ADDIS ABABA UNIVERSITY

SCHOOL OF GRADUATE STUDIES

SCHOOL OF INFORMATION SCIENCE

AND

SCHOOL OF PUBLIC HEALTH

APPLICATION OF DATA MINING TECHNIQUES TO PREDICT URINARY FISTULA SURGICAL REPAIR OUTCOME

MINALE TEFERA

JUNE 2012

ADDIS ABABA UNIVERSITY

SCHOOL OF GRADUATE STUDIES

SCHOOL OF INFORMATION SCIENCE

AND

SCHOOL OF PUBLIC HEALTH

APPLICATION OF DATA MINING TECHNIQUES TO PREDICT URINARY FISTULA SURGICAL REPAIR OUTCOME: THE CASE OF ADDIS ABABA FISTULA HOSPITAL, ADDIS ABABA, ETHIOPIA.

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

BY
MINALE TEFERA

JUNE 2012

ADDIS ABABA UNIVERSITY

SCHOOL OF GRADUATE STUDIES

SCHOOL OF INFORMATION SCIENCE

AND

SCHOOL OF PUBLIC HEALTH

APPLICATION OF DATA MINING TECHNIQUES TO PREDICT URINARY FISTULA SURGICAL REPAIR OUTCOME: THE CASE OF ADDIS ABABA FISTULA HOSPITAL, ADDIS ABABA, ETHIOPIA.

BY
MINALE TEFERA

Members of the examining board:

Name	Title	Signature	Date
_____	Chair person	_____	_____
_____	Advisor	_____	_____
_____	Advisor	_____	_____
_____	Examiner	_____	_____
_____	Examiner	_____	_____

DECLARATION

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

_____

Date

ACKNOWLEDGEMENT

First and foremost I would like to acknowledge my advisors, Ato Getachew Jemaneh and Dr Mitike Mola, for constructive comments and guidance that they have provided me throughout the study. I would like also to express my strong appreciation for the approach, treatment and the help I have got at the time of difficulties.

I thank the staff in the Addis Ababa Fistula Hospital for their generous assistance, and friendship. I couldn't have started the study, if it is not for the interest and support contributed form Dr. Catherine Hamlin, Richard Hamlin, Ato Abebe Gesit, Dr. Habtemariam Tekle, Dr Fiseha Tiku, Ato Michael Miraw. I am really indebted to Dr. Dreje Ayele for his support in evaluating the model/findings based on previously existing knowledge in the domain area; and to W/ro Eleni Aschalew and W/ro Roman Yilma on dealing with problems in the data and technical facilitations.

I would also like to thank Addis Ababa University, School of Information Science and School of Public Health for financial support and overall facilitation of the research from the beginning until the end.

Finally, I would like to thank my classmates for their comments, constructive ideas and suggestions; to my best friends and relatives for their moral support and understanding during the time I solely devoted to the study.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ACRONYMS

AAFH	Addis Ababa Fistula Hospital
ARFF	Attribute Relation File Format
AUC	Area Under (ROC) curve
CLI	Command Line Interface
CRISP-DM	CRoss-Industry Standard Process model for Data Mining
CSV	Comma Separated Value
DBMS	DataBase Management System
FDA	Food and Drug Administration
FP	Forward Pruning
GUI	Graphical User Interface
HEWs	Health Extension Workers
ICS	International Consultation on Incontinence
KBS	Knowledge Base System
KDD	Knowledge Discovery in Databases
KDP	Knowledge Discovery Process model for data mining
MB	Mega Byte
MMR	Maternal Mortality Rate
MLP	Multi Layer Perceptrones
ROC	Receiver Operating Characteristics
RVF	Recto Vaginal Fistula
SVM	Support Vector Machine
UNFPA	United Nations Population Fund
US	United States
VVF	Vesico Vaginal Fistula
Weka	Waikato Environment for Knowledge Analysis
WHO	World Health Organization

ABSTRACT

Background: The likelihood of the occurrence of incontinence after successful surgical repair makes predicting urinary fistula surgical repair outcome important for decision making during operation and for further follow up and treatment.

Objective: The purpose of this thesis is to apply data mining techniques to build a model that can assist in predicting surgical outcome of urinary fistula repair based on clinical assessments done just before surgical repair.

Methodology: The six-step hybrid knowledge discovery process model is used as a framework for the overall activities in the study. 15961 instances that have undergone urinary fistula repair in Addis Ababa Fistula Hospital are used for both predictive association rule extraction and predictive model building. Apriori algorithm is used to extract association rules while classification algorithms J48, PART, Naïve Bayes and multinomial logistic regression are used to build predictive models. Support and confidence are used as interestingness measure for association rules while area under the WROC and ROC curve for each specific outcome is sequentially used to compare performances of models from the predictive algorithms.

Results: Predictive association rules from Apriori have shown frequent co-occurrence of less severity of injury with cured outcome. The predictive model from PART-M2-C0.05-Q1 scheme has shown an area under WROC curve of 0.742. Area under the ROC curve for residual outcome ($ROC_{Residual}$=0.822) from this algorithm is better than Naïve Bayes and logistic, while the areas under the ROC curves for the other outcomes are greater than the model from J48.

Conclusion: Predictive model is developed with the use of PART-M2-C0.05-Q1. It is better in detecting residual outcome than the logistic regression model. The predictive association rules and predictive model built with the use of data mining techniques can assist in predicting urinary fistula surgical repair outcome.

CHAPTER ONE

INTRODUCTION

1.1. Background

1.1.1. Data Mining

Since the 1990s, the social and economic structure of the world has changed from industrial and product oriented environment to information and knowledge dependant one. Rapid growth of information technologies and its integration with digital networks, software, and database systems are the main characteristics of information and knowledge society. This rapid technological progress increased the storage capacity of devices, improved the available storage technology, and created new storage technologies. At the same time advances in interconnectivity of computer technology and software systems speeded up the rate at which digital data is acquired, processed, stored, retrieved and communicated. These major progresses of the information age gave rise to the availability of massive amount of electronic data stored in databases, data warehouses, or other kinds of data repositories such as the World Wide Web (1).

The explosive growth in raw data accumulation in turn widened the gap between raw data that is not yet analyzed and meaningful information available for decision making. Because of the high volume of data, activities such as finding patterns, trends, irregularity, and summarizing them with simple quantitative models became a great challenge for the information age – turning data into information and turning information into knowledge – lead to a demand for specialized tools to view and analyze the data (2). As a result of this demand, a new field of science was born, in which; statistics, machine learning, pattern recognition, and databases merge to form a multidisciplinary field called data mining. On the basis of its roots, data mining is defined in Encyclopedia Britannica online dictionary by Christopher Clifton as:

> A field that combines tools from statistics, machine learning, database management to analyze large digital collections, known as datasets (3).

Due to its multidisciplinary nature, data mining has a capability derived from the above mentioned fields and others such as: artificial intelligence, pattern recognition, data visualization, mathematical algorithms. However, the primary and major contributors to its capability are database or database management, machine learning and statistics. In short, the capability of these three primary contributors can be described as follows. Database technology manages data for convenient selection and retrieval of the data while machine learning technology learns information or patterns from data in an automatic way (i.e. algorithms that improve their performance automatically through experience, such as neural networks or decision trees), and statistics finds characteristics or statistical parameters of the data. The rest are information retrieval, algorithms and parallel processing. The former focuses on selecting and searching related documents by keywords which makes it more suited to text processing and mining while the latter ones are highly valuable in mining large amount of data (4).

