is the primary reason for hospital emergency unit visit in Addis Ababa (Larson & Dessie, 1993).

Large quantities of clinical and demographical data about the patients have been accumulated in health care databases, However there are only few tools which can be used to evaluate and analyze this data after it has been captured and stored. Evaluation of stored clinical data might lead to the discovery of trends and patterns hidden within the data that could significantly enhance understanding of disease progression and management (Prather, 2002).

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

The difficulty encountered in human beings to interpret and digest the huge amount of accumulated data and make use of them for decision-making has created a need for development of new tools and techniques for automated and intelligent huge database analysis. As a result, the discipline of knowledge discovery or data mining in databases, which deals with the study of such tools and techniques, has evolved into an important and active area of research (Raghavan, Deogun & Sever, 1998).

Data mining is an automated process employed to analyze patterns in data and extract information .It is an extension of the Relational Database Management System (RDBMS) that helps companies perform highly complicated tasks that are often beyond the capacity of human mind (Trybula, 1997).

Data mining approaches can be applied to clinical databases to assist with decision support. Commonly clinical health care data is massive. It includes basically patient centric data and resource management data. Health care organizations must have ability to analyze data. Treatment records of millions of patients can be stored and computerized and data mining techniques may help in answering several important and critical questions related to health care (Piatetsky-Shapiro, 1991).

As in the other sectors, the health-care sector encounters tremendous economic and competitive pressures which make it to start mining its data. As a result various benefits can be obtained such as improving quality of knowledge to be generated. One way in which data mining is helping health-care providers to cut costs and improve care is by showing which treatments statistically have been most effective. One fruitful area of data mining is medical data mining, here it has been applied to accurate classification and rapid prediction for prognosis and diagnosis of patients in a specialized medical area. In addition it has been also used for training unspecialized doctors to solve a specific diagnostic problem (Masuda, Sakamoto, Yamamoto & Kononenko,2002).

Predictive data mining in clinical medicine enables to derive models that use patient information to support specific clinical decisions. Data mining models can be applied for building decision-making procedures such as diagnosis, and treatment planning, which once evaluated and verified, could then be embedded in clinical information systems. Even though the application of data mining to medical data analysis has been relatively limited until recently, the term data mining has been increasingly used in the medical literature over the past few years. (Bellazzi & Zupan , 2008).

1.2 Statement of the Problem

Hospitals' especially those of the industrialized countries, have employed different types of hospital information systems to manage their healthcare or patient data. These systems typically generate vast amounts of data in the form of number, text, chart, and image. Due to the vast amount, fast growth of data content, size, and diversity, researchers have focused on techniques to find useful information from collections of data (Bellazzi & Zupan , 2008).

Tikur Anbessa Specialized Hospital has great number of patients who visit it with various emergency medical cases such as Road Traffic Accident (RTA), sudden collapse, poisoning, fall down accident ,diarrhea, chest pain, suicide, foreign body swallow and so on. These varieties of cases and the existence of greater amount of emergency case patients from different areas of the country result in an increase in the amount of emergency medical data. Hence these lots of emergency medical patients' data make difficult on finding useful knowledge.

Generally there are a lot of problems and challenges regarding to emergency medical in Ethiopia which is due to many factors such as shortage of health professionals, facilities, resources and discovery of non-trivial knowledge from the collected data .On the other hand, there is a gap or problem of knowledge generation from the massive data which is kept in the emergency unit, and which can result poor handling or planning on emergency medical that is why this research work is initiated to discover such gap regarding on the dataset and to extract knowledge such as the pattern on the data using data mining techniques.

Simple statistical techniques such as /regression technique/ have been used to analyze data in an effort to find correlations, patterns, and dependencies .The analysis made by using such traditional methods focus on problems with much more manageable number of variables and cases than may be encountered in the real world databases. They have limited capacity to discover new and unanticipated patterns and relationships that are hidden in relational databases. That is why data mining is preferable from them, especially due to its ability to explore hidden knowledge from complex databases (Kaur &Wasan, 2006).

Nowadays it is known that the amount of electronic data gathered, the size of population, the speed of disease outbreaks increasing .These make almost impossible to accomplish what the pioneers did, in terms of data analysis using traditional means. This is where data mining becomes useful (Ruben & Canlas , 2009).

Emergency medical service is one of the critical services in the Hospital. So it is important to explore which factors are essential related to patients' cause of visit to the emergency unit. Exploring the collected data through powerful analysis tools has paramount importance. However, such discovery was not done before. The emergency database could be large and complex to be easily analyzed by the traditional statistical methods. Hence, data mining is a solution for exploring such a large and complex database to discover necessary knowledge.

However, as far as the researcher knowledge in Ethiopia is concerned; data mining research especially in emergency medical is almost non-existent but it is the researcher's belief that if data mining was applied in this area, it is significant for generating valuable knowledge for health care administrators, policy makers and health professionals. As a result; the researcher is motivated to see the applicability of data mining techniques on emergency data the case of Tikur Anbessa Specialized Hospital. This Hospital was selected for the research purpose because it deals with large volume of patient data and it handles emergency cases from all over the country. Hence, it is a good representative and it has also emergency medicine registration database system which store patients' data.

The Hospital emergency medical registration database system currently, used for registering and storing emergency medical patients' data. Hence analysis on these data was not made previously. Moreover, there were also gaps identified such as being unable to discover knowledge from the data related to patients cause of visits in the emergency unit(emergency medical service unit) which brings problem of decision making ,planning and awareness. Therefore it is the researcher's belief that if data mining techniques are applied on these patients' data, there can be patterns or relationships among attributes of the database which can help the Hospital to get knowledge and improve the scenario.

1.3 Objectives of the Study

1.3.1 General objective

The general objective of the research is to explore the application of data mining techniques on emergency medical data to discover patients' cause of visit in the emergency medical unit of Tikur Anbessa Specialized Hospital that could support to identify significant knowledge in the emergency unit

1.3.2 Specific objectives

The specific objectives of this study are:

- To collect and extract the dataset required for the problem domain.
- To preprocess the raw data into a suitable dataset for model building.
- To build data mining models on the preprocessed dataset.
- To come up with the selected models to be implemented which are significant for the scenario under investigation.
- To analyze and interpret the results of the models.
- To extract interesting patterns which are relevant to the problem domain.
- Report results and make recommendations.

1.4 Research Methodology

In this research, CRISP_DM (Cross-Industry Standard Process for Data Mining) methodology was selected among other methodologies (such as SEMMA and KDD). Because it is considered to be the most popular, extremely complete, documented, and presented in many of data mining research publications. Furthermore all CRISP phases are duly organized, structured and defined, allowing that a research could be easily understood or revised (Azevedo & Santos, 2008).

The brief comparison of these methodologies is found in chapter two under section 2.5.4.
According to Chapman et al.(2000) CRISP has six phases, they are business understanding, data understanding, data preparation and pre-processing, model building, evaluation and deployment. The next section briefly discuses these phases based on Chapman et al. (2000).

1.4.1 Business or problem understanding

It is known that business or problem understanding is one of the essential activities in research work, which can support to understand the research area or problem domain. According to Chapman et al.(2000) this phase focus on understanding the research objectives, requirements and goals.
In this phase the researcher tried to understand the current business problem using different techniques such as reviewing documents, observation and discussion with domain experts so that the existing work flow, research area and underlying processes that can be mapped in to data mining process were notified. As a result, the researcher attempted to formulate the research problem in such a way it would be handled by data mining techniques. Objectives of the research were specified as well.

In general extensive literature review was conducted concerning on data mining technology that focuses on mapping business problem to be handled by data mining techniques in health care. Numerous books, journals, documents, and articles were revised to understand the research area; data mining technology in general and its application in health care in particular.
The detail of business and data understanding is available in chapter three under section 3.1 and 3.2 respectively.

1.4.2 Data collection and understanding

This phase starts with an initial data collection and proceeds with activities in order to be familiar with the data, and to discover first insight into the data. In this phase the researcher tried to understand the collected data through interviewing domain experts, observation of current existing system and referring documents related to the data.

The source data for the research was collected from emergency medical unit of the Hospital,
A total of 5,524 patients' records were taken from the emergency database for the research.

Based on the discussion, observation of the database and the researcher's view, four important files or tables were identified and used for the research purpose. They were: Personal information or Socio-demographic, Event, Trauma triage and Non-trauma triage tables. Based on the appropriateness to the problem domain and the discussion with domain experts the following fourteen attributes were selected from all four tables. They are medical card number, sex, age ,marital status ,sub city , region, cause of visit, referred from ,intent, triage assessment , activity, place of injury ,past medical illness and transferred to**.** More description for all these attributes is found in chapter three, under data source description (section 3.2.2).

1.4.3 Data preparation and preprocessing

The data preparation and preprocessing phases cover all activities to construct the final dataset (data that will be fed into the modeling tool(s)) from the initial raw data. Data preparation tasks are likely to be performed iteratively.

A data quality assessment identifies characteristics of the data that will affect the model quality. Basically, we are trying to ensure not only the correctness and consistency of values but also that all the data you have is measuring the same thing in the same way. There are different types of data quality problems such as, fields may have an incorrect value, the value for a field might be missing and inconsistencies might be identified specially when consolidating data from multiple sources(Two Crows Corporation ,1999).

In this research efforts to identify and reduce data quality problems were done. Cross checking the exported excel data files with their corresponding original data files were done before directly using the exported source data as they were. Data preprocessing activities were carried out which generate appropriate data for the research that improves the quality of data and the knowledge out of it. Tasks which were performed during data preparation and preprocessing include description of data sources, performing statistical summary measures, data transformation and reduction; handling missing, outliers, and noisy data values. The detail of each task is found in chapter three section 3.3.

The following tools were used for data preparation and preprocessing:
MS- Excel was used for data preparation, pre-processing and analysis tasks.
SPSS was used for generating statistical summary such as calculating mean, mode and median.
Weka data mining tool was used for preprocessing (such as to discretize age variable and convert numeric variable to nominal). They are selected due the researcher prior knowledge on them.

1.4.4 Model building

There are different data mining models which can be applicable for different problem domain. Hence, for this research appropriate data mining models were selected and implemented. They are classification and association rule discovery models .

Decision tree classifier (J48) and rule induction classifier (PART) were used for classification purpose. Decision tree classifier (J48) was selected due to its visualization power of the tree structure, and its simplicity to understand. Rule induction classifier (PART) was selected due to its capability of generating relatively large and understandable rules. The reason that these two algorithms were employed is to investigate the problem domain with variety of rules, options and performance which can add interesting knowledge to the area.
On the other hand, Apriori algorithm was used for association purpose as it can discover hidden patterns in addition to classification rules and it is also the most common.

Various models were carried out by changing the default parameters of the algorithms, from these four experimental models were selected and built for classification purpose using J48 and PART model also used for classification purpose. Ten, experimental models were built for association purpose. Weka was used for building models, evaluation and analysis of the models.

1.4.5 Evaluation of the models

Before proceeding to final deployment of the model, it is important to more thoroughly evaluate the model and review the steps executed to construct the model to be certain it properly achieves the business objectives.

In this phase, the researcher tried to evaluate J48 decision tree and PART model performance using various accuracy measure such as accuracy performance (correctly and incorrectly classified instance) ,confusion matrix analysis ,True Positive Rate, False Positive Rate, precision, recall, execution time , ROC area and discussion on the generated rules or models with domain experts.

Evaluation of Association rules were also made in terms of the meaning of patterns, minimum support and confidence thresholds. The minimum support were varied from 0.1 to 0.3 and confidence thresholds varied from 50% to 100%.

1.4.6 Report generation or Deployment

It is the last phase of CRISP methodology. Creation of the model is generally not the end of the study. Even if the purpose of the model is to increase knowledge of the data, the knowledge gained will need to be organized and presented in a way that the customer can use it.

In this phase, the researcher tried to prepare the thesis report and present it to Addis Ababa University School of Information Science and Public Health, Health Informatics Program.

Efforts will be made to publish it on the university website as reference material, especially for those research works related to the area.

Efforts will be made to publish it in reputable journals so as to make it accessible at large.

1.5 Scope and Limitation of the Study

The scope of the research is limited to explore the applicability of classification and association rule data mining techniques (using J48,PART and Apriori classifiers) for discovering patients' cause of visit scenario only in the emergency medical dataset for the selected tables and variables of Tikur Anbessa Specialized Hospital.

The limitations of the research include unable to explore the data set for the remaining files, variables and scenarios using the selected or other unused algorithms due to various reasons. Includes, shortage of time, delay on approval of the proposal and ethical clearance, too much delay to get emergency medical source data, lack of standardization during data entry and collection, existence of conflicting, duplicate, noisy and missed data. In general quality of the source data was among the constraints.

1.6 Significance of the Study

The research was targeted to explore the benefit of data mining technology for exploring significant knowledge from the emergency medical dataset which can be used to generate new or unseen knowledge that can improve the current status of the emergency medical health care by building applicable model and identifying dominant factors related to the cause of visit.
The result of the study could be used to investigate patterns that can be observed from the emergency medical data that can be used as input for health care providers, decision makers, policy makers and planners.
Moreover, the output of the study might be used as a benchmark for other researchers who have interest in such related issue or area.

1.7 Ethical Consideration

The research did not require personal identifiers (like ID, name) of the person about whom the data is collected and therefore, the problem of confidentiality was unquestionable and furthermore, the research did not target particular individual.

Ethical clearance was also obtained from the school of public health.

The research is fully for the purpose of academic.

The research did not expose or harm anybody by any means.

Objective of the study was clarified for the emergency unit before data collection was made.

1.8 Organization of the Thesis

The thesis is divided into five chapters. The first chapter is an introductory part, which contains basically background information, statement of the problem, and objective of the research, significance of the study, methodology used, scope and limitation.

The second chapter deals with literature review related with data mining technology, KDD, data mining phases, data mining methodologies, data mining application in health care, data mining techniques and algorithms(which includes classification, association rule discovery, decision tree and Apriori).

The third chapter deals with business understanding and data preprocessing which includes the work flow, tables and variables understanding, descriptive summary of variables and data preprocessing.

The fourth chapter deals with experimentation and analysis using different scenarios for both classification and association rule discovery.

The five chapter deals with conclusions and recommendations.

CHAPTER TWO

LITERATURE REVIEW

2.1 Data Mining and Knowledge Discovery in Database (KDD)

Recently it is hard to find examples of anything, anywhere, that has changed as fast as the quantity of stored information. Furthermore, information explosion has created new opportunities and new headaches in every field, ranging from marketing to medicine to manufacturing (Berry & Linoff, 1997).

It is surprising that, the amount of information in the world is estimated to double every 20 month as studies illustrate. As a result large number of scientific, government and corporate information systems are being overwhelmed by a flood of data that are produced and stored routinely, developing into large databases amounting to giga (and even tera) bytes of data (Fayyad, Piatetsky-shapiro & Smyth, 1996). These databases comprise potentially gold mine of valuable information, but it is beyond human capability to analyze massive amounts of data and draw meaningful patterns form such databases. In addition, the explosive growth in data and databases has resulted in an acute need for novel techniques and tools that can intelligently and automatically transform the processed data into useful information and knowledge (Deogun & Jitender ,2001).

As a result, a new generation of computerized method is emerging in recent years to help the attempt of interrogating and analyzing very large data sets automatically and efficiently, thereby extracting knowledge which is useful in decision making. This method is referred to as data mining (Levin & Zahavi, 1999). To understand the term data mining, it is useful to look at the literal translation of the word: to mine in English means to extract. The association of this word with data suggests an in-depth search to find additional information which previously went unnoticed in huge data available. Data mining is the process of selection, exploration, and modeling of large quantities of data to discover regularities or relations that are at first unknown with the aim of obtaining clear and useful results or knowledge (Guidicini**,** 2003).

There are many definitions of data mining that could be found in different literatures with variations in the name as well that carry a similar or slightly different meanings. Therefore, it is important to get familiar to those variations to have the overall picture of data mining.
According to Fayyad et al. (1996), the variation is explained as "Finding useful patterns or meaning in raw data has variously been called KDD (Knowledge Discovery in Databases), data mining, knowledge discovery, knowledge extraction, information discovery, information harvesting, data archaeology and data pattern processing". They define Data Mining as a "non-trivial process of identifying valid, novel, potentially useful and ultimately understandable patterns in data" .For this particular research their definition is used throughout this paper.

Data mining is essentially an interdisciplinary field bringing together techniques from machine learning, pattern recognition, statistics, databases, and visualization to address the issue of exploring large and complicated databases to identify "interesting" relationships, e.g., high order interactions, or very non-linear relationships that ordinarily would not be detected by standard statistical analyses (Cabena et al., 1998 ; Borok, 1997).

Data mining usually makes sense when there is large amount of data. For this reason most of the algorithms developed for data mining purpose requires large volume of data so as to build and train models that are responsible for different tasks of data mining such as classification, clustering, prediction, association and the like (Berry & Linoff ,1997).
The major reason that data mining has attracted a great deal of attention in the information industry in recent years is due to the wide availability of huge amounts of data and the imminent need for turning such data into useful information and knowledge(Han & Kamber ,2006)

Knowledge Discovery in Database (KDD) refers to the overall process of discovering useful knowledge from data and it consisting of several distinct steps .The focuses of KDD include how the data are stored and accessed; how algorithms can be scaled to massive data sets and still run efficiently; how results can be interpreted and visualized; and how the overall man-machine interaction can usefully be modeled and supported (Fayyad, Piatetsky-Shapiro& Smyth, 1996).

There still prevails misunderstanding about the terms KDD and data mining. More often, these terms have been used interchangeably. However, the two terms have distinct meanings.
The term KDD refers to the whole process of changing low level data into high level knowledge. KDD can simply be defined as the non-trivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data. On the other hand, data mining is defined as a single core step in the process of knowledge discovery that involves the application of appropriate algorithm so as to discover meaningful trends from the database under investigation (Frawley & Piatetsky-Shapiro, 1996).

Nevertheless, many authors believe that the term data mining has become more popular in industries, in media and in database researches as a synonym for knowledge discovery in database and hence the two terms are used interchangeably. As a result, in this research the term data mining and knowledge discovery in database used interchangeably (Carbone, 1997; Han & Kamber, 2001).

2.2 Data mining and Data warehousing

As one good feature data mining can have, and will continue to take place in environments not supporting a data warehouse. However, as volumes of data continue to be collected for purposes of decision support, the need for integrating the disparate systems ,the need for efficient data storage and retrieval architectures has become quite apparent .The result of this need has sparked the birth of the data warehouse.

The construction of data warehouses can be viewed as an important preprocessing step for data mining. Moreover, data warehouses provide on-line analytical processing (OLAP) tools for the interactive analysis of multidimensional data of varied granularities, which facilitate effective data mining. Hence, the data warehouse has become an increasingly important platform for data analysis and on-line analytical processing and will provide an effective platform for data mining (Han & Kamber,2001; Berry & Linoff, 1997). According to Immon (1996), a data warehouse is ‘an integrated collection of data about a collection of subjects (units), which is not volatile in time and can support decisions taken by the management.

Data warehousing is a popular business trend of collecting and cleaning transactional data to make them available for online analysis and decision support (Fayyad, Piatetsky-Shapiro & Smyth, 1996).

Most of the time, the data to be mined is first extracted from an enterprise data warehouse into a data mining database. There is some real benefit if the data mining database is already part of a data warehouse. A data mining endeavor includes the effort to identify, acquire, and cleanse data. The problems of cleansing data for a data warehouse and for data mining are very similar. If the data has already been cleansed for a data warehouse, then it most likely will not need further cleaning in order to be mined .Furthermore, if the design of the data warehouse includes support for data mining applications, the warehouse facilitates and catalyzes the data mining efforts.
The two technologies work together to deliver value. Data mining fulfills much of the promise of data warehousing by converting an essentially inert source of data in to actionable information (Two Crows Corporation, 1999; Berry & Linoff, 1997).

The use of data warehouse will have addressed many of the problems in data consolidation. But data warehouse is not the prerequisite for data mining and data analysis. Setting up a large data warehouse consolidates data from multiple sources, resolves data integrity problems, and loads the data into a query database. However, putting such a large database up can be an enormous task, sometimes taking years and costing millions of dollars. Alternatively the data miner could mine data from one or more operational or transactional databases by simply extracting it into a read-only database. This new database functions as a type of data mining database (Two Crows Corporation, 1999).

Lack of complete and well organized information is one of the obstacles to achieving efficient data mining. Very often a database is created for reasons that have nothing to do with data mining, so the important information may be missing. Incorrect data is another problem. As one possible solution, creation of a data warehouse can eliminate many of these problems. Efficient organization of the data in a data warehouse coupled with efficient and scalable data mining allows the data to be used correctly and efficiently to support company decisions (Guidicini, 2003).

2.3 Data Mining, Artificial Intelligence (AI) and Statistics

In the past two centuries traditional statistics has developed to help scientists, engineers, and later business analysts to make sense of the data they have collected. The history of data mining and data mining techniques are generally rather different, highlighting the influence of other disciplines. Data mining takes advantage of advances in the fields of AI and statistics. Both disciplines have been working on problems of pattern recognition and classification, and have made great contributions to the understanding and application of neural nets and decision trees (Two Crows Corporation, 1999; Berry & Linoff, 1997).

Statistics is very useful in providing a language and framework for quantifying the uncertainty, which results when one tries to infer general patterns from a particular sample of an overall population. However, it does not solve all data mining problems. Moreover, the computational complexity of statistical approaches does not grow well with larger data sets. On the other hand, data mining does not replace traditional statistical techniques. Rather, it is an extension of statistical methods that is in part the result of a major change in the statistics community.
The development of most statistical techniques was, until recently, based on elegant theory and analytical methods that worked quite well on the modest amounts of data being analyzed. The increased power of computers and their lower cost, coupled with the need to analyze enormous data sets with millions of rows, have allowed the development of new techniques with a powerful exploration of possible solutions (Two Crows Corporation, 1999; Edelstein, 1998; Berry & Linoff, 1997; Fayyad et al., 1996).