The development of a common and universal definition to data mining seems very challenging due to the multiplicity of the fields which created it. However, one can easily identify the variations are due to emphasis given to the methods used in data mining rather than the purpose of data mining. There are some definitions which emphasize the purpose of mining data irrespective of the methods utilized. For example, according to Nisbet, Elder and Miner (5) one of the earliest definitions of data mining is the non trivial extraction of implicit, previously unknown, and potentially useful information from data. A closer definition to this early definition is the one provided by Hand, Manila and Smyth (6) which defines data mining as the analysis of large, observational, data bases with the objective of extracting previously unsuspected relationships and summarizing the data in novel ways that are both comprehensible and useful to the data owner. In addition, Cios, Pedrycz, Swiniarski and Kurgan (7) also defines data mining, focusing on the objectives it performs, as an activity that enables us to make sense of large amounts of data.

There are definitions which emphasize on the methods used in data mining than the purposes that data mining accomplishes. According to machine learning point of view data mining is considered as a set of methods used in the knowledge discovery process to distinguish previously unknown relationships and patterns within data (2). This last definition shows that data mining is not a standalone (self sufficient) activity which results in useful, previously unknown relationships and summaries rather a step in knowledge discovery process. Discovered new knowledge should also exhibit a series of essential attributes such as understandability, validity, novelty, and usefulness.

Data mining is just a single but crucial step in a larger process known as "Knowledge Discovery in Databases" (KDD). Other steps in the KDD process, in progressive order, include data cleaning, data integration, data selection, data transformation, data mining, pattern evaluation, and knowledge presentation (8). The steps in KDD process are detailed only to show the basic difference between data mining (single step) and KDD process. Since, another very detail and explicit process model (six-step KDP process model) is used as framework to guide this data mining research.

A comprehensive definition to KDD acknowledges the existence of several stages. Knowledge discovery in databases is, therefore, a process in several stages, not trivial, interactive and iterative, for identification of new, valid, understandable and potentially useful patterns in data (9). Thus, for the purpose of this study data mining is defined as a single step in KDD in which various techniques and algorithms were applied on databases for the discovery of patterns and in order to transform information to knowledge, to assist the decision making process or to explain and justify it.

Nowadays, data mining is getting wide acceptance in different organizations. A number of advances in technology and business processes have contributed to the growing acceptance and application of data mining in both the public and private sectors. The first of these advances is associated with the growth of computer networks, that is used in interconnecting databases; followed by the development of enhanced search-related

techniques such as neural networks and advanced algorithms; the use of client-server computing which allowed users to access centralized data resources from the desktop and increased ability to combine data from different sources into a single searchable source (10).

The second reason for the wide acceptance of data mining is its applicability on data represented in different formats such as quantitative, textual, and multimedia and the variety of techniques it possesses. Techniques found in data mining include association rule mining which refers to identification of patterns where one event co-occurs with another event; sequence or path analysis where a pattern of one event leads to another event; classification referring to the identification of a class to which new instance belongs to; clustering refers to finding and visually documenting groups of previously unknown facts; forecasting which refers to another form of classification for discovering patterns based on which one can make reasonable predictions regarding future activities (11, 12).

Amongst the significant importance of data mining to medical and healthcare fields; evaluating effective courses of treatment, prediction of surgical outcome, and prediction of length of stay are of great importance for both parties i.e. the patient and the health care service provider. Data mining capacity in treatment comparison has shown its ability by identifying the best course of action that proves effective, even in determining proactive steps that can reduce the risk of affliction (13, 14). With respect to predicting surgical outcome a study by Reinbolt et al (15) demonstrated that the use of linear discriminant analysis enabled to predict whether or not a patient's knee motion will improve following rectus femoris transfer surgery based on few preoperative gait analysis measurements. The same potential can be leveraged to the prediction of urinary fistula surgical repair outcome i.e. the most common of clinical presentations of obstetric fistula.

1.1.2. General Overview of Obstetric Fistula

Maternal outcomes are good in most countries of the developed world while the same is not true in many developing and resource-poor countries. This disparity in maternal

outcomes can easily be seen from the maternal mortality rate and lifetime risk of maternal death. For example, the 2008 estimate of maternal mortality ratio for developed regions is 14 per 100,000 live births while it is 290 per 100,000 live births for developing regions. In the same line, the lifetime risk of maternal death is 1 per 4,300 births for developed regions while it is 1 per 120 births for developing regions. The above statistics for Sub Saharan Africa will rise to MMR (Maternal Mortality Ratio) of 640 per 100,000 live births and lifetime risk of maternal death of 1 per 31 births (16).

WHO defines maternal mortality as: "the death of a woman while pregnant or within 42 days of termination of pregnancy, irrespective of the duration and site of the pregnancy, from any cause related to or aggravated by the pregnancy or its management, but not from accidental or incidental causes" (17). In turn, maternal morbidity is defined as "a condition outside of normal pregnancy, labour, and childbirth that negatively affects a woman's health during those times" (18).

Generally, throughout the world, half a million women die from complications of pregnancy or childbirth every year, most of which occurs in resource-poor countries. In 2008 alone, an estimated 358,000 maternal deaths occurred worldwide because of complications related to pregnancy and childbirth from which developing countries accounted for 99% of the deaths. Furthermore, the analysis of the maternal mortality data for Sub-Saharan Africa and South Asia alone has shown that 87% of the global maternal deaths occurred in countries of these regions (16).

Among those complications of pregnancy and childbirth, the contribution of obstructed labour to maternal mortality is negligible in developed countries while it accounts significant percent of maternal deaths in developing world (19, 20). Due to shortage or absence of emergency obstetric care; obstructed labour ends with stillbirth and maternal mortality in many developing countries (21). Even if the woman survives obstructed labour, she often sustains the multidimensional effects of the obstruction. Arrowsmith et al cited in (21) described the multidimensional effects of obstructed labour using the term "obstructed labour injury complex" to represent the broad scope of injuries resulting from prolonged obstructed labour, representing a syndrome that often involves multiple organ

systems. The obstetric labour injury complex includes urologic and gynaecologic injuries, rectovaginal fistulas, orthopaedic trauma, neurologic, dermatologic and psychosocial injury in addition to the vesico-vaginal fistula (22).

A fistula is an abnormal opening between the vagina and the bladder, the most common and the one which dominates the clinical presentations i.e. vesico vaginal fistula (VVF), and/or between the vagina and rectum i.e. recto vaginal fistula (RVF). The immediate manifestation of obstetric fistula is, therefore, persistent leakage of urine or feces, or both; while the secondary consequences include a range of devastating physical co-morbidities, psychological, social and economic problems (23, 24).

A woman's obstetric history is the most significant element in the development of fistula: hence the term "obstetric fistulas". Other direct causes of fistula formation include iatrogenic injury, accidents, sexual abuse and rape and genital mutilation (25). However, the development of obstetric fistula is also commonly associated with background factors such as early age at first marriage, illiteracy, rural residence and living far from the nearest health care facility (26).