Furthermore, data mining is the application of AI and statistical techniques to common business problems. Data mining is, therefore, a tool for increasing the productivity of people trying to build predictive models by making AI and statistical techniques available to the skilled knowledge workers as well as the trained professionals (Berry & Lineoff, 1997).
Statistics has always been about creating methods to analyze data. The main difference between statistical methods and machine learning methods is that statistical methods are usually developed in relation to the data being analyzed but also according to a conceptual reference paradigm (Guidicini, 2003).

Statisticians have recently shown an interest in data mining and this could help its development. Statistical analysis traditionally concerns itself with analyzing primary data that has been collected to check specific research hypotheses, data mining can also concern itself with secondary data collected for other reasons. Furthermore, statistical data can be experimental data, but in data mining the data is typically observational data (Guidicini, 2003).

2.4 The Process of Data Mining

Data mining research undergoes through varies steps as depicted in figure 2.1.The following are the most commonly employed steps in the process of data mining; Identifying source of data, preparing data for analysis, building and training a computer model, and evaluating the computer model. The next section outlined broadly the basic steps in the process of data mining (Berry & Linoff, 1997).

2.4.1 Identifying the target data set (Selection)

In general as in many problem solving researches, the data mining research starts with a clear definition of the business problem involved and the objective function. Once the data mining problem is identified, the next step is to identify the target data set in which the solution of the problem can be found. To accomplish a successful knowledge discovery, reliable and consistent data is a prerequisite (Han & Kamber, 2001).

2.4.2 Preparing the data for analysis (Preprocessing and Transformation)

Real world databases usually contain incomplete, noisy and inconsistent data such data may cause confusion for the knowledge discovery process. Hence, data cleaning is mandatory to improve the quality of data and so as to improve the accuracy and efficiency of the knowledge discovery process. Preprocessing is a routine task that usually consumes much of the efforts exerted in the entire data mining process. Generally speaking, preparing the data is a step where much time is devoted, which can take up to 80% of the total project effort (Han & Kamber, 2001; Saarenvitra, 2001).

To start with a data mining problem, it is important to bring all the data together into a set of instances with their corresponding attributes. Specially, in the case of relational databases (one type of data modeling technique in database management where records are placed in a tabular form), it is necessary to denormalize (merge related tables) the relations (tables) so that one can describe any relationship among attributes of a given instance. There are various activities to be done in this step such as data sets containing missing values, inaccurate values; duplicate, null, noisy and inconsistent data have to be preprocessed before any data mining tool is applied. For instance, missing values are frequently indicated by out of entering values like negative one in a numeric field that are normally positive (Witten & Frank, 2000).

On the other hand, there are different approaches to handle missing values such as filling mean values, using the most probable value, filling manually, or ignoring the record containing missing values. In summary, real-world data tend to be dirty, incomplete, and inconsistent. Data preprocessing techniques can improve the quality of the data, thereby helping to improve the accuracy and efficiency of the subsequent mining process. The data mining modeling is the next step once we get the preprocessed dataset. It is only at this point that one invokes data mining models and tools to interrogate the data and convert it into knowledge for decision making. This model building step involves selecting data mining tools, transforming the data if the tool requires it, generating samples for training and testing the model, and finally using the tools to build and select a model. The resulting models might have important patterns to be analyzed and interpreted to be used as decision support knowledge (Han & Kamber, 2006).

2.4.3 Building and testing the model

Before building and training a model, appropriate data mining technique has to be selected. The problem of overfitting is another issue that deserves a due consideration at this step. According to Mitchell (1997), overfitting is an attempt to create overly complex data mining model that fits noise in the training data or unrepresentative features of the particular training data that decreases the generalization accuracy of the model over other unseen instances.

In practice, most data mining models have a tendency to suffer from overfitting and the solution for this problem is to provide the model with the test dataset. In fact, the model that yields a promising result for the training data set will at first come up with disappointing results on the test data. So the next step in the process is to refine the model to produce a second model that can work as well on the test set as it does on the training set (Thearling, 2003).

2.4.4 Evaluating the model

In order to evaluate the performance of the model for new data, there is a need to examine the error rate on the data set that did not take part in the process of model formulation, test set (Thearling, 2003). Then a model with a high success rate or low error rate is considered as a good model. This holds true assuming that both the training and test set data are representatives of the underlying population (Witten & Frank, 2000). Figure 2.1 illustrates data mining Processes.

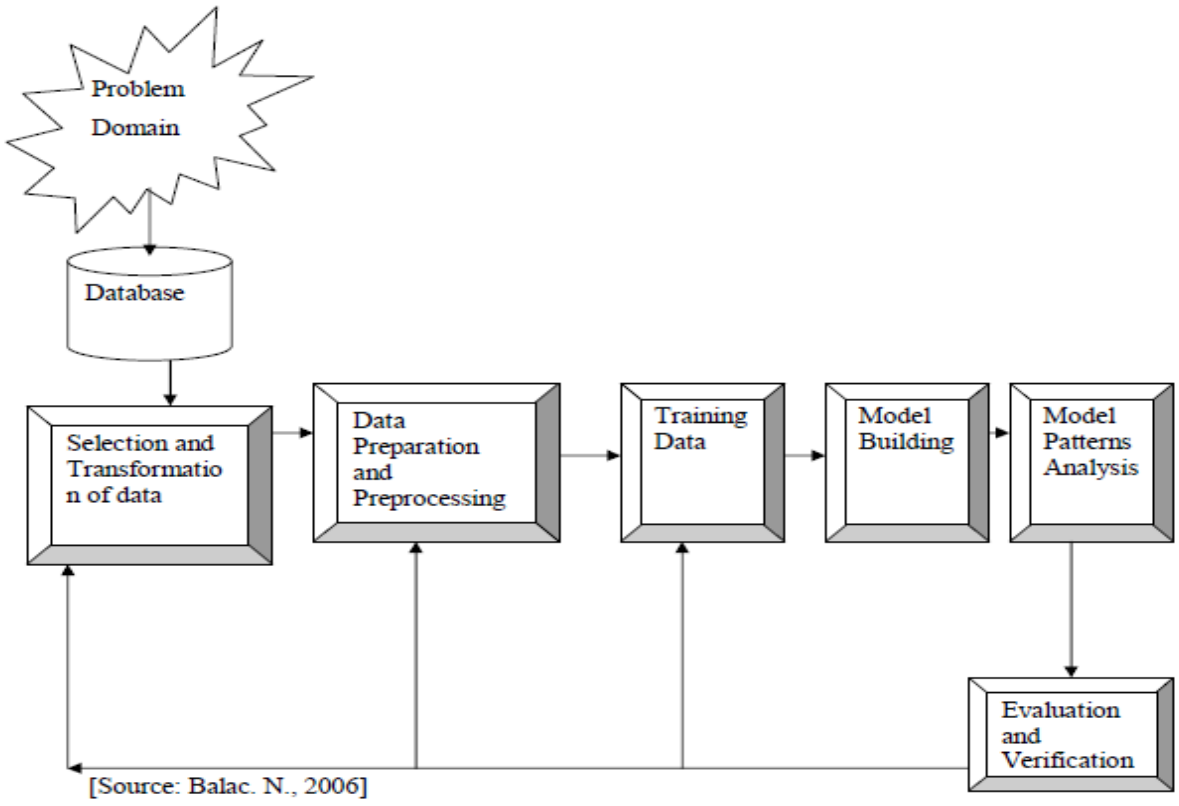


Fig. 2.1: Data Mining Processes

2.5 Data Mining Techniques and Algorithms

The goals of data mining can be varied based on the intended use of the system. There are two high-level primary goals of data mining, they are prediction and description. Prediction involves finding the patterns to predict the future characteristics of some entities that the user interests using variables or fields in the dataset. In the other words predictive mining tasks perform inference on the current data for the seek of predicting the target that the user interested on it. Descriptive mining involves on presenting patterns that describe the data in the form that could be understood by the user. The relative importance of prediction and description for particular data mining applications can vary considerably. In the context of KDD, description tends to be more important than prediction (Fayyad, Piatesky-Shapiro, & Symth, 1996).

There are two terms that are used in predictive models, dependent and independent variable.
The values or classes that are going to be predicted are known as the response, dependent or target variables and values used to make prediction are said to be the predictor or independent variables. Predictive models are built, or trained, using data for which the value of the response variable is already known. This kind of training is sometimes referred to as supervised learning or directed data mining, the reason is that calculated or estimated values are compared with the known results. On the other hand, descriptive techniques are sometimes referred to as unsupervised learning or undirected data mining because there is no already known result that can be used to guide algorithms or it is a form of learning by observation, rather than learning by examples (Two Crows Corporation, 1999).

Descriptive data mining differs from predictive data mining, which analyzes data in order to construct one or a set of models and attempts to predict the behavior of new data sets. Predictive data mining includes classification, regression analysis, and trend analysis. Clustering algorithms, pattern recognition models, visualization methods, and link analysis are the major members of descriptive models(Han & Kamber, 2006; Levin & Zehavi ,1999).

2. 5.1 Clustering technique and algorithms

Unlike with that of classification here in clustering the class label of each object is not known when we start or by which attributes the data will be clustered .

Clustering is the process of grouping the data into classes or clusters, so that objects within a cluster have high similarity in comparison to one another, however are very dissimilar to objects in other clusters. This is similar to assigning animals and plants into families where the members are alike. Clustering is used to place data elements into related groups without advance knowledge of the group definitions. There are many areas in which clustering can be found such as data mining, statistics, biology, and machine learning .Furthermore, the goal of clustering is to find groups in which their members are similar to each other but are very different for other clusters. Each cluster that is formed can be viewed as a class of objects from which rules can be derived (Two Crows Corporation, 1999; Han & Kamber, 2006).

In general, the class labels are not present in the training data simply because they are not known to begin with. As a result, experts' knowledge on that domain is required to interpret the clusters and it is up to the data miner to determine what meaning, if any, to attach to the resulting clusters. Once the clusters are found that reasonably segment the database, then these clusters may then be used to classify new data (Berry & Linoff, 1997).

Clustering algorithms can be classified in two general categories: Hierarchical and Non-hierarchical. Hierarchical procedures involve the construction of a hierarchy or tree like structure in a bottom-up and top-down approach. They have the advantage of being fast and take less computing time. But they could be misleading and unreliable because undesirable early combinations may persist throughout the analysis and lead to artificial results. On the other hand non-hierarchical clustering procedures do not involve the tree-like construction process. Instead, here the first step is to select a cluster center or seed, and all objects (data points) within a pre-specified threshold distance are included in the resulting cluster. The most well known non-hierarchical clustering algorithm is the K-Means algorithm (Bounsaythip et al, 2001).

2. 5.2 Classification technique and algorithms

Classification is one of the most common data mining task which enable us to understand the world since we are commonly classifying, grouping and ranking things around us. It explores the features of a newly presented object and assigning it to one of a predefined set of classes. For any object or instance, classes are predefined according to the value of a specific field (Berry & Linoff, 1997).

It used to extract models or functions for describing important data classes or to predict future data trends whose class label is unknown previously. Classification models are created based on the analysis of already classified data or a set of training data (i.e., data objects whose class label is known) and can be represented in various forms, such as If-Then rules, decision trees, and neural networks. It has numerous applications, including fraud detection, target marketing, manufacturing, and medical diagnosis (Han & Kamber, 2006; Two crow corporation, 1999).

There are two steps in classification process: Learning (training) and classification. In the training step, a classifier is built describing a predetermined set of data classes , this step is also known as supervised learning (i.e., it is told to which class each training records belongs). In the second step the model is used for classification. The dataset is split into training and testing data or it will follow iterative process through k-fold cross validation to evaluate accuracy of the classifier. The accuracy of a classifier on a given test set is the percentage of test set records that are correctly classified by the classifier .The class label of each test records is compared with the learned classifier's class prediction of the records (Han & Kamber, 2006).

The common difference between classification and clustering is in the case of clustering unlike classification there are no predefined classes in which the data is categorized. As there are no predefined classes and examples, in clustering records are grouped together on the basis of similarity among the instances (Beaza-yates & Ribeiro-Neto, 2000). The most widely used techniques for classification are decision trees and neural networks (Berry &Linoff, 1997).

2.5.2.1 Decision tree

Decision tree classifiers or algorithms such as J48 and neural networks are the common algorithms that are used in classification tasks.

A decision tree is a flowchart-like tree structure that contains different components such as internal node (decision node), branch and leaf node. Where branch shows an outcome of the test, internal or non leaf node denotes a test on an attribute, and leaf node (or terminal node) holds a class label. The topmost node in a tree is called the root node (Han & Kamber, 2006).

According to Bramer (2007), decision tree structures are a common ways to organize classification schemes. In classifying tasks, decision trees visualize what steps are taken to arrive at a classification. Every decision tree begins with what is termed as a root node which is considered to be the parent of every other node. Each node in the tree evaluates an attribute in the data and determines which path it should follow. Typically, the decision test is based on comparing a value against some constant.

A decision tree grows from the root node. At each node the data is split to form new branches, until reaching a node that cannot be split any more (leaf node).Traversing the tree from the best leaf node to the root provides the rule that classifies the target variable. All records that arrive at a given leaf of a tree are classified the same way. Trees could grow in binary trees of non-uniform depth, means each node has two children and the distance of a leaf to the root varies.

Figure 2.2, represents sample decision tree.

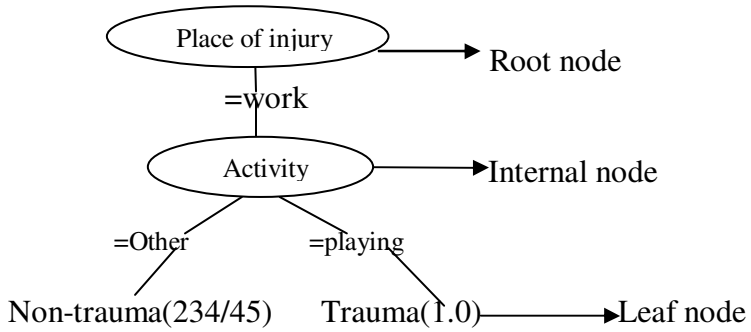


Figure 2.2 Decision tree

If the attribute that is tested at a node is a nominal one like in figure 2.2, the number of children is usually the number of possible values of the attribute. Because there is one branch for each possible value. If the attribute is numeric, the test at a node usually determines whether its value is greater or less than a predetermined constant, giving a two-way split. Alternatively, a three-way split (split into less than, equal to, and greater than) may be used (Witten & Frank, 2000).

There are various reasons where decision tree classifiers are so popular. To mention some; the construction of decision tree classifiers do not require domain knowledge or parameter setting; it can handle high dimensional data, their representation of acquired knowledge in tree form is intuitive and generally easy to assimilate by humans; the learning and classification steps of decision tree induction are simple and fast. In general, decision tree classifiers have good accuracy and offered graphical user interface (Han & Kamber, 2001).

Decision tree induction algorithms such as Classification and Regression Trees (CART), C4.5 algorithm operate recursively. First, an attribute must be selected as the root node. In order to create the most efficient (i.e., smallest) tree, the root node must effectively split the data. Each split attempts to pare or trim down a set of instances (the actual data) until they all have the same classification (Bramer, 2007).

The best split is the one that provides the most information gain. Information in this context comes from the concept of entropy(An information-theoretic measure of the ‘uncertainty’ of a training set, due to the presence of more than one classification), as developed by Claude Shannon. Although "information" has many contexts, it has specific mathematical meaning relating to certainty in decision making. Ideally, each split in the decision tree should bring us closer to a classification. One way to conceptualize this is to see each step along the tree as removing randomness or entropy (Han & Kamber, 2006). According to Bramer(2007) the process of decision tree generation by repeatedly splitting on attributes is equivalent to partitioning the initial training set into smaller training sets repeatedly, until the entropy of each of these subsets is zero (i.e. each one has instances drawn from only a single class). The ‘entropy method’ of attribute selection prefers to split on the attribute that gives the greatest reduction in (average) entropy, i.e. the one that maximizes the value of information gain.

Information gain attribute selection measure allows multi-way splits (i.e. two or more branches to be grown from a node) unlike gini index attribute selection measure which force the tree branching to be binary .Many algorithms attempt to "prune", or simplify, their results; such as J48 algorithm which is actually a slightly improved and the latest version of a very popular C4.5 algorithm. Decision tree pruning produces smaller tree, less complex and more easily interpretable results. They are usually faster and better at correctly classifying independent test data than unpruned trees. Decision tree pruning attempts to identify and remove many of the branches that may reflect noise or outliers in the training data (problem of overfitting*)*, with the goal of improving classification accuracy on unseen data (Witten & Frank, 2000).

There are two common approaches to tree pruning which are prepruning(forward pruning) and postpruning(backward pruning). In the case of prepruning approach, a tree is "pruned" by halting its construction early (e.g.,by deciding not to further split or partition the subset of training tuples at a given node).Upon halting, the node becomes a leaf. The leaf may hold the most frequent class among the subset tuples or the probability distribution of those tuples. On the other hand postpruning removes subtrees from a "fully grown" tree. It is the more common approach. Here a subtree at a given node is pruned by removing its branches and replacing it with a leaf. The leaf is labeled with the most frequent class among the subtree being replaced. Postpruning requires more computation than prepruning, However it generally leads to a more reliable tree.
No single pruning method has been found to be superior over all others. Although some pruning methods do depend on the availability of additional data for pruning, this is usually not a concern when dealing with large databases. Even though, pruned trees tend to be more compact than their unpruned counterparts, they may still be rather large and complex (Han &Kamber , 2006).

There are two different operations have been considered for postpruning: subtree replacement and subtree raising. At each node, a learning scheme might decide whether it should perform subtree replacement, subtree raising, or leave the subtree as it is, unpruned (Witten & Frank, 2000; Han &Kamber , 2006).

There are two pruning operation employed in the influential decision tree-building system such as C4, J48. The first is known as subtree replacement, is the primary pruning operation. Here the idea is to select some subtrees and replace them with single leaves, basically reducing the number of tests along a certain path. This process starts from the leaves of the fully formed tree, and works backwards toward the root. On the other hand, the second type of pruning operation is more complex, and it is not clear that it is necessarily always worthwhile or often has a negligible effect on decision tree models. In this case, a node may be moved upwards towards the root of the tree, replacing other nodes along the way (Witten & Frank, 2000).

2.5.2.2 Rule induction

Rule induction rules are a popular alternative to decision trees. The antecedent, or precondition, of a rule is a series of tests just like the tests at nodes in decision trees, and the consequent, or conclusion, gives the class or classes that apply to instances covered by that rule, or perhaps gives a probability distribution over the classes. Generally, the preconditions are logically ANDed together, and all the tests must succeed if the rule is to fire. It is easy to read a set of rules directly off a decision tree. One rule is generated for each leaf (Witten & Frank, 2000).

The antecedent of the rule includes a condition for every node on the path from the root to that leaf, and the consequent of the rule is the class assigned by the leaf (Witten & Frank, 2000).
Rule induction algorithms generate a model as a set of rules. The rules are in the standard form of IF THEN rules. Witten et al., (1999) proposed PART, separate-and-conquer rule induction algorithm. This algorithm generates sets of rules called ‘decision lists’ which are ordered set of rules and it builds a partial C4.5 decision tree in each iteration and converts the "best" leaf into a rule. The algorithm is a mixture of C4.5 and RIPPER (Repeated Incremental Pruning to Produce Error Reduction) rule learning (Nilgün & Özgür, 2010).
In this research PART rule induction algorithm was used PART rule induction algorithm was used, due to its ability and potential to produce accurate and readable rules (Daud & Corne, 2007; Witten & Frank, 2000).

2.5.3 Association rules

For descriptive task, association rule mining is one of the most important and well researched techniques of data mining. It is initially used for market basket analysis. It finds all the rules existing in the transactional database that satisfy some minimum support and minimum confidence constraints (Vyas, 2010).
Association and classification rules are similar except that association rules can represent association of any attribute including the class and this gives them the freedom to see the combinations of attributes. Whereas classification rules predict only the class. Different association rules express different regularities that underlie the dataset, and they generally predict different things (Witten & Frank, 2000).

The association data mining problem involves finding all of the rules (or at least a critical subset of rules) for which a particular data attribute is either a consequence (outcome) or an antecedent (precursor). This type of problem is very common in health care professionals who are looking for relationships between diseases and life-styles or demographics or between survival rates and treatments, as an example. The goal of using associations is to find common relationship among the items, or variables, existing in a collection of records (Houston, 2000).

One of the common practical application of association can be on market basket analysis .Here association rules can be developed in order to determine arrangements of items on store shelves in a given supermarket so that items often bought together will be found together .Association, unlike classification, can predict any field; not only the class and it can forecast more than one attribute's value at a time. For this reason we can find a number of association rules than classification rules (Berry & Linoff, 1997).

We can write associations as A ⇒B, where A is called the antecedent or left-hand side (LHS), and B is called the consequent or right-hand side (RHS).
For example, in the association rule “If people buy a hammer then they buy nails,” the antecedent is “buy a hammer” and the consequent is “buy nails”.

The frequency with which a particular association (e.g., the item set "hammers and nails") appears in the database is called its support or prevalence. If, say, 15 transactions out of 1,000 consist of "hammer and nails," the support for this association would be 1.5%.

To discover meaningful rules, however, we must also look at the relative frequency of occurrence of the items and their combinations (Two crow corporation, 1999).

Given the occurrence of item A (the antecedent), how often does item B (the consequent) occur? That is, what is the conditional predictability (Confidence) of B, given A? using the above example, this would mean asking "When people buy a hammer, how often do they also buy nails?". Confidence is calculated as a ratio: (frequency of A and B)/(frequency of A).

To clarify the above concepts of confidence and support, it is better to use the next example.

Let's consider hypothetical database that has the following features:

Total hardware-store transactions: 1,000, Number which include "hammer": 50.

Number which include "nails": 80, Number which include "hammer" and "nails": 15

Then from the database we can now calculate:

Support for "hammer and nails" = 1.5% (15/1,000), Confidence of "hammer ⇒nails" = 30% (15/50) ,Confidence of "nails ⇒hammer" = 19% (15/80).

The prevalence of this hammer-and-nails association (i.e., the support is 1.5%) is high enough to suggest a meaningful rule (Two crow corporation, 1999).

The main challenge when mining association rules is the immense number of rules that must be considered theoretically. The number of rules grows exponentially with itemsets since it is neither practical nor desirable to mine such a huge set of rules, the rule sets are typically restricted by minimal thresholds support and confidence. Support and confidence are two probability measures, which are used to assess the associations of itemsets(Levin & Zahavi ,1999).

In general association rules are the most popular kind of rules generated from frequent patterns. Typically, such mining can generate a large number of rules, many of which are redundant or do not indicate a correlation relationship among itemsets. Thus, the discovered associations can be further analyzed to discover statistical correlations, leading to correlation rules(Han & Kamber,2006).

Computational efficiency is one of the greater concern in association rule mining, since each of the attributes can appear with any of its possible values and anywhere in the association rule. Therefore; there might be a very large number of possible rules. To identify the best quality of association rules, we need to see a kind of rule interestingness measures (Bramer, 2007).

Rule interestingness measures adopted in this research is the one proposed in Bramer (2007), let us see its detail below.

In association rule mining, the rules appear in the form of:

If Antecedent Then Consequent or If LEFT then RIGHT

Let us start by defining four numerical values which can be determined for any rule simply by counting:

Let; $N_{Antecedent}$ be the number of instances matching Antecedent.

$N_{consequent}$ be the number of instances matching Consequent.

N_{BOTH} be the number of instances matching both Antecedent and Consequent.

N_{TOTAL} be the total number of instances.

Measures of rule interestingness (support and confidence),can be computed from these four numerical values. The proportion of both consequent and antecedent occurring together to the occurrence of the antecedent is called confidence and it is computed as:

Confidence (Predictive Accuracy, Reliability) = N_{BOTH} / Antecedent

The proportion of the training set correctly predicted by the rule is called support and computed as:

Support (Prevalence) = N_{BOTH} / N_{TOTAL}

As stated in Han and Kamber (2006), association rules are considered interesting if they satisfy both minimum support and confidence thresholds. Rules that satisfy both a minimum support threshold and a minimum confidence threshold are strong rules.

Apriori algorithm is used for mining association rules. That is, given a database consisting of tuples, it finds association rules that frequently and reliably predict which items occur together. Since the association algorithm results in several rules, it is imperative that mining be limited by using certain parameters so that only interesting association rules with high coverage will be found (Agrawal, 1994).

2.6 Data mining Methodologies

Data mining methodologies that can be used while data mining research is conducted includes, SEMMA, KDD and CRISP .Here under a brief discussion about these methodologies together with their summarized comparison is given.

2.6.1 Knowledge Discovery in Databases (KDD) methodology

The term KDD was coined in 1989 to refer to the broad process of finding knowledge in data, and to emphasize the "high-level" application of particular data mining methods.KDD process is the process of using the database along with any required selection, preprocessing, subsampling, and transformations of it; to apply data mining methods (algorithms) to enumerate patterns from it; and to evaluate the products of data mining to identify the subset of the enumerated patterns.
The five stages of KDD Process are listed bellow (Fayyad et al., 1996).

Selection: this stage consists of creating a target data set, or focusing on a subset of variables or data samples, on which discovery is to be performed.
Preprocessing: this stage consists of the target data cleaning and pre processing in order to obtain consistent data.
Transformation: this stage consists of the transformation of the data using dimensionality reduction or transformation methods.
Data mining: this stage consists of the searching for patterns of interest in a particular representational form, depending on the DM objective (usually, prediction).

Interpretation or Evaluation: this stage consists of the interpretation and evaluation of the mined patterns. The KDD process is described in a graphical form in Figure 2.3(Fayyad, Piatetsky-Shapiro & Smyth,1996).

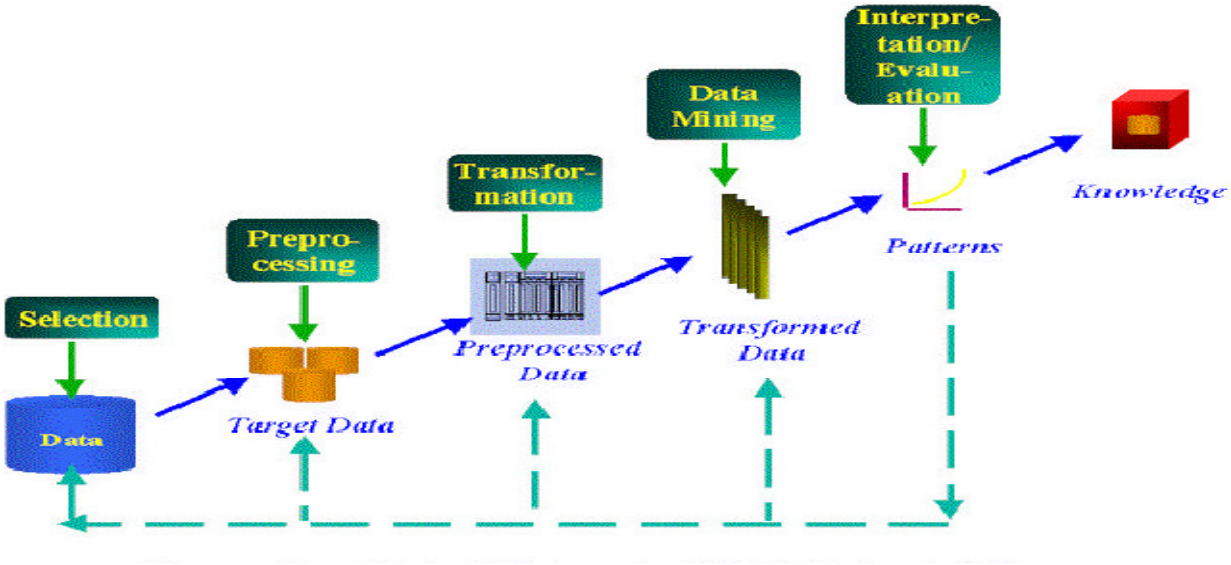


Figure 2.3 KDD Process steps [source Fayyad, Piatetsky-Shapiro & Smyth]

2.6.2 SEMMA methodology

SEMMA stands for Sample, Explore, Modify, Model, and Assess. It was developed by SAS(Statistical Analysis System) Institute. According to SAS Institute SEMMA is a cycle methodology with 5 stages. These are:

Sample: this stage consists of sampling the data by extracting a portion of a large data set big enough to contain the significant information, yet small enough to manipulate quickly.

Explore: this stage consists of the exploration of the data by searching for unanticipated trends and anomalies in order to gain understanding and ideas.

Modify: this stage consists of the modification of the data by creating, selecting, and transforming the variables to focus the model selection process.

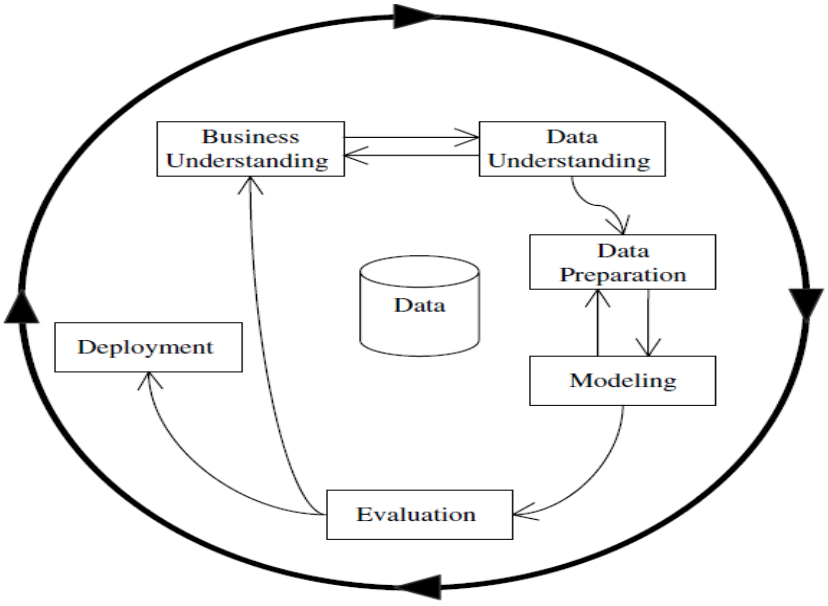
Model: this stage consists of modeling the data by allowing the software to search automatically for a combination of data that reliably predicts a desired outcome.

Assess: this stage consists on assessing the data by evaluating the usefulness and reliability of the findings from the data mining process and estimate how well it performs.

SEMMA offers an easy to understand process, allowing an organized and adequate development and maintenance of data mining projects. It thus confers a structure for conception, creation and evolution, helping to present solutions to business problems as well as to find data mining business goals (Azevedo &Santos, 2008).

2.6.3 CRISP (CRoss-Industry Standard Process) methodology

CRISP which is a methodology applied for this research purpose has different phases. It was first established in the late 1990s by four companies: Integral Solutions Ltd. (a provider of commercial data mining solutions), NCR (a database provider), DaimlerChrysler (an automobile manufacturer), and OHRA (an insurance company). The last two companies served as data and case study sources. The life cycle of a data mining research consists of six phases according to CRISP. The sequence of the phases is not rigid. Moving back and forth between different phases, as necessary, is required.The steps of CRISP methodology briefly summarized below based on Chapman et. al.(2000). The phases of CRISP are graphically depicted in Figure 2.4.



[Source: Chapman et al., (2000)]

Fig. 2.4: phases of CRISP

The first phase is known as Business understanding. It is known that business or problem understanding is one of the essential activities in research work, which can support to understand the research area or problem domain. According to Chapman et al.(2000) this phase focus on understanding the research objectives, requirements and goals from a business perspective.

The second phase is data understanding which starts with an initial data collection and proceeds with activities in order to get familiar with the data, to identify data quality problems and to discover first insights into the data.

The third phase is known by the name data preparation; it covers all activities to construct the final dataset from the source data. Real world databases usually contain incomplete, noisy and inconsistent data such data may cause confusion for the knowledge discovery process. Hence, data cleaning is mandatory to improve the quality of data and so as to improve the accuracy and efficiency of the knowledge discovery process .Data preprocessing is a routine task that usually consumes much of the efforts in the entire data mining process. Generally speaking, preparing the data is a step where much time is devoted (Han & Kamber, 2001). According to Saarenvitra's (2001) estimation, this step can take up to 80% of the total project effort.

The fourth phase is data mining modeling. It is only at this point that one invokes data mining models and tools to interrogate the data and convert it into knowledge for decision making. There are different activities to be done at this step such as selecting data mining tools, transforming the data if the tool requires it, generating samples for training and testing the model, and finally using the tools to build and select a model. The resulting models might have important patterns to be analyzed and interpreted to be used as decision support knowledge (Han & Kamber, 2006).

The fifth phase is Evaluation. At this stage, there is evaluation on the model(s) and review of steps executed to construct the model to be certain that properly achieves the business objectives.

The last phase is known as deployment. It often involves applying selected model(s) within an organization's decision making processes. However, 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 organization.

In many cases it is the user, not the data analyst, who carries out the deployment phase. Even if, the analyst will not carry out the deployment effort it is important for the user to understand up front what actions need to be carried out in order to actually make use of the created models.
The activity done at each phase is found in chapter one section 1.4

2.6.4 Comparison among the above data mining methodologies

When we compare KDD and SEMMA phases, they are equivalent or correspond each other; i.e Sample stage of KDD corresponds with Selection stage of SEMMA. The five stages of SEMMA process can be seen as a practical implementation of the five stages of the KDD process.
Comparing the KDD phases with the CRISP phases is not as straightforward as in the SEMMA situation. Nevertheless, we can first of all observe that the CRISP methodology incorporates the steps that precede and follow the KDD process that is to say: The Business understanding phase can be identified with the development of an understanding of the application domain. Concerning the remaining stages, we can say that: The data understanding phase of CRSIP can be identified as the combination of selection and pre processing stage of KDD; the data preparation phase of CRISP can be identified with transformation stage of KDD; the modeling phase of CRISP can be identified with DM stage of KDD and The Evaluation phase of CRISP can be identified with interpretation or evaluation stage of KDD(Azevedo &Santos, 2008).
Table 2.1 presents a summary of their correspondences based on Azevedo &Santos, 2008.

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

Table 2.1. Summary of KDD, SEMMA and CRISP Methodologies phases.

Both SEMMA and CRISP can be viewed as an implementation of the KDD process. They guide people to know how data mining can be applied in practice in real systems. At first sight, we can get to the conclusion that CRISP is more complete than SEMMA. However, analyzing it deeper, we can integrate the development of an understanding of the application domain, the relevant prior knowledge and the goals of the end-user, on the sample stage of SEMMA, because the data can not be sampled unless there exists a truly understanding of all the presented aspects (Azevedo & Santos, 2008).

2.7 The Application of Data mining Technology

When we consider the last few years, Data Mining tools have been used mainly in experimental and research environments. But recently the application of data mining also answers business questions that were previously impossible, impractical or unprofitable to address (Goebel & Gruenwald, 1999). Moreover, data mining has been providing substantial contribution to the business environment, and becoming increasingly popular. Many organizations are using data mining to help manage all phases of the customer life cycle, including acquiring new customers, increasing revenue from existing customers, and retaining good customers. By determining characteristics of good customers (profiling), a company can target prospects with similar characteristics. By profiling customers who have bought a particular product it can focus attention on similar customers who have not bought that product .Other business sectors such as insurance companies, telecommunications, credit card companies and stock exchanges are also interested in applying data mining technology to reduce fraud (Two Crows Corporation,1999).

As a result of vast amount of textual data available, Information retrieval has typically been concerned with finding better techniques to query for and retrieve textual documents based on their content. Data mining is being applied to this area so that the vast amounts of electronic publications currently available may be brought to users' attention in a more efficient manner Application of data mining also shows fruitful result in medical applications. Here, data mining can be used to predict the effectiveness of surgical procedures, medical tests or medications, to characterize patient behavior so as to predict office visits, and to identify successful medical therapies for different illnesses(Carbone, 1997; Two Crows Corporation, 1999).

The major reason that data mining has attracted a great deal of attention in the information industry in recent years is due to the wide availability of huge amounts of data and the imminent need for turning such data into useful information and knowledge (Han & Kamber, 2001).

2.7.1 Application of data mining in Healthcare

In today's healthcare system specialized tools are required for storing, accessing, analysis, and effective use of data as extensive amounts of data gathered in heath care databases. Medical informatics may use data mining technology. It provides a user- oriented approach to novel and hidden patterns in the health care data. The discovered knowledge can be used by the healthcare administrators to improve the quality of service, and by the medical practitioners to reduce the number of adverse drug effect and to suggest less expensive therapeutically equivalent alternatives. Hence, it is possible to say data mining techniques can be applied to create knowledge rich health care environment (Kaur & Wasan , 2006).

Healthcare systems contain huge amount of data and they are generally perceived as being information rich, yet knowledge poor environment since there is a lack of effective analysis tools to discover hidden relationships and trends in data. Data mining have found numerous applications in healthcare system as business and scientific domain. Its techniques can be used to investigate important knowledge in healthcare system that could improve the situation (Srinivas , Kavihta & Govrdhan ,2010).

According to Kaur and Wasan (2006) the following are some of the important areas of interests in which data mining techniques are applicable in health care .

Data modeling for health care applications
Executive Information System for health care
Forecasting treatment costs and demand of resources
Anticipating patient's future behavior given their history
Public Health Informatics and Health Insurance.

According to Mary & Mat (2004) the history of data mining applications in business and marketing organizations or sectors were a head of healthcare on using it to derive knowledge from data. But, now this thing is quickly changing because successful mining applications have been implemented in the healthcare arena, three of which are described below.

Hospital Infection Control

Early recognition of outbreaks and emerging resistance requires proactive surveillance. Computer-assisted surveillance research has focused on identifying high-risk patients and possible cases.

Data mining techniques can be applied in surveillance system to identify new and interesting patterns in infection control.

Ranking Hospitals

Data mining techniques have been implemented to examine reporting practices. With the use of International Classification of Diseases (ICD), codes (risk factors) and by reconstructing patient profiles, cluster and association analyses can show how risk factors are reported.

Identifying High-Risk Patients

A robust data mining and model-building solution identifies patients who are trending toward a high-risk condition.

This information gives nurse care coordinators a head start in identifying high-risk patients so that steps can be taken to improve the patients' quality of healthcare and to prevent health problems in the future.

2.8 Issues and Challenges of Health care Data mining

There are various problems that hamper the application of data mining on efficient manner. From these problems lack of information is one of the greatest obstacles to achieving efficient data mining .One reason about the lack of appropriate data for mining is that very often a database is created for reasons that have nothing to do with data mining. That means, mostly databases are built for other intended purpose; as a result, the important information may be missing. Incorrect data is another problem (Giudici, 2003).

Applying data mining in the medical field is a very challenging undertaking due to the idiosyncrasies of the medical profession. There are several inherent conflicts between the traditional methodologies of data mining approaches and medicine. In medical research, data mining starts with a hypothesis and then the results are adjusted to fit the hypothesis. This diverges from standard data mining practice, which simply starts with the data set without an apparent hypothesis. Traditional data mining is concerned about patterns and trends in data sets, data mining in medicine is more interested in the minority that do not conform to the patterns and trends. What heightens this difference in approach is the fact that most standard data mining is concerned mostly with describing but not explaining the patterns and trends. In contrast, medicine needs those explanations because a slight difference could change the balance between life or death. Even if data mining results are credible, convincing the health practitioners to change their habits based on evidence may be a bigger problem. Most doctors prefer to listen to a respected opinion leader in the medical profession, rather than to the result of data mining.Privacy of records and ethical use of patient information is also one big obstacle for data mining in healthcare (Mary & Mat, 2004).

2.9 Related Works

There are a lot of health care-related works which were conducted using traditional statistical analysis and data mining technology. The following are some of them:

Tsegaye et. al (2010) conducted a statistical research with a title "Pattern of Fatal Injuries in Addis Ababa, Ethiopia: A One-year Audit." in Addis Ababa, Ethiopia.
In their introduction they explained about injury which is one of the leading causes of death and disability in the developed countries and it is common but largely neglected health problem in developing countries including Ethiopia .It is the primary reason for an emergency hospital visit in Addis Ababa hospitals such as Tikur Anbessa Specialized Hospital.

Their audit was devised mainly to assess the burden of fatal injuries together with identifying common causes of fatal injuries in Ethiopia. It was specifically designed to determine the profile or pattern of commonly occurring fatal injuries, and medical attention received before death.

They utilized data from the pathology department of Menilik II hospital, during one year period, between July 1, 2006 and June 30, 2007. In their method, they stated that they used a prospective descriptive study where the data on injury was extracted from the registration book of the department by using structured question of Fatal Injury Surveillance Data Collection Form prepared by Addis Ababa City Administration Health Bureau. In their result they stated a total of 2985 dead body was seen at the pathology department of Menilik II hospital, of which 2,107 (70.4%) were related with injuries.

Injuries occurred on the road account the larger amount that is 868(41.2%), followed by home 253 (12.0%) and only 68 (3.2%) occurred at work place but in 374 (17.8%) was not possible to know the place. Above 75% which is exactly 1715 (81.4%) of the victims did not receive post crash care at any level either pre-hospital or hospital/health facility and around 1139 (54.1%) of victims died because of accident, 641 (30.4%) homicide, 234 (11.1%) suicide, and 93 (4.4%) was not possible to determine the circumstance of death.
Health care-related literatures which were done by data mining technology include the following:

Vararuk et al. (2008) conducted a data mining research to investigate patterns in HIV/AIDS patient data in Thailand through the use of data mining techniques. As the researchers stated, these patterns can be used for better management of the disease and will be more appropriate for targeting of resources.

In their methodology part, stated they took a total of 250,000 records from HIV/AIDS patients imported into a database. IBM's Intelligent Miner was used for clustering and association rule discovery. On their finding they stated that clustering highlighted groups of patients with common characteristics. Unexpected association rules were identified that were not expected in the data and were different from traditional reporting mechanisms utilized by medical practitioners. The significance of the research was stated as to show that providing a realistic and targeted approach to the management of resources available for HIV/AIDS treatment can provide a much better service, while at the same time reducing the expense of that service.

The study could also be used as a means of implementing a quality monitoring system to target available resources.

There are some health care related works that apply data mining technology in Ethiopia.

Abraham Tesso (2005) had done a research in health with a title- "Application of Data Mining Technology to identify determinant factors of HIV Infection and to find their Association Rule": Case of Center for Disease Control and Prevention (CDC).

The objective of the study was to see the potential applicability of data mining on VCT data to broaden the insight regarding determinant factors of HIV/AIDS infection. The researcher used 5267 records of VCT service visitors for the research and he did various data preprocessing activities to generate the final dataset that was used to build the model.The researcher used association rule that was done by Apriori Algorithm. The researcher reported that his research output (discovered) promising result, that can benefit and is useable by health professionals, government, policy makers and the society at large.

Shegaw (2002) has conducted research in health care that apply data mining with the title "Application of Data Mining Technology to Predict Child Mortality Patterns: The Case of Butajira Rural Health Project (BRHP)" .The objective of the research was to investigate the potential applicability of data mining technology in developing a model that can support primary health care providers, policy makers, planners that used to identify the major determinants of child mortality,to prevent and control child mortality in the district of Butajira.