Despite its devastating effects the exact prevalence of obstetric fistula is unknown while it is estimated to affect thousands of women in developing countries. The most frequently reported global prevalence of obstetric fistula shows that approximately 2 million women have untreated fistula in Asia and sub-Saharan Africa alone and additional 50,000 to 100,000 women develop obstetric fistulas each year (16). Other estimates show that as many as 130,000 new cases of fistula are occurring annually in Africa and globally up to 3.5 million women may be living with the condition (21). In Ethiopia also obstetric fistula is a health challenge to thousands of women where 9,000 are affected each year (23).

1.1.3. Treatment to Obstetric Fistula Victims and Its Outcomes

Once fistula has already developed, surgical repair is necessary since it does not heal by itself. But, fistula patient often has more than just a fistula and treating a woman in whom

a fistula is suspected requires a holistic approach. For example, victims arriving at fistula repair centers have often had the fistula for months or even years, and are suffering from malnutrition and anemia, which must be improved before surgery. Intensive physical therapies are provided for lower limb weakness, muscular contractures, and foot drop in order to make them able to walk again (23, 27). Psychological and emotional counseling is also part of the treatment which begins preoperatively and will continue postoperatively (23, 28). Recommended laboratory tests in preoperative care period include a check for sexually transmitted diseases because all infections need to be treated prior to surgery.

During preoperative care a complete physical examination is done to locate fistulas and a thorough recto-vaginal examination is also performed to determine the coexistence of recto-vaginal fistulas (29). In general, the clinician looks for evidence of the entire "obstetric labour injury complex" whose clinical assessments include assessments not only of urologic and gynecologic injury, but also for evidence of the presence of orthopedic trauma, neurologic, dermatologic injury, and psychological impact (30). Prior to their operation women are examined at the outpatient clinic or in the operation theatre for the site, the size and number of fistulas, amount of scarring and presence of other complications (29). The information gained from these examinations and assessments are used during preoperative care, during surgical intervention, and in post operative care to make the victims physical and mental health ready for surgical repair, to decide on surgical repair approach and finally to decide on the length of stay and the various postoperative cares that need to be provided to the victim.

Collectively, all the efforts in preoperative, operative and postoperative care are made with the objective of restoring continence to enable the victim to resume and lead a full and active life. Studies show that in areas where surgical repair services are available and accessible, 80- 95% of vaginal fistula can be closed surgically (31), however, up to 33% of these women may suffer residual incontinence even after successful closure of the fistula (32, 33). There are a number of studies which show repair success and incontinence between the ranges specified above. For example, research conducted on

surgical success rates shows fistulas can be closed successfully in 80% to 95% of cases (34). A research for comparison of abdominal and vaginal routes of repair by Rakesh et al in India shows an overall success rate of 94.2% (35). A study in Nigeria also has shown that an overall repair success rate of 79.2% (36). A descriptive study of 716 vesico-vaginal and recto-vaginal fistulae victims managed at the Addis Ababa Fistula Hospital and Birmingham and Midland Hospital showed that 84.6% cured at the first repair attempt, 6.3% failed and 9.1% had stress incontinence (37). A research by Pierre et al (2010) in Cameron has shown that complete closure of vesico-vaginal fistula (VVF) was achieved in 92% of 25 victims and among those with complete closure 74% had good continence showing the remaining 26 % incontinent despite successful closure (38). However, the remaining rates in the above studies indicate that there is a chance of failure of surgical repair and persistent incontinence despite successful closure of the fistula which has been termed as "the continence gap" (39). Estimates of persistent urinary incontinence after a successful closure of the fistula comes from case series, ranging from 16.3% in a large retrospective review of victims by (40) to 33% in a small series of complex fistulas in which the proximal urethra was lost (41).

Further studies with the objective of identifying the risk factors associated with surgical repair failure and incontinence despite success in repair were conducted in different countries. A Retrospective cross-sectional study in fistula referral hospital in eastern Democratic Republic of Congo shows closure rate of 87.1%, with 15.6% remaining incontinent. The attributes considered during the study are amount of fibrosis, overall repair attempt, distance to fistula, size of fistula, place of fistula (vesico-vaginal fistula high, vesico-vaginal fistula low, circumferential), duration of fistula, parity, age at fistula repair, site of fistula (high and urethra not involved, low and urethra involved , circumferential). This study utilized multivariate regression models to predict surgical outcome and has found that failure to close the fistula was significantly associated with previous repairs, amount of fibrosis and fistula size; and incontinence was significantly associated with previous repairs, amount of fibrosis and fistula location (42).

Goh also has shown that, out of a total of 131 genital tract fistula repairs performed in Addis Ababa Fistula Hospital, 98.5% of fistulas were successfully repaired in first attempt and 7.5% of women suffered from urinary incontinence. The author indicated the factors contributing to incontinence following obstetric fistulas repair as the destruction of the urethra; large fistulas may result in gross vesical contraction; fistulas destroying the bladder neck and nerve supply to the bladder; and marked scarring may distort bladder and urethral anatomy. The study has also noted that most of the women who complained of incontinence were associated with juxta-urethral or urethral fistulas and significant vaginal scarring (29).

Another study by Browning in Addis Ababa Fistula Hospital considers a larger set of attributes in order to identify risks factors in those women who remain incontinent after fistula repair. The attributes included in this study are: urethral involvement, women having repeat procedure (number of repeats), fistula diameter, small bladder, ureter outside bladder, vaginoplasty at operation, flap vaginal reconstruction, recto-vaginal fistula present, multiple vesico-vaginal fistulae present, parity, average fistula diameter (cm), time in labour (days), age (years), time since delivery (months), delivery method (vaginal delivery, caesarean, caesarean, hysterectomy), birth-outcome. After doing statistical tests by chi-square, t test, and Mann–Whitney U test and odds ratio: urethral involvement, small bladder, vaginoplasty (vaginal scarring), diameter of fistula were reported as significant independent predictors of developing urinary incontinence following fistula repair. But the study itself has noted that circumferential injuries to the urethra were not examined separately in the series which may even be stronger predictor of residual incontinence (43).

Goh et al have studied the possibility of predicting the risk of failure of closure and post-fistula urinary incontinence using a classification system. Out of a total of 987 women who had genito-urinary fistula repair performed in Addis Ababa Fistula Hospital, 97.2% had successful closure of their fistulae and of those with successful closure 23.8% complained of urinary incontinence following surgery. The attributes considered in this study were type, size and special considerations: Type: distance from external urinary

meatus (2.5–3.5 cm, 1.5–2.5 cm, <1.5 cm), Size: diameter in centimeters (<1.5 cm, 1.5–3 cm, >3 cm), Special considerations (I: None or mild fibrosis and/or vaginal length>6 cm, normal capacity; II: Moderate or severe fibrosis and/or marked reduction in vaginal length and/or capacity; III: Special circumstances, e.g. ureteric involvement, circumferential fistula, previous repair). This study used logistic regression and showed that the location of the fistula is not associated with a risk of unsuccessful anatomical closure while the risk of residual urinary incontinence is location dependent. The size of the fistulae is also related to risk of residual incontinence, with an increased risk of incontinence in women with larger fistulae. Special considerations such as scarring or circumferential fistulae are significant factors for incontinence and failure of closure (44).

1.2. Statement of the Problem and Justification

Due to shortage of emergency obstetric care, obstructed labour ends with still birth and leaves a fistula which makes holistic treatment a necessity to relieve the direct and related consequences of the injury. Addis Ababa Fistula Hospital is exemplary in the provision of this type of treatment to victims of fistula of various causes and types, in addition to obstetric fistula.