The researcher selected and extracted the source data from the ten years surveillance dataset of the BRHP epidemiological study which contains a total of 64,077 records, from this data a sample dataset consisting of 1,100 records of both alive and dead children selected randomly to build predictive models using neural network and decision tree techniques. He had gone through different activities such as data collection, preparation, model building and testing. The researcher used Neural Network and Decision tree Classifiers for classification purpose using BrainMaker tools.

The researcher stated that several neural networks and decision tree models were built and tested for their classification accuracy and many models with encouraging results were obtained.
Furthermore, he compared the two methods used as “unlike the neural network models, the results obtained by using the decision tree approach provided simple rules that can be used by non- technical health care professionals to identify cases for which the rule is applicable”.
The researcher also showed that an epidemiological database could be successfully mined to identify public health and socio-demographic determinants (risk factors) that are associated with infant and child mortality in rural communities.

CHAPTER THREE

BUSINESS UNDERSTANING AND DATA PREPROCESSING

3.1 Business Understanding

Both business and data understanding are phases in the life cycle of a data mining project. Business understanding is considered as the initial phase in CRISP methodology which is concerned with project objectives understanding, requirements from a business perspective, changing business knowledge or understanding a data mining problem definition and a preliminary plan designed to achieve the objectives. Understanding the business domain is a key in addition to using appropriate techniques and tools for ending up in a good result .More- over having an in-depth knowledge in the business domain enables data analysts to clearly set the objectives (Chapman et. al., 2000).

Business understanding is one of the essential phases in data mining process. Hence, in this research there were some measures taken to understand the business by using different methods such as reviewing documents, observation and discussion with domain experts. Business understanding measures, carried out by the researcher, start from a brief overview of the Hospital.

Tikur Anbessa Specialized Hospital is the largest, referral and teaching hospital in the country, with 560 beds and is located in Ledeta sub-city, Addis Ababa, Ethiopia. The hospital has different departments such as emergency medical service unit, diabetics unit, gynecology unit,cancer unit, orthopedic surgery department and so on .It receives referred and some directly visiting patients from all parts of the country and provides emergency service(Yishak,2009).

3.1.1 Work flow (the whole process) in Tikur Anbessa Specialized Hospital emergency medical service unit

The entire treatment process in emergency medical unit of the hospital may take short or long time depending on the type of emergency case, number of patients, availability of resources and other factors. In general, the overall processes followed to get service at emergency unit is presented below:

Step 1: Separation and registration in triage section: The patient whose case is suspected as emergency first comes to the Hospital emergency medical service unit Triage area. Here the nurse differentiates whether the patient is emergency case or not by observing the case, checking vital signs and asking the patient or his/her contact person. If the nurse believes the case is emergency, he or she provides give card to patient.Then, the patient takes out medical card from the hospital card room. In the card room, various socio- demographic information such as (name, address, age, sex, marital status and so on) is recorded on the medical card of the patient.

After separation, triage nurse fills various information such as socio demographic information, event information (i.e, activity, place of injury and intent) and some clinical information (i.e vital sign data and history of past medical illness) on the triage form by observation, asking the patient or his/her contact person, observing medical card information and values obtained from vital sign medical instruments.
Finally the nurse indicates or shows the patient where to go for treatment.

Step2: Treatment: In the treatment section of the emergency unit responsible doctors diagnose the patient for further investigation. Depending on the case, the doctor may order laboratory,
X-ray, surgery, medicine, refer for admission in emergency department, to other units even to other Hospitals .In this section, the doctor records the details of the investigation on the patient's medical card.

Step3: Entering patient information on the emergency medical registration data base:
In this section, the data clerk (nurse in profession) enters or fills the details of the patent socio-demographic, event and triage data from the medical card and triage form to the database.
After the clerk finishes entering all the necessary data, the medical card together with triage form is returned and stored to the card room.
Step4: Follow up: If the patient is admitted, then, normal follow up process is continued .Here, nurses and doctors carry out the process.

As statistics shows emergency medical cases are among the main causes of disability and mortality in Ethiopia. There are large number of patients who visit Tikur Anbessa specialized Hospital. Since it is one of the specialized Hospital in the country, it is also a referral Hospital that can accept patients all over the country and its location also make it the center for many of the emergency medical cases. These and other reasons make the hospital very crowded with patients in general and emergency case patients in particular. As a result, there is a huge volume of emergency medical patients' data which is kept unprocessed with regard to knowledge generation.

Therefore the main reason that necessitated this research is to explore these huge volumes of data that can be used to discover knowledge and patterns that can play a role in awareness, planning and decision-making in emergency care environment. The data can be used beyond simple statistical analysis such as data mining. It uncovers important data patterns that contribute greatly to business strategies in providing a novel knowledge that can be used as a base for guidance and decision making.

Identifying the important factors or variables among the emergency patient variables have significant impact for the problem domain. Moreover, predicting or classifying which age group, sex, place, activity and other independent variables are more likely to be associated with the cause of visit dependent variable which is important for the situation. Hence, data mining technology can offer enormous potential in predicting and mining the hidden characteristics and patterns that exist within the dataset.

3.2 Data Understanding

Data understanding is one of the critical phases in data mining research. It starts with an initial data collection and proceeds with activities in order to get familiar with the data, to identify data quality problems, to discover first insights into the data or to detect interesting subsets to form hypotheses for hidden information (Chapman et. al., 2000).

The researcher consulted domain experts to understand the situation related to emergency and to have an overview of the problem domain. The domain experts that were communicated include doctors, nurses, data clerks and database administrator from emergency unit of the Hospital. Among these experts, doctors and nurses are concerned in the process of patient treatment and the consultation of them provides what exactly the business is and what kind of data is captured during emergency medical care.

The Data clerks are actually concerned on entering the data from the medical card and triage form to the database through user interfaces. The consultation of them provides the overall picture of data encoding to the system such as the flow or logic, problems encountered and so on.

The Database administrator is concerned with the database design and management basically taking backup, maintaining the system and training users. The consultation of the administrator provides the overall picture of emergency medical registration system such as the database tables, relationships among tables, attributes and interfaces .The source data is also obtained by consulting the administrator. Data quality problems were identified with the discussion of data base administrator, triage nurses and data clerks. As a result the following factors can be raised; inconsistent data encoding, incorrect source data value, missed or unknown source data value, wrongly recorded data value on the patient card or database and so on.

3.2.1 Data collection

The source data employed in this research was collected from Tikur Anbessa Specialized Hospital emergency medical registration database. A backup of the database tables which were exported in to excel format was taken from emergency unit.

The Hospital emergency medical registration database contains around 5708 total records during the time of data collection (until May 2011 G.C) but 5524 records were used for the research experimentation since 184 records were removed due to inconveniences because some were duplicate and did not have medical card number.

3.2.2 Data source description

The emergency medical database is designed by MySQL Database Management System (DBMS). The data clerks (who are nurses by profession) have the interface (designed by web-based system) and through which they can enter patient demographic, Event, Triage, Trauma and Non-trauma related data. The data clerk first enters his or her user name and password for authentication purpose. Once the user is authorized the system displays personal information or socio demographic page and the user fills the necessary data from the card and triage form. The database contains more than four tables or files, but for the research purpose four tables were selected based on significance to the problem domain. The tables were Personal information or socio demographic, Event,Trauma triage and Non-trauma triage.

Each table has more than five attributes (variables) but based on the appropriateness to the problem domain the following fourteen were selected from all tables. The selected attributes from the Personal information table include Medical Record Number,Sex ,Age ,Marital status ,Subcity and Region more description for all selected attributes is given in this section on Table 3.1. The selected attributes from the Event table include Medical Record Number, cause of visit, intent, activity, place of injury and referred from. The selected attributes from each of the two tables (Trauma and Non-trauma triage tables) include Medical Record Number, past medical illness, Triage assessment and Transferred to. These four tables were joined by Medical Record Number (i.e it is primary key which uniquely identifies records)to get the full integrated attributes for each individual patient record.

Full description for the above selected attributes is given in the table 3.1.

S.No	Attributes	Description	Values	Data Type
1	MRNo	Medical Record Number	Numeric value	numeric
2	Sex	Gender of the patient	Female and Male	Nominal
3	Age	Numeric age value	Age of the patient	Numeric
4	Marital Status	Marital Status of the patient	Married, Divorced, Single, Other	Nominal
5	Subcity	The sub city of the patient	Addis Ketema, Bole, Akaki Kaliti, Gulele,Yeka ,Kirkos ,Ledeta, Arada Nefas SelkLafto and Kolfe Keranio	Nominal
6	Region	The region of the patient	Addis Ababa,Amhara, Oromia, Afar, Harrari, Somale, SNNPR, Dire Dawa,Benshangu Gumez,and Tigray	Nominal
7	Cause of visit	The cause or event for patient's Hospital visit	Trauma Non-Trauma	Nominal
8	Triage Assessment	The level of emergency. Values are arranged from the most critical (Red) to lowest(Green) Black value represents the patient dead on arrival	Red Orange Yellow Green Black	Nominal
9	Intent	Describes how the injury or illness occurred	Accident, Violence and Other	Nominal
10	Activity	Describes the activity of the patient during injury or illness	Traveling, Working, Studying, Playing and Other	Nominal
11	Referred From	Describes the referral of the patient	Governmental Health Facility , Non Governmental Health Facility and Self	Nominal
12	place of injury	The place where the injury or illness occurred	Home, Street, Work, School, Recreational Place and Other	Nominal

S.No	Attributes	Description	Values	Data Type
13	Transferred to	The place or Special unit that the patient was transferred	EmergencyOPD,Stabilization room, ,Resuscitation room, Home, Regular OPD, and Referred to other Hospital	Nominal
14	Past medical illness	Describe whether the patient has any previously (Past) known medical illness	Yes, No and Unknown	Nominal

Table 3.1: Description of Attributes

3.3 Data Preprocessing

There a lot of problems that affects the quality of large data bases. These problems include noisy, incomplete, missing, and inconsistent data. Incomplete data can occur for a number of reasons. Attributes of interest may not always be available, due to misunderstanding while recording, or because of equipment malfunctions. There are many possible reasons for noisy data (having incorrect attribute values). The data collection instruments used may be faulty. There may have been human or computer errors occurring at data entry. Therefore; data preprocessing to clean noisy data, to handle missing values and inconsistent attribute values has an immense importance in data mining research. There are a number of data preprocessing techniques. Which includes data cleaning, it can be applied to remove noise and correct inconsistencies in the data; data integration, it merges data from multiple sources; data reduction, it remove irrelevant and redundant attributes (Han and Kamber, 2006).

As a result; this section is devoted to describe the different statistical summary measures which helps to see what is missing and data cleaning activities performed to reduce the data quality problems**.**

3.3.1 Descriptive statistical summary of attributes

Descriptive data summarization serves as a foundation for data preprocessing. It helps us to study the general characteristics of the data and identify the presence of noise or outliers, which is useful for successful data cleaning and data integration (Han & Kamber, 2006). The following sub section provide statistical summary of each attributes of the above table.

Sex attribute

It is nominal type attribute with two values (Male and Female). The modal value is 'male', which accounts for 64.5%. There is a 0.3 % missing value in this attribute. The statistical summary of the attribute is presented in Table 3.2.

	Frequency	percent
Missing value	19	0.3
Male	3562	64.5
Female	1943	35.2
Total	5524	100.0

Table 3.2: Statistical summary of sex attribute

Age attribute

Age is numeric type attribute and holds numeric age values in years of individual patient. Its Inter Quartile Range (IQR) is 23. The most frequent age is 25 year and the mean is 35 year. There is no missing value in this attribute. Its statistical summary is presented as in Table 3.3.

	Attribute type:Age		Statistic
	Missing value		0
	Mean		35
	Median		30
	Mode		25
	Variance		279.805
	Inter Quartile Range		23
	Minimum		1
	Maximum		127
	Range		126
	percentiles	25	22
		50	30
		75	45

Table 3.3: Statistical summary of age attribute

To find the records with outlier values, five number summary was done accordingly; first Quartile (Q1) is 22 and third quartile (Q3)is 45. The values bellow Q1- (1.5*IQR) which is 22-(1.5*23)=13 can be considered as outlier values(lower limit) as many literatures recommend ,similarly the values above Q3 + (1.5*IQR) which is 45+(1.5*23)=80 is the upper limit for outliers which means age beyond 80 can be considered as outlier values. Therefore, values for age bellow 13 and above 80 can be considered as outlier values. As a result around 47 outlier values for this attribute were detected. The researcher decides on the outlier values to be removed and replaced by the mean value (35 year).

Marital Status

It is nominal type attribute with 4 distinct values. They are married, single, other and divorced. It has 7.4% missed values .The most frequent (modal) value is married (49.9%) and the least frequent values are other and divorced with 0.1 %.

The statistical summary of this attribute is as shown in Table 3.4.

		frequency	percent
	Missing	407	7.4
Marital status	Married	2754	49.9
	Single	2351	42.6
	Divorced	4	0.1
	Other	8	0.1
	Total	5524	100

Table 3.4: Statistical summary of marital status attribute.

Region attribute

It is nominal type attribute that consist of the regional states of Ethiopia. The modal value is Addis Ababa (69.3%) and the least frequent value is Benishangul Gumuz with 0.1 %.

The statistical summary of this attribute is as shown in Table 3.5.

Region attribute		
	Frequency	Percent
AddisAbaba	3829	69.3
Afar	13	0.2
Amhara	251	4.5
Benishangul Gumuz	7	0.1
DireDawa	16	0.3
Gambela	3	0.1
Harrar	22	0.4
Oromia	1066	19.3
SNNPR	264	4.8
Somalia	12	0.2
Tigray	41	0.7
Total	5524	100.0

Table 3.5: Statistical summary of region attribute

Referred from attribute

It is nominal type attribute with 3 distinct values, Governmental Health Facility, Non - Governmental Health Facility and Self. The modal value is Governmental Health Facility (44.13%) and the least frequent value is Non -Governmental Health Facility with 17.65 %.

The statistical summary of this attribute is as shown in Table 3.6.

	frequency	percent
Missing value	844	15.27
Governmental Health Facility	2438	44.13
Non -Governmental Health Facility	975	17.65
Self	1267	22.93
Total	5524	100

Table 3.6: Statistical summary of referred from attribute

Intent attribute

It is nominal type attribute with 3 values.Violence, other, and accidental. The modal value is accidental (49.7%) and the least frequent value is violence with 18.8 %.

Statistical summary of the attribute is as shown in Table 3.7.

		Frequency	Percent
Values	Missing value	649	11.7
	accidental	2747	49.7
	Other	1090	19.7
	Violence	1038	18.8
	Total	5524	100.0

Table 3.7: Statistical summary of intent attribute

Cause of visit Attribute

This attribute is nominal type. It refers to for what cause or event the patients visit or come to the Hospital. The two distinct values for the attribute are 'Trauma' to mean the patient came to the Hospital due to trauma and 'Non-trauma' to mean the patient came to the hospital due to Non trauma case . This attribute is a class label (Dependent Variable).

The modal value is trauma (50.6%) and the least frequent value is Non- trauma with 49.4 %. Statistical summary of this attribute is presented in table 3.8.

Attribute name: Cause of visit			
Value		Frequency	Percent
	Trauma	2797	50.6
	Non-Trauma	2727	49.4
	Total	5524	100.0

Table 3.8: Statistical summary of the Cause of visit attribute

Activity Attribute

This attribute is nominal type with 5 values. They are working, playing or recreating, studying. travelling, and other. The modal value is working (25.7%) and the least frequent value is studying with 12.5 %. Statistical summary of this attribute is presented in table 3.9.

		Frequency	Percent
Values	Missing value	713	12.9
	Other	759	13.7
	Playing Or Recreating	697	12.6
	Studying	689	12.5
	Travelling	1246	22.6
	Working	1420	25.7
	Total	5524	100.0

Table 3.9: Statistical summary of the activity attribute

Sub city attribute

This attribute is nominal type that contains of the ten sub city of Addis Ababa and other value called outside Addis Ababa. The modal value is outside Addis Ababa (28.2%) and the least frequent value is Akaki Kaliti with 2.8 %.

Statistical summary of this attribute is presented in table 3.10.

		Frequency	Percent
Distinct Values	Missing	135	2.4
	Addis Ketema	298	5.4
	Akaki Kaliti	153	2.8
	Arada	409	7.4
	Bole	322	5.8
	Gulele	335	6.1
	Kirkos	499	9.0
	Kolfe Keranio	407	7.4
	Ledeta	679	12.3
	Outside Addis Ababa	1560	28.2
	Nefas Silk Lafto	419	7.6
	Yeka	308	5.6
	Total	5524	100.0

Table 3.10: Statistical summary of the subcity attribute

Where outside Addis Ababa: refers to Patients whose address is outside Addis Ababa since there is no sub city category there.

Triage Assessment Attribute

This attribute is nominal type with five distinct values. They are red, orange, yellow,black, and green in their urgency level. The attribute refers the level of emergency, the values are arranged in the table 3.11 based on risky (urgency) level to mean patient with red triage assessment (Indictor) is the most risky patient. So, needs immediate treatment; on the contrary green Indicator shows the least risky patient. Black triage assessment level refers the patient dead on arrival at the Hospital. The modal value is orange (50.4%) and the least frequent value is black with 0.4 %.

Statistical summary of triage assessment attribute is presented in table 3.11.

		Frequency	Percent
Values	Missing value	61	1.1
	Red	635	11.49
	Orange	2814	50.94
	Yellow	1724	31.21
	Green	265	4.79
	Black	25	0.4
	Total	5524	100.0

Table 3.11: Statistical summary of the triage assessment attribute

Transferred to Attribute

This attribute is nominal type. The modal value is emergency OPD (83.07%) and the least frequent value is home with 0.0036 %. Statistical summary is presented in table 3.12.

		Frequency	Percent
Values	Missing value	230	4.16
	Emergency OPD	4589	83.07
	Stabilization room	80	1.44
	Resuscitation room	582	10.54
	Home	2	0.0036
	Regular OPD	37	0.67
	Referred to other Hospital	4	0.072
	Total	5524	100.0

Table 3.12: Statistical summary of the transferred to attribute

Past Medical illness

It is nominal type attribute with three values. Yes, no, and unknown. The modal value is no(51.3%) and the least frequent value is unknown with 15.41 %.The statistical summary of this attribute is as shown in Table 3.13.

Values	Frequency	Percent
Missing value	649	11.75
Yes	1190	21.54
No	2834	51.3
Unknown	851	15.41
Total	5524	100

Table 3.13: Statistical summary of past medical illness attribute

3.3.2 Data cleaning

In today's dynamic world data tends to be incomplete, noisy, and inconsistent. As a result data cleaning is important to handle such data quality problems and as a result tasks which require quality data can get benefit from it. Data cleaning is a routine attempt to fill missing values, smooth out noise while identifying outliers, and correct inconsistencies in the data. It also refers to the preprocessing of data in order to remove or reduce noise (i.e by applying smoothing techniques) and the handling of missing values (e.g. by replacing a missing value with the most commonly occurring value for that attribute, or with the most probable value based on statistics) (Han & Kamber, 2006).

From the above fact data cleaning task or process is essential. Hence, the researcher tried to carry out different data cleaning tasks in this subsection.

3.3.3 Handling missing values

There are a number of reasons in which missing data might occur such as the value might not be relevant to a particular case; value could not be recorded when the data was collected, or is ignored by users because of privacy concerns and an availability of data(Agrawal & Srikant, 2000). The data set that is used in this research has missing values which were already shown in the above statistical summary section. There are missing values on the variables Marital status, Sex ,Place of injury, Intent, Activity , Referred From **,**Triage assessment ,Transferred to ,Past medical illness and Sub city .There are also attributes that do not have missing values which include Region, Age and Cause of visit .

There are different mechanisms to handle missing values in data mining techniques For instance, if the variable is nominal replacing missing values with modal (the most frequent) value it is one option and if it is numeric type replacing missing values with mean is one option as it is recommended by Two Crows Corporation (1999).

As a result, the missing values of the dataset were handled in accordance with the above suggestion. The missing value of Sex ,Place of injury, Activity , Intent, Referred From ,Triage assessment ,Transferred to ,Past medical illness and Sub city variables were filled by their modal values since they are nominal type. For Marital Status attribute special consideration was made to handle its missing values that are based on the patients' age and the fact or trend seen from the data set. Accordingly for those patients whose age is between 13 and 24 years, their Marital Status missing values were replaced by 'single'. For those male patients whose age is between 25 and 27 years, their Marital Status missing values were replaced by 'single'. For those patients whose age is greater than 32, their Marital Status missing values were replaced by 'married' and other Marital Status missing values replaced by modal value.

3.3.4 Handling outlier values

The quality of output will often be sensitive to outliers which are data values that are different from the typical values in the database.

Outliers are data objects that do not comply with the general behavior or model of the data in the database. They may be the result of incorrectly entered data. Deviation-based methods identify outliers by examining differences in the main characteristics of objects in a group. The degree to which numerical data tend to spread is called the dispersion, or variance of the data. The most common measures of data dispersion are range, the five-number summary (based on quartiles), the Inter Quartile Range(IQR), and the standard deviation. Box plots can be plotted based on the five-number summary and are a useful tool for identifying outliers (Two Crows Corporation, 1999;Han & Kamber, 2006).

In this research, to handle outlier value that exists in the dataset, the recommendation stated by Han & Kamber (2006) was used. As Han & Kamber (2006) stated a common rule of thumb for identifying suspected outliers is to single out values falling at least 1.5*IQR above the third quartile or below the first quartile. In other words, it is to mean that the values outside the limits: Q3+ (1.5*IQR) and Q1-(1.5 *IRQ) will be considered as outlier values. Accordingly Age attribute which has outlier value is handled by the above recommendation. Hence, the Age outlier values replaced by mean values of the attribute (35). The lower and upper limit for the outliers in the Age attribute is 13 and 80 respectively as described under sections 3.2.3.
We can see section 3.2.3.for the detail of the calculation to find the outliers.

3.3.5 Handling noisy values

A noisy data is one of the common problem or challenge in data mining and knowledge discovery research. Noise is a random error or variance in a measured variable. It includes misclassified data or information as well as missing data or information .It can occur due to various factors such as an erroneous instrument measuring and human error when registering data. Noisy data will definitely minimize the accuracy of any data mining system. (Al Shalabi, 2009).

In this research noisy values were encountered in Age, sub city, activity and Medical Record Number attributes. In the case of age those patients whose age are less than 13 years are not treated in emergency medical service unit of the Hospital rather they are treated in separate unit (pediatrics unit) as mentioned by database clerks. In age variable 1and127 were encountered which are noise value. Since patients whose ages are bellow 13 are treated in other unit and age value 1and 127 also detected as outlier value. In the variables sub city and activity noisy values were obtained which were removed and replaced by their modal value. There were also noisy values (such as individual name and no card number as a value) in Medical Record Number variable which were removed by deleting the tuples since joining, files or tables having unique and correct value, for this variable is a must.

3.3.6 Data transformation and reduction

There are two purposes as far as data transformation is concerned .One, is to fix problems with the data such as missing values and categorical variables that take on too many values, and the second one is to bring information to the surface by creating new variables to represent trends and other ratios (Berry & Linoff,1997).

Data transformation can involve, smoothing or feature (attribute) construction, which works to remove noise from the data. Smoothing techniques include binning, regression, and clustering. It can also serve as data reduction, for example in the case of smoothing through binning, since the number of the distinct values for a certain attribute is reduced. Data reduction can reduce the data size much smaller in volume by using various strategies such as aggregating, eliminating redundant features, or clustering , removing irrelevant attributes, data discretization and dimensionality reduction (e.g., using encoding schemes). But it produces the same (or almost the same) analytical results. Data discretization techniques can be used to reduce the number of values for a given continuous attribute by dividing the range of the attribute into intervals. Interval labels can then be used to replace actual data values. Replacing numerous values of a continuous attribute by a small number of interval labels thereby reduces and simplifies the original data. This leads to a concise, easy-to-use, knowledge-level representation of mining results(Han & Kamber, 2006).

Based on the above suggestion, data transformation and reduction were applied during data preprocessing to improve the dataset for mining task. As a result, 'Age' attribute discretized(binned) using data discretization techniques to reduce its 79 distinct values so that it will suit to mining tool and obtain meaningful patterns. Data reduction was also applied on Medical Record Number and other unnecessary attributes (such as phone number, contact person data).Hence they were removed due to their irrelevance for mining task.

Binning is a top-down splitting technique based on a specified number of bins. Equal width (distance) partitioning divides the range into N intervals of equal size or uniform grid,this approach leads into bins with non uniform distribution of data elements per bin (Han & Kamber, 2006).

Since there is no standard way to categorize age, variable in the business domain (Emergency unit), binning method of weka was used. From the table below we can see the age attribute has 7 distinct categories (7 bins) which is convenient to mining task .The bin was selected to 7 based on the distribution of source data with respect to age and the generated model.

The next tables 3.14 summarize the discretized labels of age attribute.

Age Category	Frequency
13-25	2072
26-37	1453
38-49	795
50-61	715
62-73	337
74-85	136
>85	16

Table 3.14: Discretized result of age attribute

3.3.7 Summary of original and target datasets

The final data sets used for this research has the following attributes sex ,age ,marital status ,subcity , region ,cause of visit, intent, activity, place of injury , referred from ,past medical illness, triage assessment and transferred to. Table 3.15 presents the comparative summary of the original and the final datasets.

Parameters	Original dataset	Target dataset		
Number of attributes	20	13		
Total number of records	5708	5524		
File format	Excel(.xls)	.xls	.csv	.arff
File size	654KB	306KB	674KB	771KB

Table 3.15 Dataset summary for the original and target data

CHAPTER FOUR

EXPERIMENTATION AND ANALYSIS

This chapter is devoted to discuss on the models to be built and experiments carried out together with their analysis. For this research classification and association rule data mining models were selected and experimented as discussed under chapter one section 1.4.4. Thirteen attributes were selected using weka information gain, discussion with domain experts and subjective judgment.

4.1 Model Building

4.1.1 Attribute ordering

One of the essential tasks in data mining decision model is attribute selection which can have an impact on the models to be built. As a result the researcher tried to rank the attributes based on information gain, which was already reviewed under chapter two, section 2.5.2.1

The result of attribute ordering is shown in the figure 4.1.

```
=== Attribute Selection on all input data ===
Search Method:
        Attribute ranking.
Attribute Evaluator (supervised, Class (nominal): 13 Cause of visit):
        Information Gain Ranking Filter
Evaluation mode:    evaluate on all training data
Ranked attributes:
 0.58941      8 Place of injury
 0.45414      9 Activity
 0.2672       7 Intent
 0.06244     10 Past medical illness
 0.03404      6 Referred from
 0.02796      1 Sex
 0.01334     12 Transferred to
 0.01267      2 Age
 0.00795      4 Subcity
 0.00443      5 Region
 0.00377      3 Marital Status
 0.00269     11 Triage assessment

Selected attributes: 8,9,7,10,6,1,12,2,4,5,3,11 : 12
```

Fig. 4.1: Attribute ordering using information gain

Where the numbers that precede the attribute name(follow selected attributes:) are the attributes' indices in the dataset and the number 12 at the bottom shows total number of attributes ordered in determining the class label "Cause of visit", which is the last(13^{th}) attribute. The numbers in the left most side(such as 0.58941 ,0.45414) indicate information gain value.
From figure 4.1 ,the top three dominant attributes to determine the cause of visit are place of injury, activity and intent with information gain of 0.58941 ,0. 45414 and 0.2672 respectively**.**

Selecting or ordering attributes' in data mining task, as the one depicted above helps, for later experimentation like which attribute is relevant for the class level and which one is the least relevant attribute that can be removed step by step. The excel(.xls) dataset format first converted in to comma-separated values(.csv) then to attribute relation file format(.arff) for model building.

4.1.2 Building classification models

For the classification model five, experiments were built. Which are unpruned decision tree with confidence factor 0.15, pruned decision tree with confidence factor 0.15, pruned decision tree with default confidence factor and unpruned decision tree with default confidence factor. The fifth one is PART rule induction model. These experiments were also compared for their efficiency. The objective of conducting these five experiments were to explore the dataset with representative models which produce rules in acceptable performance. One or more algorithms may be better or worse for the data under investigation in data mining. Hence, it is important to select appropriate algorithm (Han & Kamber, 2006). Furthermore, it is good to select appropriate validation method for the dataset which categorizes the data in to training and testing purpose.

4.1.2.1 Validation Method Selection for decision tree models

Validation (measuring classifier accuracy on unseen data) is one thing that needs attention in classification model building. There are three main validation methods which are full training set (it divide the data into a training set and a test set), k-fold cross-validation (10-fold cross validation in weka) and N-fold cross-validation (or leave-one-out, where *n* is the number of instances in the dataset). It is important to check the appropriateness of dataset for selecting certain validation method (Bramer, 2007).

Generally making a model performance based on the training set is definitely not a good indicator of performance on an independent test set. If we have vast supply of data available, this is no problem: just we can make a model based on a large sample training set, and try it out on another independent large sample test set, provided that both samples are representative, the error rate on the test set will give a true indication of future performance generally, the larger the training sample the better the classifier accuracy. However, a problem occurs to choose which validation method is appropriate if the volume of the data source is limited and still controversial, one. The standard way of predicting the error rate of a learning technique given a single, fixed sample of data is to use stratified 10-fold cross-validation ,10 is selected because extensive tests on numerous datasets, with different learning techniques, have shown that 10 is about the right number of folds to get the best estimate of error. Leave-one-out cross-validation it increase the chance of classifier accuracy, but computationally expensive (Witten & Frank, 2000).

In k-fold cross-validation, the initial data are randomly partitioned into k mutually exclusive subsets or "folds," D_1, D_2,….. , D_k, each of approximately equal size. Training and testing is performed k times.In stratified cross-validation, the folds are stratified so that the class distribution of the tuples in each fold is approximately the same as that in the initial data.
In general, stratified 10-fold cross-validation is recommended for estimating accuracy (even if computation power allows using more folds) due to its relatively low bias and variance (Han & Kamber, 2006). Based on the above recommendation, stratified 10-fold cross validation was used to minimize the effect of the limited dataset used.

4.1.2.2 Building pruned decision tree

Pruned decision tree is one of the experiments to be implemented for decision tree classification model in this research.
According Han &Kamber(2006) decision tree pruning produces smaller tree, less complex and more easily interpretable results. They are usually faster and better at correctly classifying independent test data than unpruned trees. On the other hand, even if, pruned trees tend to be more compact than their unpruned counterparts, they may still be rather large and complex.

Due to this fact and the interest to see the variety of the generated rules with their accuracy measure, the researcher was experimented on both pruned and unpruned models.

More detail discussion about pruned and its counterpart is given in chapter 2, section 2.5.2.1

In the pruned decision tree two experiments were done: pruned decision tree with default confidence factor and pruned decision tree with confidence factor 0.15.The confidence factor 0.15 was selected after experimenting on many confidence factor values(such as 0.5,0.4,0.35, 0.3,0.28,0.2,0.18) and the model with confidence factor 0.15 has slightly better performance than the other in addition, as confidence factor decrease more pruning incurred(see table 4.1), hence complexity of the tree decreases.

4.1.2.3 Modeling unpruned decision tree

Unpruned decision tree model is built by setting the 'unpruned' parameter of J48 to 'True'.

Here the researcher tried to carry out two experiments: unpruned decision tree with default confidence factor and unpruned decision tree with confidence factor 0.15. They were built to see if there might be interesting rule generated with acceptable accuracy and complexity; They were compared also, if they have advantage with that of their pruned counter parts.

For practical implementations of the above four decision tree experiments, weka has J48 classifier implementation as one option. Figure 4.2, shows the visual graphical explorer window of weka .

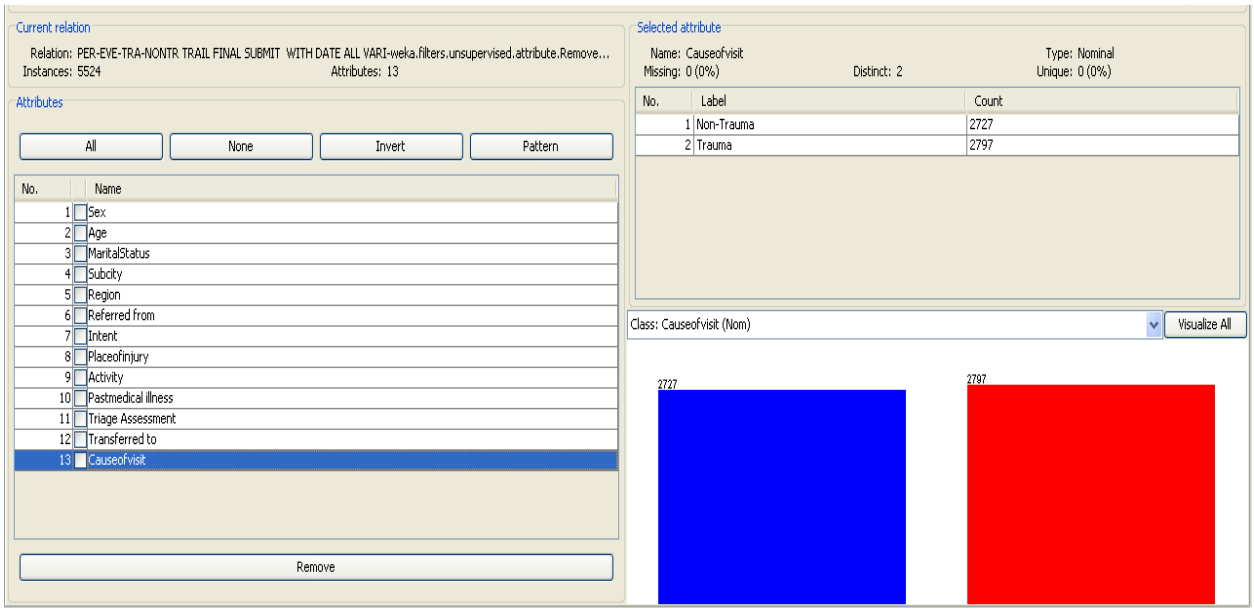


Fig. 4.2: Weka explorer window

Weka generic object editor window enables us to modify options or parameters that enable us to build different decision tree experiments or scenarios. Figure 4.3, shows generic object editor window.

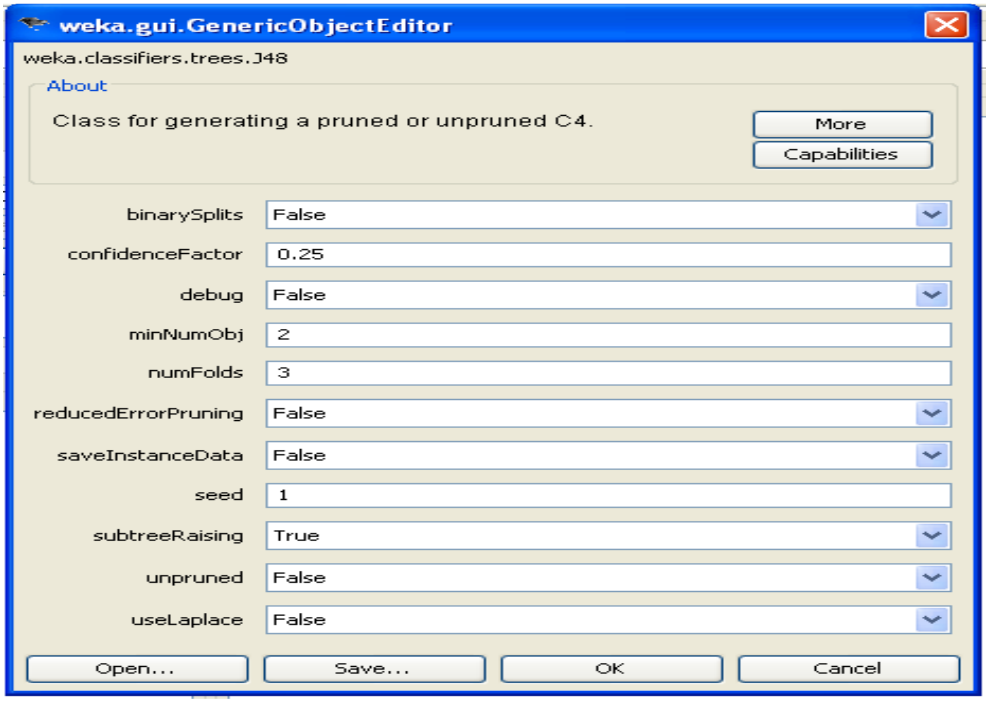


Fig. 4.3: Weka generic object editor window for J48 classifier.

Table 4.1 describes the J48 classifier parameter options found in figure 4.3.

Parameter(Option)	**Description**
binary splits	Whether to use binary splits on nominal attributes when building the trees.
confidence factor	The confidence factor used for pruning (smaller values incur more pruning).
debug	If set to true, classifier may output additional info to the console.
minNumObj	The minimum number of instances per leaf.
numFolds	Determines the amount of data used for reduced error pruning. One fold is used for pruning, the rest for growing the tree.
reduced error pruning	Whether reduced-error pruning is used instead of C.4.5 pruning.
save instance data	Whether to save the training data for visualization.
seed	The seed used for randomizing the data when reduced-error pruning is used.
sub tree raising	Whether to consider the subtree raising operation when pruning.
unpruned	Whether pruning is performed
use laplace	Whether counts at leaves are smoothed based on laplace.

Table 4.1: Description of J48 classifier parameter options in Weka

If we need to build decision tree with different confidence factor we can change the parameter value of 'confidence factor'.

'Unpruned' parameter (the most basic) is also another parameter which is used in this research.It has 'True' and 'False' value. Making the value of 'unpruned' to 'True' enable to tell the classifier we don't want to prune the tree where as setting it to 'False' is to mean we need to prune.

4.1.3 Building association rule model

Another data mining technique to be used for this research is Association rules . There are different algorithms found that are used for implementing association rule mining. Apriori algorithm is the most commonly used association rule mining algorithm which is also applied in this research.

General information regarding association rule mining and how it finds out the best rules based on the interestingness measures (minimum support and minimum confidence thresholds) is given in chapter two sections 2.5.3.

For this research the minimum support starts at 0.3 and experimented until 0.1 with varying the minimum confidence thresholds from 50% to 100%. The actual experiment and analysis of association rules presented is presented later. These different ranges of support and confidence were tested and used because they generally produce similar acceptable patterns to the problem domain.

Figure 4.4. Shows the weka window that used to change parameters such as minimum support (lower bound min support), confidence thresholds (minMetric) and so on for association rule.

Fig. 4.4: Weka generic object editor for Apriori parameters setting

The meanings of some of the parameters on this window are presented in table 4.2 .

NumRules	Number of rules that can be displayed as output or the number of rule required
MinMetric	Confidence
Delta	The delta at which the minimum support is decreased at each iteration
UpperbondMinSupport	The upper bound for minimum support
LowerBondMinSupport	The Lower bound for minimum support

Table 4.2: Meanings of the Apriori parameters for association rule

4.2. Experimentation and Analysis of Classification models

4.2.1 Experimentation and analysis using decision tree

Analysis of the decision tree models were made in terms of detailed accuracy of the classifier based on a confusion matrix of each model resulted.

The confusion matrix is a useful tool for analyzing how well our classifier can recognize tuples of different classes (Non-trauma and Trauma classes in the case of this research) or it is the base for calculating accuracy measures (Correctly Classified Instances and Incorrectly classified Instances) .

The experiments for decision tree classification models are listed below:

Experiment 1: Unpruned decision tree with default confidence factor
Experiment 2: Pruned decision tree with confidence factor 0.15
Experiment 3: Pruned decision tree with default confidence factor
Experiment 4: Unpruned decision tree with confidence factor 0.15

These experiments were analyzed to compare them in terms of different performance matrices values, accuracies, number of leaves, and size of tree generated, ROC curve and execution time. The models were also compared with regard to the patterns or knowledge discovered.

Let us see each of the experiments

Experiment 1: **Unpruned decision tree with default confidence factor**

The result of this model is as follows;

Test mode: 10-fold cross-validation
=== Classifier model (full training set) ===

J48 unpruned tree
Number of Leaves: 882
Size of the tree: 744
Time taken to build model: 0.31seconds

=== Summary ===

Correctly Classified Instances	5137	92.9942 %
Incorrectly Classified Instances	387	7.0058 %
Total Number of Instances	5524	

The detailed accuracy measure for this experiment is as follows:

=== Detailed Accuracy by Class ===

Class	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area
Non-Trauma	0.953	0.093	0.909	0.953	0.931	0.966
Trauma	0.907	0.047	0.952	0.907	0.929	0.966
Weighted Avg.	0.93	0.069	0.931	0.93	0.93	0.966

To understand more about the above detail accuracy measure for all experiments and their confusion matrix let us consider the two by two confusion matrix in the table 4.3.

	Predicted class	
Actual class	True Positives(TP)	False Negative(FN)
	False Positives(FP)	True Negative(TN)

Table 4.3 Confusion matrix

From the table 4.3 confusion matrix we can say the following:

True positives refer to the positive tuples that were correctly labeled by the classifier. True negatives are the negative tuples that were correctly labeled by the classifier. False positives are the negative tuples that were incorrectly labeled (tuples which are actually incorrect or false but the classifier predicted as correct). Similarly, false negatives are the positive tuples that were incorrectly labeled (tuples which are actually correct but the classifier predicted incorrectly).

True Positive Rate (Sensitivity) or Recall for TRUE Class = TP/TP+FN

True Negative Rate (Specificity) or Recall for FALSE Class=TN/FP+TN.

False Positive Rate for TRUE Class = FP/TP+FN

False Positive Rate for FALSE Class = FN/FP+TN.

Precision for TRUE Class = TP/TP+FP, Precision for False Class = FN/FN+TN.

Correctly classified instance=TP+TN and Incorrectly classified instance=FP+FN.

Now based on the above formula we can easily get the value for each parameter in the experiments confusion matrix and their detail accuracy measures as well. The confusion matrix for this experiment is presented as follows:

```
=== Confusion Matrix ===
   a     b   <-- classified as
2599   128 |   a = Non-Trauma
 259  2538 |   b = Trauma
```

Performance analysis for the result of experiment 1 model and its confusion matrix:

The confusion matrix of experiment one depicts that the number of true positives (TP) were 2599 instances which are classified correctly in the class of Non-trauma ,means those instances which were predicted as 'True' class by the classifier and also happened actually true(Non-trauma) when tested on the test data.

The number of the instances which were classified to "False' class by the classifier and they are actually False(TN) as tested on the test data is 2538 or which are classified correctly in the class of Trauma. The sum of TP (2599) and TN (2538) gives us correctly classified instances. Therefore, the total number of instances which were correctly classified to true and false classes of patient cause of visit in the emergency unit was 5137(92.9942 %). On the other hand, 128 records were incorrectly classified as Trauma while actually they were supposed to be in the Non-trauma class(FN) and 259 records were classified incorrectly as Non-Trauma while actually they were in Trauma class(FP). This represents that from the total 5524 records 387 were classified incorrectly.

These misclassification values were obtained due to the classifier way of classifying records. It classify records based on information gain value and attribute similarity that means if most of the attributes have similar value ,specially the attributes with higher information gain then the classifier classify the records in the same class even if they differ with certain attribute value. However, this classification strategy do not work correctly for all instances because the attribute that has a potential to determine the class in real scenario(actual classification) can be ignored or underestimated. The number of leaves for the experiment 1 model was 882 and its tree size was 744 .These values were larger than the other three models this is due to unpruned(set to true) and confidence factor (set to 0.25) parameters. These parameter values have a capability of increasing tree size and number of leaves.