The hospital keeps the record of the services provided in the different courses of the treatment process to victims both in manual and electronic form. Various types of assessments and clinical examinations taken from each victim, the type of fistula and its surgical repair outcome is part of the information found in the database. Thus, knowledge extracted from this rich information source can be used to help the effort of the hospital in the provision of high quality care. The use of data mining for prediction of surgical outcomes helps knowledge discovery with the end goal of further improvement in quality of care. According to Wall, Arrowsmith, Briggs, Browning, Lassey, the necessity of providing high quality care for women who already developed obstetric fistula is stated as:

> Although the solution to the fistula problem will ultimately come from the provision of essential obstetric services for all of the world's women, the current needs of those women who have already developed an obstetric fistula cannot be ignored …. specialized centers should be created in all countries where obstetric fistulas are prevalent …. they should serve as centers of compassionate excellence and should provide high quality patient care, medical and nursing education, and clinical research as part of their mandate for existence (34).

Acknowledging the role that research plays in improving quality of care and its importance in medical and nursing education, the Addis Ababa Fistula Hospital has set an objective of being a research center in addition to being treatment center. Initial literature review also showed this reality since a number of studies, with different objectives, were conducted in the hospital. Some of these studies were done to identify or predict the risk factors associated with failure of repair of urinary fistula and urinary incontinence despite successful repair. For example, the research conducted by Goh et al (44) has shown the association of fistula characteristics suggested in specific classification system with the possibility of successful surgical repair and urinary incontinence following successful fistula repair.

A research by Browning has also used a multivariate analysis to determine factors that were significant independent predictors for developing incontinence following fistula repair (45). Though, it was done on a small number of patient records, the research has incorporated the interventions undertaken in addition to the characteristics of the fistula for the purpose of determining the significant predictors of incontinence following fistula repair (32).

The efforts of these researchers have contributed a lot in the identification of the risk factors for both surgical failure and urinary incontinence despite surgical repair success. But the logistic regression models used were limited to predicting only two outcomes (success or failure and incontinence or continence). However, surgical outcome for a

urinary fistula victim who have undergone surgical repair is measured by categorical values such as: cured, failed, stress and residual.

The main advantage of using data mining techniques over the statistical methods used is that they enable us to predict outcomes at patient level than only identifying risk factors for a specific outcome at group level. Typically logistic regression used in the above studies helped to identify risk factors for urinary incontinence, thereby making easy to identify victims at high risk and those that are not. But, models developed with the use of data mining techniques can be used to predict health outcomes at specific patient level (45).

It is therefore reasonable enough to study what attribute values are associated with those listed surgical repair outcomes of urinary fistula and develop a model that assists in predicting future surgical outcome of a victim based on the values of significant attributes identified. The ability of the tools and algorithms of data mining to deal with datasets characterized by thousands of instances and high dimensionality (large number of attributes) coupled with the understandability of models produced at the end and their ease of use makes data mining suitable for this study.

The purpose of the research is therefore, to apply data mining techniques and build a model that maps clinical examination attributes with the outcome of surgical repair for urinary fistula. This research will also compare the performance measures of logistic regressions with that of Decision Trees, Decision rules, Naïve bayes, multinomial logistic regression so as to come up with a model of relatively higher area under the ROC (Receiver Operating Characteristics) curve. To this end, this research will try to answer the following questions:

1. What values of predictive factors (attributes) are associated with each outcome of urinary fistula repair?
2. Would it be possible to draw association rules among the attributes and the classes of urinary fistula surgical repair outcomes?

3. Can models from other algorithms predict urinary fistula surgical repair outcome with better sensitivity and specificity expressed as area under ROC curve than logistic regression?
4. Would it be possible to build a model that can be used to predict new victim's future surgical repair outcome for urinary fistula on the basis of clinical examination or assessment?

1.3. Objective

The general objective of this research is to apply data mining techniques to develop a model that can assist in predicting the outcome of surgical repair for urinary fistula victims based on clinical examination or assessment data in Addis Ababa Fistula Hospital.

1.3.1. Specific Objectives:

Some specific objectives are derived from the general objective of the research stated above for the purpose of ease of manageability and monitoring. Thus, the specific objectives of this research include:

a) Conduct a thorough review of literature on data mining methods and techniques that can be used to attain the objective of surgical repair outcome prediction for urinary fistula.
b) Select and extract the dataset required for analysis from the database of Addis Ababa Fistula Hospital.
c) Prepare the data by performing activities like resolving for inconsistencies, managing missing values.
d) Extract different predictive association rules that show relationship in between the attributes and the class attribute in the dataset.
e) Compare the performance of model from logistic regression with models built by other classification algorithms in predicting surgical repair outcome.

f) Build a predictive model that can be applicable on a new instance in order to determine the surgical repair outcome for urinary fistula.
g) Report research findings and make recommendations.

1.4. Research Methodology

The hybrid model (six-step KDP model) is chosen to be used as a framework to guide the overall activities in the current study. The main reason for the selection of this model is its capacity in providing research-oriented description of steps in knowledge discovery process.

A. Understanding Treatment to Obstetric Fistula Victims.

Review of journals and committee proceedings are main sources of information that were used to understand and describe the problem (the need for predicting urinary fistula surgical repair outcome). In addition domain experts are consulted in order to clearly specify the benefits from predicting surgical repair outcome. This business problem is translated into data mining objectives as shown in the objectives section of this chapter. Weka 3.6.4 is selected and used in the research process.

B. Understanding and Preparation of Obstetric Fistula Victims' Treatment Data.

The initial dataset is acquired from the Addis Ababa Fistula Hospital database that is used to store demographic, social, physical, and clinical assessment and treatments of victims of fistula. General description of the obstetric fistula victim treatment dataset is given; attributes and instances that are used for the objective (urinary fistula surgical repair outcome predictive model building) are selected. Exploratory data analysis with the use of statistical summary measures is done on the selected attributes. The quality of the data is assessed based on the information obtained from the statistical summaries. The data is cleaned from missing values, noises and inconsistencies.

C. Experimentation, Analysis and Evaluation of Discovered Knowledge.

Algorithms like J48, PART, Naïve Bayes, and logistic regression were used for predictive model development while Algorithm Apriori is used to drive predictive association rules. Two successive experiments are done on each algorithm, first, by altering parameters then on the best scheme by increasing SMOTE from 100-500%. Models from these experiments were compared based on area under the WROC curve.

The ROC curves for each surgical repair outcome from the best schemes of each algorithm are compared to select the best model. Consultation with the surgeons is conducted in evaluating the novelty, interestingness, and understandability of the discovered knowledge. Recommendation is given on how and when to deploy and implement the discovered knowledge.

1.5. Significance of the Study

The findings from this research can be used to assist in predicting the urinary fistula surgical repair outcome of a victim. The study also supplements those few researches which used the logistic regression method for the purpose of predicting surgical repair success and identifying the risk factors for residual incontinence after successful surgery.

Obtaining assistance while predicting the success or failure of surgical intervention can help surgeons during the time of surgery; and, the predictions can also be used in giving special counseling post-operatively. For example, predicting residual and stress incontinence despite successful repair will help identify victims who need to be offered post-operative follow-up assistance. In addition to these direct benefits of the model during treatment of victims, the model and rules can be used for planning and decision making in the hospital management.