The detail accuracy measure for experiment 1 model is as follows:

True Positive Rate (Sensitivity) or Recall for TRUE Class = TP/TP+FN

=>2599/2599+128=0.953 (95%)

True Negative Rate (Specificity) or Recall for FALSE Class=TN/FP+TN

=>2538/259+2538=0.907(91%)

False Positive Rate for TRUE Class = FP/TP+FN =>259/2599+128=0.093

False Positive Rate for FALSE Class = FN/FP+TN=>128/259+2538=0.047

Precision for TRUE Class = TP/TP+FP=>2599/2599+259=0.909

Precision for False Class = FN/FN+TN=>128/128+2538=0.952

Table 4.4 shows the overall summary of experiment1 model results.

Pruning	Tree Size	No.of leaves	Time (sec.)	AVG TPR	AVG FPR	Precision	Recall	Sensitivity	Specificity	ROC Area	CCI
No	882	744	0.31	0.93	0.069	0.931	0.93	0.953	0.907	0.966	92.99%

Table 4.4: Performance Summary of experiment 1

Where CCI is to mean Correctly Classified Instances.

The summary comparison of all decision tree experimental models is given in table 4.8.

Experiment 2: Pruned decision tree with confidence factor 0.15

The result of this model is as follows:

```
Test mode:    10-fold cross-validation
=== Classifier model (full training set) ===
J48 pruned tree

Number of Leaves: 35
Size of the tree: 43
Time taken to build model: 0.35 seconds
               === Summary ===
Correctly Classified Instances        5154               93.302 %
Incorrectly Classified Instances      370                6.698 %
Total Number of Instances             5524
```

The confusion matrix for this experiment is as follow:

```
=== Confusion Matrix ===
   a       b   <-- classified as
 2653    74|   a = Non-Trauma
 296    2501|   b = Trauma
```

From this confusion matrix we can see the correctly classified instances were 2653 + 2501 which is 5154(93.302 %). On the other hand, 74 records were incorrectly classified as Trauma while actually they were supposed to be in the Non-trauma class(FN) and 296 records were classified incorrectly as Non-Trauma while actually they were in Trauma class(FP). This represents that from the total 5524 records 370 were classified incorrectly. In this experiment the correctly classified instances were slightly increased (93.302%) when we compare to experiment one and four this is due to pruning.

This experimental model is built by making the unpruned parameter value false, which makes the tree less complex and increase in the accuracy. The misclassification values (370) were obtained due to the classifier way of classifying records as explained in experiment 1.

Table 4.5 shows the overall summary of experiment2 model results.

Pruning	Tree Size	No.of leaves	Time (sec.)	AVG TPR	AVG FPR	Precision	Recall	Sensitivity	Specificity	ROC Area	CCI
Yes	43	35	0.35	0.933	0.066	0.936	0.933	0.973	0.894	0.945	93.302%

Table 4.5: Performance summary of experiment 2

If we want to calculate the detail of the accuracy measures for the other experiments we can follow similar operations as in experiment#1.

Experiment 3: Pruned decision tree with default confidence factor

The researcher selected this experimental model because it has slightly better performance and accuracy measure (with specificity,ROC parameters) than other experiments. The model of this experiment is presented in the annex section of the report to avoid complexity here.

The result of this model is as follows:

```
Test mode:    10-fold cross-validation
=== Classifier model (full training set) ===
J48 pruned tree

Number of Leaves: 107
Size of the tree: 135
Time taken to build model: 0.22 seconds
              === Summary ===
Correctly Classified Instances        5154               93.3 %
Incorrectly Classified Instances       370                6.69 %
Total Number of Instances              5524
```

The confusion matrix for this experiment is as follows:

```
=== Confusion Matrix ===
   a     b   <-- classified as
 2596   131 |   a = Non-Trauma
  239  2558 |   b = Trauma
```

From this confusion matrix, 2596 + 2558 records which is 5154(93.3 %) were correctly classified. On the other hand, 370(6.69%) records were classified incorrectly by the classifier. The misclassification values (370) were obtained due to the classifier way of classifying records as explained in experiment 1.

Table 4.6 shows the overall summary of experiment 3 model results.

Pruning	Tree Size	No.of leaves	Time (sec.)	AVG TPR	AVG FPR	Precision	Recall	Sensitivity	Specificity	ROC Area	CCI
Yes	135	107	0.22	0.933	0.067	0.934	0.933	0.952	0.915	0.968	93.3%

Table 4.6: Performance summary of experiment 3

From the table 4.6 we can see some of the accuracy measures such as specificity and ROC Area are slightly better when we compare to all the other experiments. The tree size and number of leaves also moderate with better patterns.

The ROC area of the model is above 0.5(which is the minimum possible acceptable value for ROC curve). It is plotted from True positive Rate (TPR) on the y-axis against the False Positive Rate (FPR) on the x-axis. The ROC area of the model is given in the figure 4.5.

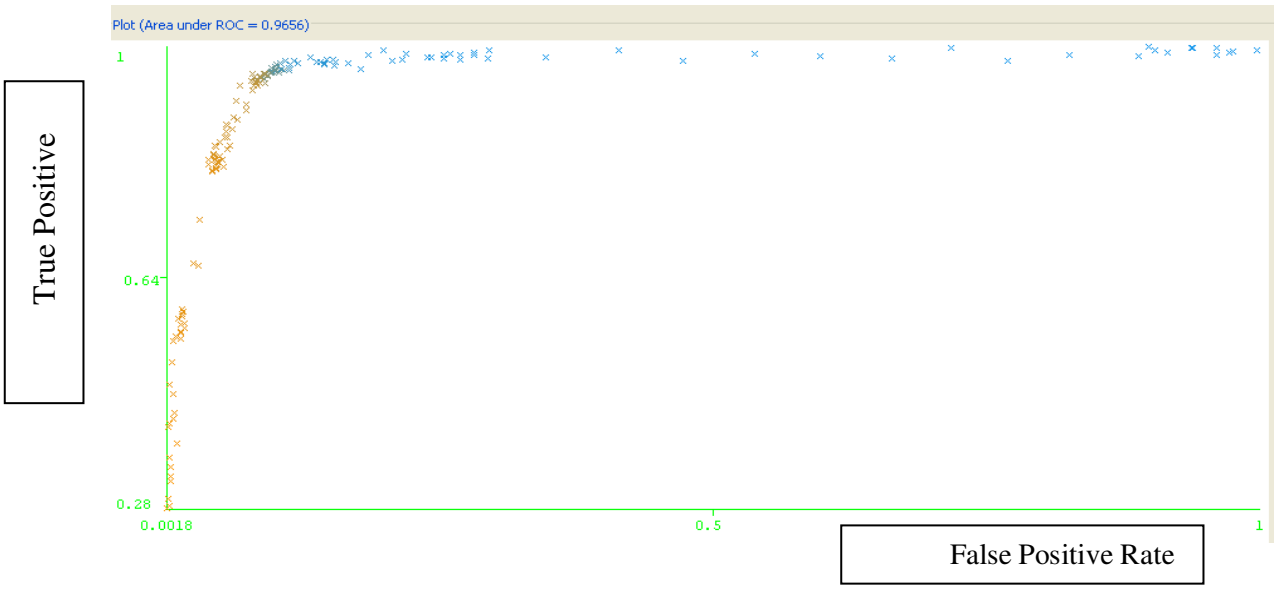


Fig. 4.5: ROC Area curve for experiment 3

As we can see from the ROC area of True class, the curve is above the diagonal line.

Experiment 4: Unpruned decision tree with confidence factor 0.15

The result of this model is as follows;

```
Test mode:    10-fold cross-validation
=== Classifier model (full training set) ===
J48 pruned tree
Number of Leaves: 644
Size of the tree: 782
Time taken to build model: 0.28 seconds
               === Summary ===
Correctly Classified Instances        5120               92.6865 %
Incorrectly Classified Instances       404                7.3135 %
Total Number of Instances             5524
```

The confusion matrix for this experiment is as follows:

```
=== Confusion Matrix ===
   a     b   <-- classified as
 2542  185 |   a = Non-Trauma
 219  2578 |   b = Trauma
```

From this confusion matrix, 5120(92.68 %) were correctly classified. On the other hand, 404(7.13%) records were classified incorrectly by the classifier. The misclassification values (404) were obtained due to the classifier way of classifying records as explained in experiment 1.

Table 4.7 shows the overall summary of experiment 4 model results.

Pruning	Tree Size	No.of leaves	Time (sec.)	AVG TPR	AVG FPR	Precision	Recall	Sensitivity	Specificity	ROC Area	CCI
No	722	644	0.28	0.927	0.073	0.927	0.927	0.932	0.92 2	0.944	92.68%

Table 4.7: performance summary of experiment 4

This model has better accuracy measure than experimental model one and two such as in terms of CCI and Specificity but has lower accuracy measure than experimental model three.

Table 4.8 summarizes the accuracy and performance of all experimental models.

Experiment	Pruning	Conf. factor	Tree Size	No.of leaves	Time (sec.)	AVG TPR	AVG FPR	Precision	Recall	Sensitivity	Specificity	ROC Area	CCI In %
1	No	0.25	882	744	0.31	0.93	0.069	0.931	0.93	0.953	0.907	0.966	92.99
2	Yes	0.15	43	35	0.3	0.933	0.066	0.936	0.933	0.973	0.894	0.945	93.302
3	Yes	0.25	135	107	0.2	0.93	0.067	0.934	0.93	0.952	0.915	0.968	93.3
4	No	0.15	722	644	0.28	0.927	0.073	0.927	0.927	0.932	0.922	0.944	92.68

Table 4.8: summary of all models measure of performance and accuracy

Comparison summary of the experimental models

The accuracy measures of all models are very close to each other for instance, in terms of TPR or recall the first three models are the same.In terms of precision experimental model two is slightly better than the others. In terms of ROC area experimental model three is slightly better than the others, this model is selected because, its CCI, ROC area, and precision is good. Its tree size and number of leaves is moderate and the rules generated were better than others in addition it is the model built with all default parameters setting. Generally it is better to stick with a default parameters setting model, if there is no significant change with the other models to be save from complexity of tree size, number of leaves, execution time, accuracy measures and other issues. For instance if we unprun, mostly the tree size, number of leaves and accuracy measures of the model goes decreasing.

4.2.2 Experimentation and analysis using rule induction

PART algorithm generates rules in plain text form, which is simple to understand. The model of this experiment has an accuracy of 93.12%.

The result of this Model is as follows;

```
Time taken to build model: 2.5 seconds
=== Summary ===
Correctly Classified Instances        5144               93.1209 %
Incorrectly Classified Instances       380                6.8791 %
                Total Number of Instances                5524
```

The detailed accuracy measure for this experiment is as follow:

=== Detailed Accuracy by Class ===

Class	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area
Non-Trauma	0.967	0.104	0.901	0.967	0.933	0.943
Trauma	0.896	0.033	0.965	0.896	0.93	0.943
Weighted Avg.	0.931	0.068	0.934	0.931	0.931	0.943

The confusion matrix for this experiment is as follow:

```
=== Confusion Matrix ===
   a     b   <-- classified as
 2637   90 |   a = Non-Trauma
  290 2507 |   b = Trauma
```

From this confusion matrix, 5144(93.12 %) were correctly classified. On the other hand, 380(6.87%) records were classified incorrectly by the classifier. The calculation of the detail accuracy measure is similar with J48 models.

4.2.3Analysis and interpretation of classification model rules

The following are some of the rules or patterns which were discovered between the class label attribute (cause of visit) and other independent attributes from J48 classification models.

If Place of injury is Street and Intent is accidental then the Cause of visit is likely to be for 'Trauma' (the patient will come to the Hospital emergency unit due to Trauma case).

If Place of injury is Street and intent is violence and Activity is working and whose Region is Addis Ababa then the Cause of visit is likely to be for 'Trauma'.

If Place of injury(illness) is Home and Intent is accidental and Transferred to the Emergency Out Patient Department (OPD) and Activity is Playing then the Cause of visit is likely to be for 'Non-Trauma'.

If Place of injury is Home and Intent is violence then the Cause of visit is likely to be for 'Non-Trauma'.

If Place of injury is Home and intent is accidental and transferred to the emergency OPD and activity is working then the Cause of visit is likely to be for ‘Non-Trauma’.

If Place of injury is working area (work) and intent is accidental and activity is working then the Cause of visit is likely to be for ‘Non-Trauma’.

If Place of injury is other and intent is accidental then the Cause of visit is likely to be for ‘Trauma’.

If Place of injury is other and Intent is violence and Past medical illness is No then the Cause of visit is likely to be for ‘Trauma’.

If Place of injury is recreational place then the cause of visit is highly likely to be predicted as ‘Trauma’.

If Place of injury is school and activity is studying then the Cause of visit is likely to be for ‘Non-Trauma’.

If Place of injury is working area and Activity is other and intent is accidental and transferred to emergency OPD and Past medical illness is no and whose Subcity is Arada then the Cause of visit is likely to be for ‘Non -Trauma’.

If Place of injury is work and activity is other and intent is accidental and transferred to Resuscitation(recovery) room then the Cause of visit is likely to be for ‘Non -Trauma’.

If Place of injury is school and Activity is playing or recreating and marital status is married then the Cause of visit is likely to be for ‘Non- Trauma’.

If Place of injury is Home and Intent is accidental and Transferred to the Emergency OPD and activity is other and Triage assessment is orange and whose Age is between 13

and 25 years, and whose Region is Addis Ababa and who are referred from Governmental Health Facility, then, Cause of visit is likely to be for 'Trauma' .

If Place of injury is Home and intent is accidental and transferred to the emergency OPD and activity is other and triage assessment is orange and whose Age is between 25 and 37 and Past medical illness is no and whose sex is female then the Cause of visit is likely to be for 'Trauma'.

The following are some of the rules or patterns which were generated by PART algorithm.

Place of injury = Street and Intent = accidental: Trauma

Place of injury = school and Activity =studying: Non-Trauma

Place of injury = street and Intent = violence and Activity = recreating: Trauma

Place of injury = recreational place: Trauma

Place of injury = street and Intent = violence and Activity = travelling: Trauma

Most of the rules that obtained from PART algorithm is similar with that of the rules obtained from J48 models. Such as Place of injury = Street and Intent = accidental: Trauma, and Place of injury = school and Activity =studying: Non-Trauma .

When we compare the accuracy measure and the results obtained from both classification algorithm(J48 and PART) models , we can say they are nearly similar. In terms of accuracy, execution time and ROC, J48 is slightly better than PART. The rules generated using both algorithms have similar patterns. We can select J48 model because it has slightly better performance and accuracy measure than PART classification model.

4.2.4 Discussion on classification models generated rules

From the generated rules it is observed that attributes such as place of injury, activity, intent and past medical illness are the basis for the classification.

The following discussion or explanation on the generated rules was made with the domain experts.

Some of the rules shows known pattern as the domain experts opinion (the rules agrees with domain experts view) .If place of injury is street and intent is accidental then it is obvious or common that the cause of visit is most likely to be 'Trauma' .Because most trauma case is occurred accidental, specially on the street due to road traffic accident.

To add another known rule, if place of injury is recreational place(such as bar, grocery) then it is common that the cause of visit is likely to be 'Trauma'. Because commonly in recreational place there are a lot of factors that leads to trauma case such as fighting.

It also known, if place of injury is school and activity is studying then it is common that the cause of visit is likely to be 'Non-Trauma'. Because there are students who have past medical illness, some of the illness (i.e endocrine problem such as sudden collapse and loss of conscious) can be revived due to stress and some unknown factors while studying.

On the other hand an interesting rule shows, patient's whose activity is playing at home encounters non-trauma case. This case might be due to some pervious medical problem(s) that the patient have, these medical problem might be triggered/revived by some recreational activities.

Another interesting rule, shows female patients whose age are between 25 and 37 ; whose triage assessment is orange and intent is accidental at home, encounters trauma case. This case might be due to some factors that might be related to blunt object (such as knife), fire while cooking or other activity and obstetrics or gynecology related problems or other factors .
Generally it is possible to say that the some of the rules obtained from the model provide a pattern or knowledge and have got meaningful contributions for exploring patients' cause of visit in the Hospital emergency unit and these findings also got acceptance by the domain experts.

4.3 Experimentation and Analysis of Association models

Another data mining technique to be experimented in this research was association rule mining. It is carried out to look the data set in different angle because, it can support to extract patterns which were not obtained by classification models. The researcher tried to investigate various rules by altering the support and confidence .Some combinations did not provide even a single rule such as in the case of support 0.4 (with confidence 90%)and support >=0.5 with any confidence .As a result, the researcher selected those scenarios that provide rules and hence the association rule models were built at minimum support of 0.3 to 0.1 and with different confidence varying from 50% to 100% thresholds.

The Table 4.9, shows the number of experiments to be carried out with their thresholds.

Experiment	Minimum Thresholds	
	Support	Confidence
1	0.3	50%
2	0.3	70%
3	0.25	75%
4	0.25	80%
5	0.2	75%
6	0.2	85%
7	0.2	90%
8	0.1	80%
9	0.1	90%
10	0.1	100%

Table 4.9: Association rule mining experiments

Experiment 1: Minimum support of 0.3 and minimum confidence of 50%

To be consistent and to have good flow, the researcher first would present all the experiments starting from the first and then later, when all experimentation is over the meaning of the rules will be given for those selected rules at the end of the experimentation.

The model obtained from the 1st experiment is as follows:

```
        Apriori
        =======
Minimum support: 0.4 (2210 instances)
Minimum metric <confidence>: 0.5
Number of cycles performed: 12
Generated sets of large itemsets:
Size of set of large itemsets L(1): 11
Size of set of large itemsets L(2): 5
Best rules found:
 1. Cause of visit=Trauma 2797 ==> Intent=accidental 2351    conf :(0.84)
 2. Past medical illness=No 3242 ==> Sex=Male 2267    conf:(0.7)
 3. Intent=accidental 3396 ==> Cause of visit=Trauma 2351    conf:(0.69)
 4. Intent=accidental 3396 ==> Sex=Male 2323    conf:(0.68)
 5. Intent=accidental 3396 ==> Region=Addis Ababa 2323    conf:(0.68)
 6. Sex=Male 3573 ==> Region=Addis Ababa 2380    conf :(0.67)
 7. Sex=Male 3573 ==> Intent=accidental 2323    conf :(0.65)
 8. Sex=Male 3573 ==> Past medical illness=No 2267    conf:(0.63)
 9. Region=Addis Ababa 3829 ==> Sex=Male 2380    conf:(0.62)
10. Region=Addis Ababa 3829 ==> Intent=accidental 2323    conf:(0.61)
```

Where the number inside the parenthesis show confidence (conf)level. Confidence(conf) refers the proportion(percentage) of both Consequent (the right hand side of the arrow)and Antecedent (the Left hand side of the arrow)occurring together to the occurrence of the antecedent. Which means the number on the right side of the arrow /Left side number. For instance the confidence for rule #3 is 2351/3396=0.69. Support refers to the proportion (percentage)of number of times the antecedent and consequent occurred together divided by the total number of dataset.

For instance consider rule #3 the support is 2351 /5524=0.4 or 40%.

Support and confidence issues were discussed in chapter 2, section 2.5.1.3. You can refer to it for more information. As we see from the above model ten sample association rule results were obtained , from these rules only one rule indicates patterns in explaining the association that exists between intent variable as antecedent and Cause of visit as consequent at minimum support of 0.3 and confidence of 50% .

Experiment 2: Minimum support of 0.3 and minimum confidence of 70%

From this experiment we can see all the top ten rules were different and newer than that of the first experiment. The rules were sorted by their confidence.

The model obtained from this experiment is as follows:

Apriori

=======

Minimum support: 0.3 (1657 instances)

Minimum metric <confidence>: 0.7

Number of cycles performed: 14

Generated sets of large itemsets:

Size of set of large itemsets L(1): 17

Size of set of large itemsets L(2): 28

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

Best rules found:

1. Intent=accidental Placeofinjury=Street 1696 ==> Causeofvisit=Trauma 1673 conf:(0.99)
2. Placeofinjury=Street 1859 ==> Causeofvisit=Trauma 1822 conf:(0.98)
3. Age='[13-25]' 2072 ==> MaritalStatus=Single 1995 conf:(0.96)
4. Placeofinjury=Street Causeofvisit=Trauma 1822 ==> Intent=accidental 1673 conf:(0.92)
5. Placeofinjury=Street 1859 ==> Intent=accidental 1696 conf:(0.91)
6. Placeofinjury=Street 1859 ==> Intent=accidental Causeofvisit=Trauma 1673 conf:(0.9)
7. Placeofinjury=Home 2210 ==> Causeofvisit=Non-Trauma 1976 conf:(0.89)
8. Causeofvisit=Trauma 2797 ==> Intent=accidental 2351 conf:(0.84)
9. Sex=Male Causeofvisit=Trauma 2068 ==> Intent=accidental 1709 conf:(0.83)
10. MaritalStatus=Single 2569 ==> Age='(13-25]' 1995 conf:(0.78)

Experiment 3: Minimum support of 0.25 and minimum confidence of 75%

The model obtained from this experiment is as follows:

```
Apriori
=======
Minimum support: 0.3 (1657 instances)
Minimum metric <confidence>: 0.75
Number of cycles performed: 14
Generated sets of large itemsets:
Size of set of large itemsets L(1): 17
Size of set of large itemsets L(2): 28
Size of set of large itemsets L(3): 2
Best rules found:
1. Intent=accidental Placeofinjury=Street  1696 ==> Causeofvisit=Trauma 1673    conf:(0.99)
 2. Placeofinjury=Street  1859 ==> Causeofvisit=Trauma 1822    conf:(0.98)
 3. Age='[13-25]' 2072 ==> MaritalStatus=Single 1995    conf:(0.96)
4. Placeofinjury=Street  Causeofvisit=Trauma 1822 ==> Intent=accidental 1673    conf:(0.92)
5. Placeofinjury=Street  1859 ==> Intent=accidental 1696    conf:(0.91)
6. Placeofinjury=Street  1859 ==> Intent=accidental Causeofvisit=Trauma 1673    conf:(0.9)
7. Placeofinjury=Home 2210 ==> Causeofvisit=Non-Trauma 1976    conf:(0.89)
 8. Causeofvisit=Trauma 2797 ==> Intent=accidental 2351    conf:(0.84)
 9. Sex=Male Causeofvisit=Trauma 2068 ==> Intent=accidental 1709    conf:(0.83)
10. MaritalStatus=Single 2569 ==> Age='[13-25]' 1995    conf:(0.78)
```

In this experiment ten rules were obtained even if the support and the confidence differ from the rules obtained they were similar with that of experiment two.

Experiment 4: Minimum support of 0.25 and minimum confidence of 80%

As we observe in this model the minimum support threshold doesn't change i.e. it is 25% for the generated rules. However, the confidence provided by the model is greater than the supplies minimum confidence threshold, which is 99% to 97% for the rules generated. This means the proportion of the number of occurrence of antecedent and consequent together to the number of occurrence of antecedent is higher than the given minimum threshold.

The model obtained from this experiment is as follows:

Apriori

=======

Minimum support: 0.25 (1381 instances)

Minimum metric <confidence>: 0.8

Number of cycles performed: 15

Generated sets of large itemsets:

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

Size of set of large itemsets L(2): 51

Size of set of large itemsets L(3): 13

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

Best rules found:

1. Intent=accidental Activity=Travelling 1497 ==> Causeofvisit=Trauma 1479 conf:(0.99)
2. Intent=accidental Placeofinjury=Street Activity=Travelling 1455 ==> Causeofvisit=Trauma 1437 conf:(0.99)
3. Intent=accidental Placeofinjury=Street 1696 ==> Causeofvisit=Trauma 1673 conf:(0.99)
4. Placeofinjury=Street Activity=Travelling 1560 ==> Causeofvisit=Trauma 1538 conf:(0.99)
5. Activity=Travelling 1609 ==> Causeofvisit=Trauma 1586 conf:(0.99)
6. Placeofinjury=Street 1859 ==> Causeofvisit=Trauma 1822 conf:(0.98)
7. Intent=accidental Activity=Travelling 1497 ==> Placeofinjury=Street 1455 conf:(0.97)
8. Intent=accidental Activity=Travelling Causeofvisit=Trauma 1479 ==> Placeofinjury=Street 1437 conf:(0.97)
9. Activity=Travelling Causeofvisit=Trauma 1586 ==> Placeofinjury=Street 1538 conf:(0.97)
10. Activity=Travelling 1609 ==> Placeofinjury=Street 1560 conf:(0.97)

In this experiment, all most all (eight) rules were new when we compared to the above models.In addition, the number of important rules that contain the dependent variable as consequent with higher confidence was also increased.

Experiment 5: Minimum support of 0.2 and minimum confidence of 75%

The model obtained from this experiment is as follows:

```
Apriori
=======
Minimum support: 0.3 (1657 instances)
Minimum metric <confidence>: 0.75
Number of cycles performed: 14
Generated sets of large itemsets:
Size of set of large itemsets L(1): 17
Size of set of large itemsets L(2): 28
Size of set of large itemsets L(3): 2
Best rules found:
 1. Intent=accidental Placeofinjury=Street  1696 ==> Causeofvisit=Trauma 1673    conf:(0.99)
 2. Placeofinjury=Street  1859 ==> Causeofvisit=Trauma 1822    conf:(0.98)
 3. Age='(-inf-25]' 2072 ==> MaritalStatus=Single 1995    conf:(0.96)
 4. Placeofinjury=Street  Causeofvisit=Trauma 1822 ==> Intent=accidental 1673    conf:(0.92)
 5. Placeofinjury=Street  1859 ==> Intent=accidental 1696    conf:(0.91)
 6. Placeofinjury=Street  1859 ==> Intent=accidental Causeofvisit=Trauma 1673    conf:(0.9)
 7. Placeofinjury=Home 2210 ==> Causeofvisit=Non-Trauma 1976    conf:(0.89)
 8. Causeofvisit=Trauma 2797 ==> Intent=accidental 2351    conf:(0.84)
 9. Sex=Male Causeofvisit=Trauma 2068 ==> Intent=accidental 1709    conf:(0.83)
10. MaritalStatus=Single 2569 ==> Age='(-inf-25]' 1995    conf:(0.78)
```

In this experiment ,most of the rules were new when we compared them to the above models. There were a total of 91 rules generated out of which the top ten of them were presented above in this experiment.

Experiment 6: Minimum support of 0.2 and minimum confidence of 85%

The model obtained from this experiment is as follows

```
Apriori
=======
Minimum support: 0.25 (1381 instances)
Minimum metric <confidence>: 0.85
Number of cycles performed: 15
Generated sets of large itemsets:
Size of set of large itemsets L(1): 21
Size of set of large itemsets L(2): 51
Size of set of large itemsets L(3): 13
Size of set of large itemsets L(4): 1
Best rules found:
 1. Intent=accidental Activity=Travelling 1497 ==> Causeofvisit=Trauma 1479    conf:(0.99)
 2. Intent=accidental Placeofinjury=Street  Activity=Travelling 1455 ==> Causeofvisit=Trauma
    1437    conf:(0.99)
 3. Intent=accidental Placeofinjury=Street  1696 ==> Causeofvisit=Trauma 1673    conf:(0.99)
 4. Placeofinjury=Street  Activity=Travelling 1560 ==> Causeofvisit=Trauma 1538    conf:(0.99)
 5. Activity=Travelling 1609 ==> Causeofvisit=Trauma 1586    conf:(0.99)
 6. Placeofinjury=Street  1859 ==> Causeofvisit=Trauma 1822    conf:(0.98)
 7. Intent=accidental Activity=Travelling 1497 ==> Placeofinjury=Street  1455    conf:(0.97)
 8. Intent=accidental Activity=Travelling Causeofvisit=Trauma 1479 ==> Placeofinjury=Street
    1437    conf:(0.97)
```

In this experiment, some of the rules were new when we compared to the above models. There were a total of 21 rules generated out of which the top ten of them were presented above in this experiment.

Experiment 7: Minimum support of 0.2 and minimum confidence of 90%

The model obtained from this experiment is as follows

```
Apriori
=======
Minimum support: 0.25 (1381 instances)
Minimum metric <confidence>: 0.9
Number of cycles performed: 15
Generated sets of large itemsets:
Size of set of large itemsets L(1): 21
Size of set of large itemsets L(2): 51
Size of set of large itemsets L(3): 13
Size of set of large itemsets L(4): 1
Best rules found:
 1. Intent=accidental Activity=Travelling 1497 ==> Causeofvisit=Trauma 1479    conf:(0.99)
 2. Intent=accidental Placeofinjury=Street  Activity=Travelling 1455 ==> Causeofvisit=Trauma
  1437    conf:(0.99)
 3. Intent=accidental Placeofinjury=Street  1696 ==> Causeofvisit=Trauma 1673    conf:(0.99)
 4. Placeofinjury=Street  Activity=Travelling 1560 ==> Causeofvisit=Trauma 1538    conf:(0.99)
 5. Activity=Travelling 1609 ==> Causeofvisit=Trauma 1586    conf:(0.99)
 6. Placeofinjury=Street  1859 ==> Causeofvisit=Trauma 1822    conf:(0.98)
 7. Intent=accidental Activity=Travelling 1497 ==> Placeofinjury=Street  1455    conf:(0.97)
 8. Intent=accidental Activity=Travelling Causeofvisit=Trauma 1479 ==> Placeofinjury=Street
  1437   conf:(0.97)
 9. Activity=Travelling Causeofvisit=Trauma 1586 ==> Placeofinjury=Street  1538    conf:(0.97)
10. Activity=Travelling 1609 ==> Placeofinjury=Street  1560    conf:(0.97)
```

This model presents similar top ten rules with that of the above model.

Experiment 8: Minimum support of 0.1 and minimum confidence of 80%

The model obtained from this experiment is as follows

```
Apriori
=======
Minimum support: 0.25 (1381 instances)
Minimum metric <confidence>: 0.8
Number of cycles performed: 15
Generated sets of large itemsets:
Size of set of large itemsets L(1): 21
Size of set of large itemsets L(2): 51
Size of set of large itemsets L(3): 13
Size of set of large itemsets L(4): 1
Best rules found:
 1. Intent=accidental Activity=Travelling 1497 ==> Causeofvisit=Trauma 1479    conf:(0.99)
 2. Intent=accidental Placeofinjury=Street  Activity=Travelling 1455 ==> Causeofvisit=Trauma
    1437    conf:(0.99)
 3. Intent=accidental Placeofinjury=Street  1696 ==> Causeofvisit=Trauma 1673    conf:(0.99)
 4. Placeofinjury=Street  Activity=Travelling 1560 ==> Causeofvisit=Trauma 1538    conf:(0.99)
 5. Activity=Travelling 1609 ==> Causeofvisit=Trauma 1586    conf:(0.99)
 6. Placeofinjury=Street  1859 ==> Causeofvisit=Trauma 1822    conf:(0.98)
 7. Intent=accidental Activity=Travelling 1497 ==> Placeofinjury=Street  1455    conf:(0.97)
 8. Intent=accidental Activity=Travelling Causeofvisit=Trauma 1479 ==> Placeofinjury=Street
    1437    conf:(0.97)
 9. Activity=Travelling Causeofvisit=Trauma 1586 ==> Placeofinjury=Street  1538    conf:(0.97)
10. Activity=Travelling 1609 ==> Placeofinjury=Street  1560    conf:(0.97)
```

This model presents similar top ten rules with that of the experiment seven and six models.

Experiment 9: Minimum support of 0.1 and minimum confidence of 90%

The model obtained from this experiment is as follows

Apriori

=======

Minimum support: 0.25 (1381 instances)

Minimum metric <confidence>: 0.9

Number of cycles performed: 15

Generated sets of large itemsets:

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

Size of set of large itemsets L(2): 51

Size of set of large itemsets L(3): 13

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

Best rules found:

1. Intent=accidental Activity=Travelling 1497 ==> Causeofvisit=Trauma 1479 conf:(0.99)
2. Intent=accidental Placeofinjury=Street Activity=Travelling 1455 ==> Causeofvisit=Trauma 1437 conf:(0.99)
3. Intent=accidental Placeofinjury=Street 1696 ==> Causeofvisit=Trauma 1673 conf:(0.99)
4. Placeofinjury=Street Activity=Travelling 1560 ==> Causeofvisit=Trauma 1538 conf:(0.99)
5. Activity=Travelling 1609 ==> Causeofvisit=Trauma 1586 conf:(0.99)
6. Placeofinjury=Street 1859 ==> Causeofvisit=Trauma 1822 conf:(0.98)
7. Intent=accidental Activity=Travelling 1497 ==> Placeofinjury=Street 1455 conf:(0.97)
8. Intent=accidental Activity=Travelling Causeofvisit=Trauma 1479 ==> Placeofinjury=Street 1437 conf:(0.97)
9. Activity=Travelling Causeofvisit=Trauma 1586 ==> Placeofinjury=Street 1538 conf:(0.97)
10. Activity=Travelling 1609 ==> Placeofinjury=Street 1560 conf:(0.97)

Experiment 10: Minimum support of 0.1 and minimum confidence of 100%

The model obtained from this experiment is as follows

```
Apriori
=======
Minimum support: 0.1 (552 instances)
Minimum metric <confidence>: 1
Number of cycles performed: 18
Generated sets of large itemsets:
Size of set of large itemsets L(1): 34
Size of set of large itemsets L(2): 199
Size of set of large itemsets L(3): 377
Size of set of large itemsets L(4): 217
Size of set of large itemsets L(5): 56
Size of set of large itemsets L(6): 5
Best rules found:
 1. Intent=Other 1090 ==> Causeofvisit=Non-Trauma 1090    conf:(1)
 2. Region=Oromia 1066 ==> Subcity=Outside Addis Ababa 1066    conf:(1)
 3. Intent=Other Placeofinjury=Home 923 ==> Causeofvisit=Non-Trauma 923    conf:(1)
 4. Intent=Other Activity=Working 834 ==> Causeofvisit=Non-Trauma 834    conf:(1)
 5. Region=Addis Ababa Intent=Other 787 ==> Causeofvisit=Non-Trauma 787    conf:(1)
 6. Sex=Male Region=Oromia 762 ==> Subcity=Outside Addis Ababa 762    conf:(1)
 7. Intent=Other Placeofinjury=Home Activity=Working 757 ==> Causeofvisit=Non-Trauma 757
    conf:(1)
 8. Region=Oromia Referred from=Governmental Health Facility 746 ==> Subcity=Outside Addis
    Ababa 746    conf:(1)
 9. Referred from=Governmental Health Facility Intent=Other 734 ==> Causeofvisit=Non-Trauma
    734    conf:(1)
10. Region=Oromia Pastmedical illness=No 691 ==> Subcity=Outside Addis Ababa 691
    conf:(1)
```

In this experiment most of the rules were new when we compared them to the above models. There were a total of 24 rules generated out of which the top ten of them were presented above in this experiment.

4.2.3 Analysis and interpretation of the association rule model results

The following are sample rules selected from different experiments.

Intent=accidental Place of injury=Street 1696 ==> Cause of visit=Trauma 1673 conf:(0.99)
Meaning: If the injury occurred is due to accidental and place of injury is street,
then, the patient came to or visit the hospital for trauma case ,with confidence of 99% and Support of 30%.
Performance: (30% Support, 99% Confidence).

Place of injury=Home 2210 ==> Cause of visit=Non-Trauma 1976 conf:(0.89)
Meaning: If place of injury is Home then the patient visit the hospital for Non- trauma case.
Performance: (30% Support, 89% Confidence).

Intent=accidental Activity=Travelling 1497 ==> Cause of visit=Trauma 1479 conf:(0.99)
Meaning: If the injury occurred is due to accidental and the activity of patient during injury is travelling then the patient visit the hospital for trauma case
Performance: (25% Support, 99% Confidence).

Intent=Other Place of injury=Home Activity=Working 757 ==> Cause of visit=Non-Trauma 757 conf:(1)
Meaning: If the illness occurred is due to other intent and place of injury is Home and the activity of patient during injury is working, then, the patient visit the hospital for Non-trauma case.
Performance: (10% Support, 100% Confidence).
Intent=accidental Place of injury=Street Activity=Travelling 1455 ==> Cause of visit=Trauma 1437 conf:(0.99).

If the injury occurred is due to accidental and place of injury is street and the activity of patient during injury is travelling ,then, the patient visit the hospital for trauma case

Performance: (25% Support, 99% Confidence).

Activity=Travelling 1609 ==> Causeofvisit=Trauma 1586 conf:(0.99)

Meaning: If the activity of patient during injury is travelling, then, the patient visit the hospital for trauma case.

Performance: (25% Support, 99% Confidence**).**

Referred from=Governmental Health Facility Intent=Other 734

==> causeofvisit=Non-Trauma 734 conf:(1)

Meaning: If the patient is referred from Governmental Health Facility and the illness occurred due to other intent ,then, the patient visit the hospital for Non-trauma case .

Performance: (10% Support, 100% Confidence**).**

Place of injury=Street 1859 ==> Causeofvisit=Trauma 1822 conf:(0.98)

Meaning: If place of injury is street,then, the patient visit the hospital for Trauma case .

Performance: (25% Support, 98% Confidence**).**

Some of the patterns were repeated in different scenarios unless the difference in the number of rules generated, support, and confidence difference. Furthermore, most of patterns were also found in the decision tree models such as, Intent=accidental Place of injury=Street Activity=Travelling ==> cause of visit =Trauma and Place of injury=Street ==> Cause of visit=Trauma.

The explain or analysis of the above generated rules is similar with that of the one provided in classification rules. Similar discussion with domain experts was made.

Some of the rules shows known patterns, if place of injury is street, activity is travelling and intent is accidental as antecedent then the consequent is most likely to be 'Trauma' ,with accepted support and confidence. Because most trauma case is occurred accidental, especially on the street due to road traffic accident.

An interesting pattern shows , if place of injury is home, activity is working and intent is other than violence and accidental as antecedent then the consequent is most likely to be 'Non-Trauma' ,with accepted support and confidence. As the discussion, the cause might be past medical illness which revived due to some activity.

As a summary, representative experiments were done in both decision tree and association rule models with different parameters, from these models the researcher investigate data mining has an applicability to discover knowledge and non-trivial patterns within the dataset . In addition, the patterns were also accepted by the domain experts. To mean some of the patterns were previously known, hence verified by the classifiers and some of the rules were interesting.

The registered accuracy (93.3%) can be improved by iterative data preprocessing activities, altering unused J48 parameters (i.e minNumobj) and employing unused algorithms such as neural nets.

There were challenges encountered while carry out this research which includes, getting the source data were a big challenge , selecting appropriate tables , selecting appropriate attributes, preprocessing the dataset (as it contains inconsistency ,noisy and missing values), J48 parameter selection and setting for modeling, extracting interesting patterns and consulting domain experts(as they are too busy) .

CHAPTER FIVE

CONCLUSIONS AND RECOMMENDATIONS

This chapter deals with conclusion of the whole research tasks and recommendations.

5.1 Conclusions

Health facilities, especially governmental ones such as Hospitals, deliver a lot of clinical services for many patients. They take or store tremendous demographic data as well as clinical history of the patients who are treated at these facilities. Hence, these data grows exponentially from time to time requiring powerful analysis tools for uncovering the hidden patterns and producing decision support information. Therefore, data mining techniques are solutions in discovering hidden and potentially useful information or knowledge that can be used for decision support out of large volume of data collected overtime from various sources.

Among trauma cases especially Injuries have been recognized as one of the most life frightening public health problems. They represent 12% of the global burden of diseases and the third cause of overall mortality. It is estimated 5.06 million people die each year as a result of some form of injuries, comprising almost 9% of all deaths. This equates to almost 14,000 injury deaths every single day.

In many developing countries, not only is the incidence of various injuries increasing but also the causative factors are changing from the historical patterns such as falling from trees to injuries due to occupational hazards, interpersonal violence and road traffic injuries. From these causative factors road traffic injuries appear to be the leading cause of traumatic injuries.
Even thought injuries are increasing as the trend suggested by available data, road traffic injuries in particular have not received the attention they deserve in most developing countries. Lack of empirical data and poor quality of the data that exist is probably part of the problem.
The situation, here, in Ethiopia is also similar like to other developing countries. That is, injuries are common but little attention is being given to this problem.

Injuries constitute half of all surgical emergencies and are the primary reason for an emergency hospital visit in Addis Ababa. Similarly Non-trauma cases also have equivalent trends with that of trauma as the data set shows almost equal number of patients visit the hospital for both cases.

Due to large number of patients in the emergency unit a lot of demographic and clinical data is collected from patients who visit the Hospital to get treatment. As a result, the data collected or stored about these patients at the Hospital may reveal some hidden but important information for decision making regarding to classify those patients based on various determining variables. For the purpose of discovering these hidden knowldge; data mining has high potential. This was the target of this research, i.e. exploring the applicability of data mining in finding out patterns which are used for classifying patients who come to the Hospital's emergency unit.

For conducting the research, data was collected from the Hospital Emergency medical service unit.Various tasks were under taken by the researcher to accomplish this research, such as Data preprocessing in preparing the data for model building , Attribute and models selections, for data mining tasks, based on its appropriateness for the problem domain and experimentations were done.

The data mining methodology used in this research was CIRSP-DM (Cross Industry Standard for Data Mining), which involves business understanding, data understanding, data preparation, model building, evaluation of the models and deployment or report generation phases.

Various literatures such as, Han & Kamber(2006) Data mining: Concepts and Techniques and Applications; Witten & Frank(2005) Data mining: practical machine learning tools and Techniques were reviewed which supports to get insight about conceptual view of data mining and related research works were also reviewed. SPSS were used for statistical summary of the data, weka 3.6.0 version for model building, experimentation and preprocessing tasks such as discretization.

The selected and implemented data mining techniques for this research were classification and association. For classification models, decision tree and rule induction were selected due to their easy structure, visualization power (for decision tree) and understandable rule generation characteristics . five experiments were employed for classification purpose.These are pruned decision tree with default confidence factor , unpruned decision tree with default confidence factor , pruned decision tree with confidence factor 0.15 , unpruned decision tree with confidence factor 0.15 and rule induction experimental model.

The above experiments were performed by using J48 decision tree implementation algorithm which is integrated in weka. It was found that all of the decision tree and rule induction classification experiments were applicable and have slight differences .From these experiments, pruned decision tree with default confidence factor was selected because it is slightly better than others in terms of accuracy measures , performance and significant patterns with moderate tree size. However we might use PART model as well because, they are very close in most aspect.

Fourteen variables were selected and ordered or ranked using ranking algorithm in weka from these variables place of injury has the first position or the most determining variable to the cause of visit.On the contrary triage assessment variable is the least determining variable.

The top four determining variables selected by the algorithm were place of injury, activity, intent and past medical illness in their placement order. These variables, especially the first three, appeared with meaningful patterns in both classification and association rule experiments.

There were ten experiments done for the association technique by varying minimum support and minimum confidence threshold .It was observed that increasing minimum support threshold above or equal to 0.3 and keeping minimum confidence above or equal to 0.9 reduces the number of rules generated even to the extent not having any single best rule.

As we decrease the confidence and increase the minimum support, more rules with meaningful pattern were obtained. Hence by decreasing the minimum support gradually from 0.3 to 0.1 and increasing the minimum confidence from 50% to 90% interesting patterns with acceptable performance (Support and Confidence) were obtained.

These patterns show strong association between the dependent and independent variables that would help to obtain knowledge out of it.

To mention a few patterns:
If the injury occurred due to accidental and place of injury is street, then, there is high chance (with 99% confidence) where the patient came to visit the hospital for trauma case.