1.6. Scope and Limitation of the Study

The scope of this research is limited to developing a model that can assist in predicting urinary fistula surgical repair outcome in Addis Ababa Fistula Hospital. While the findings of this research can pave a way for similar undertakings in other branches, the scope of the study is limited to instances found in obstetric fistula victims' treatment database found in the hospital. The study is conducted starting from October 20, 2011 to May 10, 2012. The data available in the electronic format only is considered for the study. The reason for focusing on the electronic data which is found in this main hospital is the limitations in time and budget.

The time that was given to undertake this research work is a serious limitation not only to focus on the database found in the main hospital but also has been a limitation in developing a model which works only for urinary fistula surgical repair outcome prediction. Thus, the scope of the study does not include developing a model for other types of injuries like RVF, second and third degree tear. Furthermore, the current study tried to answer the question of whether we can predict surgical outcome of urinary fistula with the use of data mining techniques based on victim assessment related data. Various areas suggested as seeking immediate researches such as: studies on techniques for fistula closure and related surgical issues which demands comparison of various surgical methodologies and techniques of repair; and the role of physical therapy and reconstructive orthopedic surgery in the management of women with obstetric foot-drop from prolonged obstructed labour suggested by Wall, Arrowsmith, Briggs, Browning and Lassey (34) are not addressed in this research. Therefore, issues like the role of surgical techniques utilized for repair, and association of clinical interventions before and after operation with outcome of repair are out of the scope of this research.

Another limitation of this study was lack of literatures related to the application of data mining techniques to prediction of outcomes after surgical interventions and the application of data mining techniques in clinical and surgical datasets. Limited prior

experience to data mining techniques and limited know how of the problem domain is the additional limitations that were faced by the researcher.

1.7. Ethical Considerations

The instances in the dataset include the victims identifying information and health information including medical history, examination results, diagnostic tests, treatment methods, and all other services provided by the hospital. Beyond explicit importance and use of the information in therapeutic process, researches like this thesis make use of it. But, the use of these medical information of instances for research and other varied purposes raises ethical issues such as: patient's privacy or confidentiality. Improper use and disclosure of any medical data negatively affects the individual's social image, social role and dictate a way of life since these instances reflect real constraints and acquired health risks. However, the research is for the purpose of professional contribution to assist obstetric fistula treatment and it will not attempt to harm anybody in any way.

Identifying information were removed from the dataset to protect the privacy and confidentiality of the victims treated in the center and of those now on treatment. The additional justification for removing identifying attribute values is that they have no value for the development of predictive model for surgical repair outcome.

Ethical clearance is obtained from the research and ethics committee of the School of Public Health of Addis Ababa University to carry out the study and analyze the dataset.

1.8. Plan of Dissemination of Findings from the Research

Findings of the research will be presented through annual students and staff research conference in the Addis Ababa University for Academia, and hard copy will be sent to Addis Ababa Fistula Hospital. The report will be placed in the libraries of the University for those who are interested in the area to make further investigation and for reference purpose.

1.9. Organization of the Thesis

This thesis is organized in six chapters; the first is concerned with introducing the concept of data mining, the reasons which gave rise to the wide acceptance of the discipline, introducing obstetric fistula and surgical repair outcomes for urinary fistula, statement of the problem, objectives of the study, research methodology, significance of the study and scope of the study.

In the second chapter of the study has discussed data mining and knowledge discovery in databases, statistics and data mining, data mining process models (focusing on the six step hybrid KDP data mining process model used as a framework for the current study), data mining techniques and algorithms used to attain the objectives of the research, measures of association rule interestingness, classifier accuracy measures, some related works on medical and health data with the use of data mining and the need for using multiple classification algorithms and comparing their results.

The third chapter is used to describe the business domain i.e. treatment to obstetric fistula victims, out of which the researcher has identified the business problem (the need for predicting surgical repair outcome of urinary fistula victims) and the description of the tool used in the study.

Chapter four explains specific activities performed to understand the data and prepare the data for experimentation. Activities performed in the study include data access, selection of attributes, descriptive statistical summaries of the selected attributes, assessing the quality of the data, managing missing values, errors and resolving inconsistencies.

In chapter five all the experiments carried out to discover association rules and to develop predictive models were explained. Multiple experiments were done by modifying the parameters of each algorithm. Experiments with Apriori algorithm is conducted to extract rules showing co-occurring values in between attributes that cumulatively predict surgical repair outcome of urinary fistula. Experiments with the use of J48, PART, Naïve Bayes,

and logistic regression were done to build models and select a model with greater area under the ROC curve. Interestingness measures such as confidence and support are used to assess the quality of association rules. The ROC curve is used to compare the predictive models built from carrying out experiments by altering parameters of same algorithm. Finally, better ones from each algorithm are compared to select the best model.

The last chapter, chapter six, shows conclusion and recommendation based on the knowledge discovered from the dataset about surgical repair outcome of urinary fistula.

CHAPTER TWO

LITERATURE REVIEW

2.1 Data Mining and Knowledge Discovery in Database (KDD)

In recent years the ever increasing accumulation of raw data in every industry has created both an opportunity and a challenge to the process of knowledge discovery. The challenge associated with the largeness of the data size is related to the limited processing capabilities of prevailing statistical tools which lead to a demand for better methods to deal with the large volume of data. But, challenges are not the only thing that large data bases have come up with, opportunities are also associated with them i.e. they possess patterns and hidden information that represent interesting and hidden knowledge (46).

As industries change their approach to discovering knowledge from the one which demands the collection of new data to the one which starts from already available data, different terms were coined to represent the approach. According to Piaketsky-Shapiro cited in (47) the phrase Knowledge Discovery in Databases (KDD) was coined at the first KDD workshop in 1989 to emphasize that knowledge is the end product of a data-driven discovery. Data mining is another popularly used term associated with KDD. Moreover, it is considered as a synonym for KDD by many people and it was defined as an automated or convenient extraction of patterns representing knowledge implicitly stored or captured in large databases, data warehouses, the Web, other massive information repositories, or data streams. On the other hand, data mining is viewed by others as single but an essential step in the larger process of knowledge discovery in databases. It is the data mining step which is concerned with the application of specific algorithms for extracting patterns from the data (48).

Cios, Pedrycz, Swiniarski and Kurgan (7) state KDD as a complete process of knowledge discovery from data, including how the data are first stored and then accessed, how

algorithms can work on massive datasets and still run efficiently, how results can be interpreted and visualized, and how the overall man-machine interaction can usefully be modeled and supported. More explicitly, the steps in the KDD process, in progressive order, include data cleaning, data integration, data selection, data transformation, data mining, pattern evaluation and knowledge presentation which include the evaluation and possible interpretation of the mined patterns to determine which patterns can be considered new knowledge (8). There are additional issues in the KDD process, such as incorporation of appropriate prior knowledge, and proper interpretation of the result of mining are essential to ensure that useful knowledge is derived from the data.

The term knowledge discovery from databases is self explanatory. It indicates the objective of approaching the database as to discover knowledge. However, the term data mining is widely accepted and in use in many industries, therefore, it is more popular than the longer term of knowledge discovery in databases. Adopting a broad view of data mining functionality Han and Kamber preferred to use the term data mining to represent the process of discovering interesting knowledge from large amounts of data stored either in databases, data warehouses, or other information repositories (8).

2.2. Data Mining and Statistical Methods

Data mining techniques explore the data so that the information gathered can be used to make decision (7). Indeed, statistical techniques and methods are also used for decision making. The nature of relationship which exists between the two is not of substitution; rather data mining is an extension of statistical techniques and methods. Therefore, it will be crucial to discuss the relative importance of data mining over and above the traditional statistical techniques.

The primary difference between classical statistical method and data mining is in the size of the dataset. As the size of data increases highly it will create challenges that may not be sufficiently solved by statistical techniques alone. Nonetheless, statistics plays a very important role in data mining: it is a necessary component in any data mining activity (6).

Large datasets contain large number of potential features which makes the number of potential combinations of features highly complex and enormous observations which challenges the capacity of traditional statistical techniques to perform the analysis in any reasonable amount of time (12). In contrast, data mining enables to deal with a dataset with huge size. Hand, Manila and Smyth (6) have also indicated that large data size to a statistician is much more less than the dataset which is considered as large by a data miner.

Another difference between traditional statistical techniques and data mining is related to the capability of data mining techniques to gainfully enhance multivariate analyses, for example, cluster analysis and regressions, which are not capable of dealing with complex interaction among input attributes (47). Altman cited in (47) indicates that logistic regression is a statistical technique particularly useful in health/medical research because of its capability to deal with events represented as binary attributes. Many events of interest in health/medical research are dichotomous for example, the presence and absence of a disease, being alive or dead, or responding to specific treatment or not. However, not all events in health /medical research cannot be dichotomous and therefore will not be answered by the uses of logistic regression alone. But because of the above mentioned capabilities, logistic regression is found useful in making predictions about client's future health condition; and for assisting clinical decision making for diagnosis (disease identification) and prognosis (health status in response to treatment) (47).

Finally, in the data mining approach, the number of rarely occurring instances or cases will be sufficiently large due to the large size of data so that over-sampling still results in a sufficiently large sample. Therefore, it allows changing the focus from the identification of significant risk factors to the prediction of diagnosis or prognosis, and to make use of the discovered knowledge i.e. to apply the model for predicting the outcome of newer cases (45). Thus, the inapplicability of statistical models due to its non representativeness to rarely occurring instances can be avoided by the ability of data mining techniques to deal with large data size having enough number of rare cases.

2.3. Knowledge Discovery Process Models

Since knowledge discovery is explained as a process, a thorough understanding of those sequences of steps that should be followed to discover knowledge in data is necessary before one goes into extracting knowledge from data. There are many process models designed to provide a roadmap to follow while planning and executing knowledge discovery projects. Effective use of these models helps to save cost and time, enables to better understand the project and leads to acceptable results (7). In general, these process models are systematic approaches essential for successful data mining (48).

2.3.1. Academic Research Models and Industrial Models

According to Cios, Pedrycz, Swiniarski and Kurgan (7) knowledge discovery process models are classified based on the area of application; into those that are not concerned with industrial issue (the academic research models) and those that take industrial issues into account (industrial models). These authors (7) have noted that two academic research models; one with nine steps and the other with eight steps were developed by Fayad, Piatetsky and Smyth in 1996 and Anand and Buchner in 1998 respectively. The model developed by Fayad, Piatetsky and Smyth in is perceived as the leading research model and its steps can sequentially be presented as: developing and understanding the application domain, creating a target dataset, data cleaning and preprocessing, data reduction and projection, choosing the data mining task, choosing the data mining algorithm, data mining, interpreting mined patterns and at last consolidating discovered knowledge (7).

Soon after academic research models were developed, the industrial models followed. Industrial models range from those developed by individuals to those proposed by large industrial consortiums. The two representative industrial models are the five-step model by Cabena et al., and the industrial six-step CRISP-DM model that is developed by a large consortium of European companies which has become the leading industrial model. The latter model is known for its successful and extensive application in the industrial, real-world knowledge discovery experience. The phases in CRISP-DM can be

sequentially put as; business understanding phase, data understanding phase, data preparation phase, modeling phase, evaluation phase, deployment phase (7).

There are a range of similarities between these process models. First, the processes models consist of multiple steps that are executed in a sequence and the steps are having defined inputs and outputs. Second, the initiation of each subsequent step depends upon the successful completion of the previous step. The results generated by the previous step serve as an input to its successor. Another feature that these process models share in common is the range of activities covered, which stretches from the task of understanding the project domain and data, through data preparation and analysis, to evaluation, understanding, and application of the generated results or knowledge. All the process models also are characterized by their iterative nature, many feedback loops are triggered and a revision is conducted in steps where going back to the previous phase is deemed necessary. The output at the end of the process is the generated new knowledge which is usually expressed in terms of rules, patterns, classification models, associations, trends, mathematical equations, etc. (7).

The main differences between the models described above lie in the number and scope of their specific steps. Both the above discussed models are not used as a framework to guide this current research due to the drawbacks associated with each of the methods. The nine step knowledge discovery process model indicated as suggested by Fayad, Piatetsky and Smyth for academic research is criticized for the lack of the description of business aspects (lack of the reasons why these technical data analysis tasks are carried out) and also for the lack of specific instances in which loop back is needed (7), while the CRISP-DM model is developed with in industrial circumstances with the objective of cutting costs and saving project time. Moreover, feedback loops in CRISP-DM process model exist between business understanding phase and data understanding phase; between data preparation phase and modeling phase; and from evaluation back to the business understanding (12, 48).

2.3.2. Hybrid Model

Improvements in academic and industrial models led to the creation of hybrid models, i.e., models that combine strong characteristics of both. One such model developed by Cios, Pedrycz, Swiniarski and Kurgan (7) is the six-step KDP model. It was primarily developed based on the CRISP-DM model by adopting it to academic research. The main differences in between these two models lie in the fact that the six step hybrid model provides more general, research-oriented description of the steps and introduced a data mining step instead of the modeling step. Additionally, several new feedback mechanisms were introduced in the six step hybrid model while the CRISP-DM model has only three major feedback sources but the new feedback mechanisms incorporated in the six step hybrid model are as important as the three already existing in the CRISP-DM. Finally, modification of the last step in the hybrid model differentiates it from CRISP-DM. The modification of the last step refers to extending the possibility of the application of the knowledge discovered for a particular domain to other domains (7, 49). Steps in the six step hybrid process model are shown in figure 2.1.