If the illness occurred is due to other intent and place of injury is Home and the activity of patient during injury is working, then, with 100% confidence the patient visit the hospital for Non-trauma case .

Finally it is the researcher's belief that the main objective of the research was achieved basically by identifying and applying appropriate data mining techniques ,algorithms which perform well for the problem domain and by selecting the appropriate determining variables from the dataset .Moreover, the researcher tried to show the applicability of data mining on emergency medical dataset by investigating important patterns .

5.2 Recommendations

Based on the findings of this study and the experience obtained from the research, the following recommendations can be forwarded.

Database standardization is required for the Hospital emergency medicine registration database because, there were so many duplicate, invalid, missing and inconsistent values in many of the attributes and records which consume much of the researcher's time to clean or preprocess the data set.

The above problems mainly occurred by the data clerks, due to varied understanding about the database these results, inconsistency during data entry. Moreover, mostly missing and inconsistent values were occurred because of the triage nurses; they do not fill the necessary variables due to overcrowded number of patients, lack of information and other reasons.

Hence, these problems affect researches that need appropriate data mainly like data mining researches. Therefore, further training is recommended about the database for data clerks and providing appropriate awareness is recommended for data clerks and triage nurses ,such as how to handle missing information during recording and how to be consistent during recording.

Getting health related data for data mining research is the main problem encountered even though the ethical considerations are made; this problem should be improved for health informatics and other interested researchers in the future.

There exists only very few health related data on softcopy form (computerized form) in the city. In addition, most hardcopy data are also not convenient (to mean the data recorded on the card is bulky, inconsistent, and it is mostly not understood by non professionals for the area) for data mining research because they are collected and stored for other purposes .Hence, as much as possible this situation need improvement such as using standardized data collection and storing system like using softwares in the country.

For this research, two selected data mining techniques (classification and association rule) were experimented by the researcher. However, those data mining techniques which were not used by the researcher might reveal important patterns in relation to the problem domain. So further studies might be required that apply those unused techniques.

Further study is recommended to the problem domain specifically and emergency medical service unit in general that apply those unused data mining models and algorithms such as neural nets in this research. Hence, these algorithms might discover vital knowledge to the area.

The researcher beliefs the study can be used as a reference for the research works which will be done in this related area in the future, especially those research that apply data mining.

References

Abraham, T.(2005). Application of data mining technology to identify determinant risk factors of HIV infection to find their association rules: The Case of Center For Disease Controls and Prevention (CDC).MSc. Thesis, Addis Ababa University, Ethiopia.

Agrawal, R. & Srikant R. (1994). Fast algorithms for mining association rules. IBM Almaden Research Center .Journal of computer science ,5, 131-135.

Agrawal, R. & Srikant, R. (2000). Privacy preserving data mining. ACM SIGMOD Rec., 29, 439-450.

Al Shalabi, L. (2009). Improving accuracy and coverage of data mining systems that are built from noisy datasets: A new model. Arab Open University, Kuwait.

Azevedo, A. & Santos,M.(2008). KDD, SEMMA and CRISP-DM: a parallel overview.IADIS European conference data mining.

Baeza-Yates, R. & Ribeiro-Neto, B. (1999). Modern information retrieval. Harlow, England: Addison Wesley.

Bellazzi, R. & Zupan, B. (2008). Predictive data mining in clinical medicine: current issues and guidelines. International Journal of Medical Informatics, 77(2):81–97.

Berry, M.,& Linoff, S. (1997). Mastering data mining: The art and science of customer relationship management. New York:Wiley.

Borok, L. (1997). Data mining: Sophisticated forms of managed care modeling through artificial intelligence. J. Health Care Finance, 23, 20–36.

Bounsaythip , C. & Rinta-Runsala, E.(2001). Overview of data mining for customer behavior modeling. VTT Information Technology.

Bramer, M. (2007). Principles of data mining. London: Springer.

Cabena, P., Hadjinian, P., Stadler, R., Verhees, J. & Zanasi, A. (1998). Discovering Data Mining*:* From Concept to Implementation. Upper Saddle River, New Jersey:Prentice Hall.

Carbone, P. (1997). Data Mining or "Knowledge Discovery in Databases": An Overview. Retrieved on April 10, 2011 from http://www.mitre.org/pubs/data_mgt/Papers/dmhdbk.pdf.

Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz ,T. , Shearer, C. & Wirth R. (2000). CRISP-DM 1.0: step-by-step Data mining guide. SPSS.

Daud ,M., & Corne, D. (2007). Human readable rule induction in medical data mining: A Survey of existing algorithms.

Deogun,J. & Jitender S. (2001). Data Mining: research trends, challenges, and applications.

Edelstein, H. (1998). Data mining-Let's get practical: How to identify strategic problem statement, prepare the right data, and build and apply a robust model. Database programming & design Magazine.

Fayyad,U., Piatetsky-shapiro, G. & Smyth , P.(1996). Knowledge Discovery and Data Mining: Towards a unifying framework. American Association for Artificial Intelligence.

Fromm, R., McCallum,W. ,Niziol, C., Babcock, J., Gueler, A. & Levine ,R.(1993). Critical care in the emergency department: a time-based study. Crit. Care Med., 21, 970-976.

Giudici,P.(2003). Applied data mining statistical methods for business and industry. Wiley & Sons Ltd.

Goebel, M. & Gruenwald, L. (1999). A Survey of data mining and knowledge discovery software Tools.

Han, J. & Kamber, M.(2006). Data Mining: Concepts and Techniques and Applications. San Fransisico : Morgan kufman publishers.

Houston, A.(2000). Medical data mining on the Internet: Research on cancer information system: Artificial Intelligence Review ,13,437–466.

Immon, W. (1996) . Building the Data Warehouse. New York:John Wiley & Sons, Inc.

Kaur, H., &Wasan , S.(2006). Empirical study on applications of data mining techniques in healthcare. Journal of Computer Science, 2,194-200.

Larson, P. & Dessie ,T.(1993). Unintentional and intentional injuries: Ecology of health and disease in Ethiopia. Westview press,473-82.

Larvac, N.(1998). Data Mining in Medicine: Selected Techniques and Applications.

Levin,N.&Zahavi,J.(1999). Data mining. Retrieved on may12, 2011 from http://www.urbanscience. com/data_mining.Pdf.

Mary, K. & Mat (2004).Application of data mining techniques to Healthcare data. Statistics for Hospital epidemiology.

Masuda, G., Sakamoto, N., & Yamamoto, R. (2002).A framework for dynamic evidence based medicine using data mining. In Proc. 15th IEEE Symposium on computer- based medical systems, IEEE press, 117-122.

Mitchell, M. (1997). Machine learning. New York: The McGraw-Hill companies, Inc.

Museru ,L. (1999). Injuries in Africa and the need to develop preventive strategies. East Central Africa J. Surg.,5,51-55.

Nilgün,U.& Özgür, D.(2010). Evaluation of risk of death in hepatitis by rule induction algorithms. Scientific research and essays, 5(20), 3059-3062.

Nordberg, E.(1994). Injuries in Africa: A review. East Afr Med J., 7,339-45.

Peden, M., McGee, K. & Sharma, G.(2002).The injury chart book: a graphical overview of the global burden of injuries. Geneva, world health organization.

Prather, J. (2001). Medical Data Mining: Knowledge Discovery in a clinical Data Warehouse.

Piatetsky-Shapiro, G. (1991). Knowledge Discovery in real Databases.

Raghavan, V., Deogun, J. & Sever, H. (1998). Knowledge Discovery and Data Mining: Introduction. Journal of American Society for Information Science, 49(5).

Razzak, A.& Kellerman, J. (2002).Emergency care in developing countries: Is it worth while ? Bull world health organization, 80(11),900-905.

Ruben, D. & Canlas, J. (2009).Data mining in Healthcare: Current applications and issues.

Saarenvirta, G.(2001). Operation data mining. Database programming and design Magazine.

Shagaw, A.(2002).Application of data mining technology to predict child mortality pattern: The Case of Butajira Rural Health Project. MSc Thesis, Addis Ababa University,Ethiopia.

Srinivas ,K., Kavihta B. & Govrdhan, A. (2010). Applications of data mining techniques in Healthcare and prediction of heart attacks. International journal on computer science and engineering ,2, 250-255.

Thearling ,K. (2003*).* An introduction to data mining. Retrieved on May 5, 2011 from http://www3.primuhost.com/~kht/text/whexcerpt/.

Trybula , W. (1997). Data mining and knowledge discovery. Annual review of information science and technology.

Tsegaye, F., Abdella, K., Ahmed, E., Tadesse T. & Bartolomeos K. (2010). Pattern of fatal injuries in Addis Ababa, Ethiopia: A one-year audit**.** East and Central African journal of surgery, 15, 10-17.

Two Crows Corporation .(1999). Introduction to data mining and knowledge discovery. 3rd ed. Retrieved on April 20, 2011 from: http://www.twocrows.com.

Vararuk, et.al (2008).Data mining techniques for HIV/AIDS data management in Thailand. journal of enterprise information management, 21(1),52-70.

Vyas,S.(2010).Using associative classifiers for predictive analysis in Health Care data mining: International journal of computer applications, 4, 0975 – 8887.

WHO(2001). Mental health: new understanding, new hope. Geneva, world health organization.

Witten, I., Frank, E., Trigg, L., Hall ,M., Holmes ,G., & Cunningham, S. (1999). Weka: Practical machine learning tools and techniques with java implementations. In: Emerging knowledge engineering and connectionist-based info. systems, 192-196.

Witten, I., & Frank, E. (2005). Data mining: Practical machine learning tools and techniques. Morgan Kaufmann.

Yishak ,A.(2009). Bacteriology of open fracture wounds in Tikur Anbessa Hospital, Addis Ababa, Ethiopia . MSc. Thesis, Addis Ababa University, Ethiopia.

Annex

Annex 1

A decision tree generated from J48 algorithm to the selected experiment or scenario for the target class Cause of visit.

Test mode: 10-fold cross-validation

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

J48 pruned tree

```
------------------
Place of injury = Home
|   Intent = Other: Non-Trauma (923.0)
|   Intent = violence: Non-Trauma (550.0/37.0)
|   Intent = accidental
|   |   Transferred to = Emergency OPD
|   |   |   Activity = Working: Non-Trauma (294.0/69.0)
|   |   |   Activity = Travelling: Trauma (8.0)
|   |   |   Activity = Other
|   |   |   |   Triage Assessment = Orange
|   |   |   |   |   Age = '[13-25]'
|   |   |   |   |   |   Region = Addis Ababa
|   |   |   |   |   |   |   Referred from = Governmental Health Facility: Trauma (24.0/9.0)
|   |   |   |   |   |   |   Referred from = self: Non-Trauma (5.0/1.0)
|   |   |   |   |   |   |   Referred from = Non Governmental Health Facility: Trauma (4.0/1.0)
|   |   |   |   |   |   Region = Amhara: Non-Trauma (1.0)
|   |   |   |   |   |   Region = Oromia: Non-Trauma (7.0/2.0)
|   |   |   |   |   |   Region = SNNPR: Trauma (1.0)
|   |   |   |   |   |   Region = Tigray: Trauma (0.0)
|   |   |   |   |   |   Region = Harrar: Trauma (0.0)
|   |   |   |   |   |   Region = Afar: Trauma (0.0)
|   |   |   |   |   |   Region = Dire Dawa: Trauma (0.0)
|   |   |   |   |   |   Region = Somalia: Trauma (0.0)
|   |   |   |   |   |   Region = Benishangul Gumuz: Trauma (0.0)
|   |   |   |   |   |   Region = Gambela: Trauma (0.0)
|   |   |   |   |   Age = '(25-37]'
|   |   |   |   |   |   Pastmedical illness = Yes
|   |   |   |   |   |   |   Sex = Male
|   |   |   |   |   |   |   |   MaritalStatus = Married: Non-Trauma (5.0/1.0)
|   |   |   |   |   |   |   |   MaritalStatus = Single: Trauma (2.0)
|   |   |   |   |   |   |   |   MaritalStatus = Other: Non-Trauma (0.0)
|   |   |   |   |   |   |   |   MaritalStatus = Devorced: Non-Trauma (0.0)
|   |   |   |   |   |   |   Sex = Female: Trauma (7.0)
```

| | | | | | Pastmedical illness = No: Non-Trauma (22.0/9.0)
| | | | | | Pastmedical illness = Unknown: Non-Trauma (2.0)
| | | | | Age = '(37-49]'
| | | | | | Referred from = Governmental Health Facility
| | | | | | | Region = Addis Ababa: Trauma (8.0/2.0)
| | | | | | | Region = Amhara: Trauma (0.0)
| | | | | | | Region = Oromia
| | | | | | | | Pastmedical illness = Yes: Non-Trauma (2.0)
| | | | | | | | Pastmedical illness = No: Trauma (2.0)
| | | | | | | | Pastmedical illness = Unknown: Non-Trauma (0.0)
| | | | | | | Region = SNNPR: Non-Trauma (2.0)
| | | | | | | Region = Tigray: Trauma (0.0)
| | | | | | | Region = Harrar: Trauma (0.0)
| | | | | | | Region = Afar: Trauma (0.0)
| | | | | | | Region = Dire Dawa: Trauma (0.0)
| | | | | | | Region = Somalia: Trauma (0.0)
| | | | | | | Region = Benishangul Gumuz: Trauma (0.0)
| | | | | | | Region = Gambela: Trauma (0.0)
| | | | | | Referred from = self: Non-Trauma (4.0/1.0)
| | | | | | Referred from = Non Governmental Health Facility: Trauma (4.0)
| | | | | Age = '(49-61]': Non-Trauma (15.0/7.0)
| | | | | Age = '(61-73]'
| | | | | | Pastmedical illness = Yes: Non-Trauma (5.0/1.0)
| | | | | | Pastmedical illness = No: Trauma (7.0/2.0)
| | | | | | Pastmedical illness = Unknown: Non-Trauma (0.0)
| | | | | Age = '(73-85]': Trauma (3.0)
| | | | | Age = '(85-97)': Trauma (1.0)
| | | | Triage Assessment = Yellow: Non-Trauma (57.0/15.0)
| | | | Triage Assessment = Red: Non-Trauma (9.0/3.0)
| | | | Triage Assessment = Green
| | | | | Age = '(-inf-25]': Non-Trauma (6.0/1.0)
| | | | | Age = '(25-37]': Trauma (4.0/1.0)
| | | | | Age = '(37-49]': Non-Trauma (0.0)
| | | | | Age = '(49-61]': Non-Trauma (1.0)
| | | | | Age = '(61-73]': Non-Trauma (0.0)
| | | | | Age = '(73-85]': Non-Trauma (0.0)
| | | | | Age = '(85-97)': Non-Trauma (0.0)
| | | | Triage Assessment = Black: Non-Trauma (1.0)
| | | Activity = Playing or Recreating: Non-Trauma (154.0/13.0)
| | | Activity = Studying: Non-Trauma (38.0/5.0)
| | Transferred to = Resuscitation Room: Non-Trauma (24.0/3.0)
| | Transferred to = Regular OPD: Non-Trauma (1.0)
| | Transferred to = Home: Non-Trauma (1.0)
| | Transferred to = Referred to other hospital: Non-Trauma (0.0)
| | Transferred to = Stabilization room: Trauma (6.0)
Placeofinjury = Other

| Intent = Other: Non-Trauma (28.0)
| Intent = violence
| | Pastmedical illness = Yes: Non-Trauma (4.0)
| | Pastmedical illness = No: Trauma (11.0/3.0)
| | Pastmedical illness = Unknown: Trauma (3.0/1.0)
| Intent = accidental: Trauma (37.0/5.0)
Placeofinjury = Street
| Intent = Other: Non-Trauma (9.0)
| Intent = violence
| | Activity = Working
| | | Region = Addis Ababa: Trauma (6.0)
| | | Region = Amhara: Non-Trauma (2.0)
| | | Region = Oromia: Trauma (0.0)
| | | Region = SNNPR: Non-Trauma (2.0)
| | | Region = Tigray: Trauma (0.0)
| | | Region = Harrar: Trauma (0.0)
| | | Region = Afar: Trauma (0.0)
| | | Region = Dire Dawa: Trauma (0.0)
| | | Region = Somalia: Trauma (0.0)
| | | Region = Benishangul Gumuz: Trauma (0.0)
| | | Region = Gambela: Trauma (0.0)
| | Activity = Travelling: Trauma (101.0)
| | Activity = Other: Trauma (1.0)
| | Activity = Playing or Recreating: Trauma (41.0)
| | Activity = Studying: Non-Trauma (1.0)
| Intent = accidental: Trauma (1696.0/23.0)
Placeofinjury = Work
| Activity = Working
| | Intent = Other: Non-Trauma (47.0)
| | Intent = violence: Trauma (115.0/13.0)
| | Intent = accidental: Trauma (277.0/14.0)
| Activity = Travelling: Trauma (4.0)
| Activity = Other
| | Intent = Other: Non-Trauma (44.0)
| | Intent = violence: Non-Trauma (43.0/5.0)
| | Intent = accidental
| | | Transferred to = Emergency OPD
| | | | Pastmedical illness = Yes: Non-Trauma (21.0/4.0)
| | | | Pastmedical illness = No
| | | | | Subcity = Arada: Non-Trauma (6.0/1.0)
| | | | | Subcity = Outside Addis Ababa: Trauma (15.0/4.0)
| | | | | Subcity = Bole: Trauma (1.0)
| | | | | Subcity = Yeka: Trauma (1.0)
| | | | | Subcity = Kirkos: Non-Trauma (3.0)
| | | | | Subcity = Ledeta
| | | | | | Referred from = Governmental Health Facility: Non-Trauma (2.0)

| | | | | | Referred from = self: Trauma (4.0/1.0)
| | | | | | Referred from = Non Governmental Health Facility: Trauma (1.0)
| | | | | Subcity = Kolfe Keranio: Non-Trauma (2.0)
| | | | | Subcity = Gulele: Non-Trauma (2.0/1.0)
| | | | | Subcity = Akaki Kaliti: Trauma (0.0)
| | | | | Subcity = Nefas Silk Lafto: Trauma (8.0/1.0)
| | | | | Subcity = Addis Ketema: Trauma (0.0)
| | | | Pastmedical illness = Unknown: Non-Trauma (31.0/4.0)
| | | Transferred to = Resuscitation Room: Non-Trauma (44.0/4.0)
| | | Transferred to = Regular OPD: Non-Trauma (3.0)
| | | Transferred to = Home: Non-Trauma (0.0)
| | | Transferred to = Referred to other hospital: Non-Trauma (0.0)
| | | Transferred to = Stabilization room: Trauma (3.0/1.0)
| Activity = Playing or Recreating: Trauma (1.0)
| Activity = Studying: Non-Trauma (2.0)
Placeofinjury = Recreational Place: Trauma (252.0/2.0)
Placeofinjury = School
| Activity = Working: Non-Trauma (4.0/1.0)
| Activity = Travelling: Trauma (16.0)
| Activity = Other: Non-Trauma (41.0/6.0)
| Activity = Playing or Recreating
| | MaritalStatus = Married: Non-Trauma (3.0)
| | MaritalStatus = Single: Trauma (5.0/1.0)
| | MaritalStatus = Other: Non-Trauma (0.0)
| | MaritalStatus = Devorced: Non-Trauma (0.0)
| Activity = Studying: Non-Trauma (371.0/7.0)

Annex 2: Rule generated by PART algorithm

=== Run information ===
Test mode: 10-fold cross-validation
=== Classifier model (full training set) ===
PART decision list

Placeofinjury = Street AND
Intent = accidental: Trauma (1696.0/23.0)

Placeofinjury = School AND
Activity = Studying: Non-Trauma (371.0/7.0)

Placeofinjury = Recreational Place: Trauma (252.0/2.0)

Placeofinjury = Street AND
Intent = violence AND
Activity = Travelling: Trauma (101.0)

Placeofinjury = Home AND
Intent = Other: Non-Trauma (923.0)

Placeofinjury = Home AND
Intent = violence AND
Activity = Studying: Non-Trauma (136.0)

Placeofinjury = Home AND
Transferred to = Resuscitation Room AND
Activity = Other AND
Triage Assessment = Orange: Non-Trauma (38.0/1.0)

Placeofinjury = Home AND
Transferred to = Resuscitation Room AND
Activity = Playing or Recreating: Non-Trauma (19.0)

Placeofinjury = Home AND
Transferred to = Emergency OPD AND
Intent = violence AND
Triage Assessment = Orange AND
Subcity = Outside Addis Ababa: Non-Trauma (68.0/1.0)

Placeofinjury = Street AND
Intent = violence AND
Activity = Playing or Recreating: Trauma (41.0)

Placeofinjury = Home AND
Transferred to = Emergency OPD AND
Activity = Playing or Recreating AND
Subcity = Outside Addis Ababa: Non-Trauma (49.0/2.0)

Placeofinjury = Other AND
Activity = Travelling: Trauma (19.0)

Placeofinjury = Home AND
Transferred to = Emergency OPD AND
Activity = Working AND
Triage Assessment = Orange: Non-Trauma (154.0/22.0)

Activity = Working AND
Placeofinjury = Work AND
Sex = Male: Trauma (322.0/15.0)

Placeofinjury = Home AND
Transferred to = Emergency OPD AND
Activity = Working AND
Triage Assessment = Orange: Non-Trauma (154.0/22.0)

Region = Addis Ababa AND
Triage Assessment = Yellow AND
Age = '(37-49]' AND
Sex = Female: Trauma (5.0/1.0)

Placeofinjury = Home AND
Transferred to = Emergency OPD AND
Intent = violence AND
Triage Assessment = Yellow: Non-Trauma (80.0/7.0)

Placeofinjury = Work: Trauma (17.0/6.0)
Intent = accidental AND
Age = '(25-37]' AND
Sex = Male: Non-Trauma (14.0/2.0)