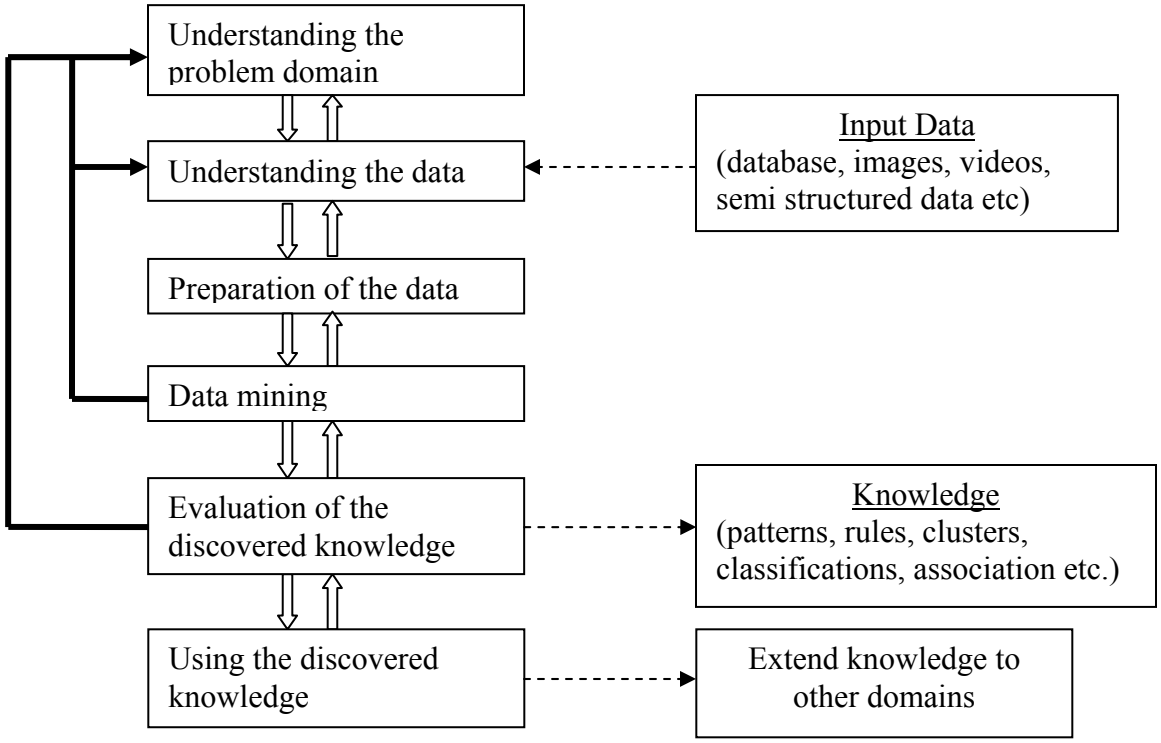

Figure 2.1. The six step hybrid KDP process model

One of the important aspects of this model is its iterative and interactive feature. The feedback loops, illustrated in figure 2.1, are necessary because any changes and decisions made in one of the steps can result in changes in subsequent steps. There are several activities and processes which can be performed in each of these steps as indicated in Table 2.1. But only those activities, methods and techniques used to achieve the objectives of this research are discussed in the subsequent chapters.

Table 2.1: Summary of tasks at each step in the six-step KDP model.

Steps	Tasks and their descriptions
Step 1: Understanding of the problem domain.	**Define the problem and determine the project goals**: working closely with domain experts, identifying key people, learning about current solutions to the problem, learning domain-specific terminology.
	Preparing description of the problem, including its restrictions/scope.
	Project goals are translated into DM goals, and initial selection of DM tools.
	Produce project plan: describe the intended plan for achieving the data mining goals and the business goals.
Step 2: Understanding of the data.	**Data Access/acquisition:** acquire the gross data that will be used as a resource to meet the data mining objective.
	Data description: examine and describe the gross properties of the dataset
	Data selection and Explore the data: select attributes based on relevance to the project goals and restrictions. Examine the distribution of values of selected attributes trough visualizations and descriptive summaries.
	Data quality assessment and verification of the usefulness of the data with respect to the DM goals: check for completeness, redundancy, missing values, plausibility of attribute values.
Step 3: Preparation of the data.	**Data integration:** collect data and attributes to a single coherent data store.
	Data cleaning: manage missing values with a suitable technique, remove or smooth noises and outliers.
	Data transformations: production of derived attributes, new instances and transformed values for existing attributes.
	Data reduction: modify the size of the dataset without changing its meaning. Reduce the dimensionality, numerousity, and the number of attributes with attribute subset selection techniques.

Steps	Tasks and their descriptions
Step 4: Data mining/ Experimentation	**Select techniques and methods**: techniques suitable to the data mining objectives are selected. Some of these are classification and numeric prediction, association, clustering.
	Selection of algorithms: multiple algorithms exist for a single type of technique. For example, classification can be done with decision trees, decision rules, Bayesian methods etc.
	Model building: calibrate parameter settings and run the model building algorithm.
	Assessing models: interpret the models/patterns according to domain knowledge and the desired level of accuracy.
Step 5: Evaluation of the discovered knowledge	**Evaluate models:** check for novelty and interestingness of the discovered knowledge and its impact. Include domain experts and/or domain knowledge.
	Review process: the entire process is revisited to identify which alternative actions could have been taken to improve the results.
	Determine the next step: decide on whether to retain approved models for use or to go back to the model building or to the first step.
Step 6: Use of the discovered knowledge.	**Planning the utilization:** plan where and how to use the discovered knowledge
	Monitoring and maintenance plan: A plan to monitor the implementation of the discovered knowledge helps avoid unnecessarily long periods of incorrect usage of data mining results.
	The entire project documented: which includes what was done well and what needs to be improved.

2.4. Knowledge Representation Models and Patterns

Variety of methods and ways are used to represent the knowledge obtained at the data mining step of knowledge discovery process models. The most common ways of knowledge representations are tables, trees, rules (classification and association), networks, equation, and graphs (clustering graphs) (2,7). These knowledge representation methods can be broadly categorized as models and patterns.

A model is a general, global description of a dataset (6, 11). The model building process is based on large sample perspective. If the rows of the data matrix are represented as p-dimensional vectors in p dimensional space, the model can make a statement about any point in this space. For example, it can be used to assign a point to a cluster or predict the

unknown value of some other attribute. Even when some of the values are missing a model typically make a statement about the row represented by the (incomplete) vector (6).

Another way of representing knowledge obtained from mining data is pattern. Cios, Pedrycz, Swiniarski and Kurgan (7) have shown that data mining is not only about finding general global model, but it is also about extracting patterns hidden in the data. In contrast to models, a pattern is a local feature of the data i.e. it describes a structure relating to a relatively small part of the data or the space in which data could occur. It holds true only for a few instances or a few attributes (or both) (6, 48). Perhaps only some of the instances behave in a certain way, and the pattern identifies and describes which they are: a pair of attributes with notably high correlation, a set of instances having extraordinarily high values on some attributes, a group of instances that always take same value on some attributes, and so on. For example, a search through a pharmaceuticals retail business database may reveal that people who buy certain combinations of drugs/items are also likely to buy others (6).

2.5. Data Mining Techniques and Algorithms

The data mining consists of many advanced techniques such as classification, numeric prediction, clustering and association with which the two practical and high level primary goals of data mining i.e. prediction and description are achieved (11).

Prediction involves using some attributes or values that potentially describe the situation to forecast outcome of a specific situation, predict the unknown or future values of other attributes of interest. On the other hand, descriptive mining focuses on finding human-interpretable patterns that describe the data (9). The basic difference between prediction and description is that prediction has unique attribute or value as its objective (for example, disease class or birth weight), while in descriptive problems no single attribute is central to the model (6).

In prediction, attributes serve two purposes for the predictive model. Those attributes that are used to make the prediction, serving as inputs to the predictive model, are commonly called as predictor or independent attribute, while the dependent or target attribute holds the values or classes that we get as an output or response from predictive model. Predictive models are built, or trained on training datasets in which the values of the dependent attribute are already known. This kind of training is referred to as supervised learning, because another dataset (test set) is prepared to compare the values calculated or estimated by the trained model with actual class values of the test set. On the other hand, descriptive techniques focus on describing data in a concise and summarative manner in order to present interesting properties of data (8). Techniques of description are sometimes considered as unsupervised learning because there is no already known result to guide the algorithms (12). Thus, descriptive techniques will consider the attributes in the dataset as single set of items which is not identified as independent and dependent.

The task of prediction itself is categorized into two based on the type of output/dependent attribute. Classification is appropriate when predicting a category or class a case falls (12), while numeric prediction (8) is used in predicting a numeric value an attribute will take. Many texts use the terms “regression” (12, 48) and “prediction-regression” (7) in spite of numeric prediction. But some other classification techniques such as backpropagation, support vector machines, and k-nearest-neighbor classifiers can also be adapted for prediction and it is preferable to use the term numeric prediction for predicting numeric values (8).

Methods of classification include decision tree induction, Bayesian classification, rule-based classification, the neural network technique of backpropagation, and many other approaches. While linear, nonlinear, and generalized linear models of regression can be used for numeric prediction. Many nonlinear problems can be converted to linear problems by performing transformations on the predictor attributes. Unlike decision trees, regression trees and model trees are also used for numeric prediction. In regression trees, each leaf stores a continuous valued prediction. In model trees, each leaf holds a

regression model (8). On the other hand, association rule discovery and clustering are unsupervised data mining methods (7) which serve the description objective (12).

It was indicated that the goals or functions of prediction and description can be achieved using a variety of data-mining methods or techniques. The methods are listed in order to acknowledge the existence of other methods outside the scope of this research. Here after, particular techniques or methods in description and prediction that are related to this research will be discussed.

2.5.1. Descriptive Methods

Descriptive methods differ in the objective that we achieve while we apply them on datasets. For example, the goal of clustering is to find groups with high internal homogeneity and high external heterogeneity. Another goal of description can be achieved by discovering the occurrence of things one after the other which is achieved by sequence discovery models (12). Han and Kamber (8) used another term "sequential pattern mining" to refer this type of frequent pattern mining in which searches are made to discover for frequent subsequences in a sequence dataset where attribute values show ordering of events. Both clustering and sequential pattern mining are among the many variants of descriptive methods that can be applied when they are found fit for data mining objectives, but they will not be discussed here because the proposed data mining objectives don't require their use. Therefore, association rule discovery alone will be discussed as there are research questions which require its use to get answered.

2.5.1.1. Association Rule Discovery

Association rule discovery or association analysis (8) is used to discover interesting relationships or dependencies hidden in large datasets. Several other terms such as affinity analysis (48); association rule mining (7, 50); association discovery (12); frequent itemset mining (8); are used to refer the task of extracting hidden patterns from an enormous data. In other words, it is concerned with finding which attributes "go together" (48) or co-occur (7). The uncovered relationship or patterns can be represented

in the form of association rule (51) which can "predict" any of the attributes, not just a specified class.

Association rules are of the form "If antecedent, then consequent." together with a measure of the support and confidence associated with the rule. Sometimes, lift and leverage are also found with these rules. Symbolically, they can be written as A => B, where A implies B i.e. the antecedent or left-hand side (LHS) implies the consequent or right-hand side (RHS) (7, 12, 48). The advantage of these forms of knowledge representations is that they are easily interpreted by humans (48).

Basic Concept

Let $I = \{i_1, i_2, \ldots, i_m\}$ be a set of items or itemset. An itemset that contains k items is referred to as a k-itemset. The total number of itemsets that the transactional dataset may contain is up to 2^m-1, excluding the empty set from the collection of all subsets of I. Let D be the set of transactions (transactional dataset) where each transaction $T \subseteq I$ is associated with an identifier TID (transaction identifier) and m is the number of items (7).

2.5.1.2. Association Rule Mining Algorithms

Algorithms used for the extraction of association rules from a set of transactions in TID-itemset format mine frequent itemset either after generating candidate itemset or without generating candidate itemset. Algorithms which generate candidate itemset for mining association rules usually have two successive phases in order to result into association rules. First, they find the frequent item sets. The aim of generating frequent itemsets is to extract all sets of items from the transaction whose percentage of occurrence is greater than a certain minimum support S_{min} value. Since the data may consist of millions of transactions, and the algorithm may have to count huge number of potentially frequent (candidate) item sets to identify the frequent ones, this phase will be computationally expensive and challenging. Next, strong rules can be generated directly from the frequent item sets by taking those items whose confidence is greater than a minimum threshold

value (7, 8, 48). Confidence as a measure of strength of a rule is the percentage of transactions in which the consequent is true when the antecedent is true.

The Naive Algorithm

The naive algorithm performs association rule discovery by considering every possible combination of antecedent and consequent, evaluate the support and confidence, and discard all associations that did not satisfy the specified constraints (50). The main drawback of this algorithm is to leave an infrequent itemset in the transaction which obliges the subsequent searches of frequent itemset to test the infrequent itemset repeatedly. Given that 2^m-1 itemsets must be searched and D transactions must be scanned each time, this algorithm requires $D(2^m-1)$ tests. This number grows exponentially with the number of items, and thus for larger data size the computations would take an unacceptably long time. Since 2^m-1 is causing the problem; we need to find a way to reduce the number of tests. Thus, the itemsets that is not produced as frequent itemsets in the preceding iteration do not need to be tested again. This reasoning resulted in the development of the Apriori algorithm (7).

The Apriori Algorithm

The Apriori algorithm uses prior knowledge about an important property of frequent itemsets— hence its name. The Apriori algorithm takes advantage of the Apriori property to shrink the search space. The Apriori property states that if an itemset T is not frequent, then adding another item A to the itemset T will not make T more frequent. That is, if T is not frequent, T U A will not be frequent (48). If a given itemset is not frequent i.e. if it does not satisfy the minimum support threshold, any superset of this itemset will also be infrequent (7).

The Apriori property is an antimonotone property i.e. if a set cannot satisfy a property, all of its supersets will also fail the same test. This helpful property is used to reduce the number of itemsets (the search space) that must be searched in every subsequent search to

find frequent itemsets. The Apriori algorithm performs repeated search for frequent itemsets through the candidate itemsets, starting with 1-itemsets, through 2-itemsets, 3-itemsets and etc (7, 8). According to Cios, Pedrycz, Swiniarski and Kurgan the Apriori algorithm will follow the steps listed bellow in order to generate frequent itemsets.

– first finds all 1-itemsets
– next, finds among them a set of frequent 1-itemsets, L1
– next extends L1 to generate 2-itemsets (C_1: candidate itemsets each with 2 items)
– next finds among these 2-itemsets a set of frequent 2-itemsets, L2
– and repeats the process to obtain L3, L4, etc.

Candidate itemset generation in apriori algorithm

Candidate itemset generation is a sub process within the larger process of frequent itemset generation. Apriori algorithm makes a pass on each frequent subsequent K itemset to generate candidate itemset. In the first pass, the algorithm directly counts item occurrence to determine the frequent 1-itemsets (item sets with 1 item). A subsequent pass, say pass k_2, consists of two steps. First, the frequent itemsets L_{k-1} found in the (k–1) pass are used to generate the entire potential candidate itemsets by a process called self joining. A self join is performed when frequent itemsets in L_{k-1} are joined with other frequent itemsets in L_{k-1} to result into the potential candidate itemsets. Then, the next step deletes or prunes all those itemsets from the join result (the potential candidate itemsets) that is not in L_{k-1}, yielding the actual candidate itemsets C_k (11).

Generating Strong Association Rules from Frequent Itemsets

Generating strong association rules from frequent itemsets is the second phase in association rule discovery. The association-rule mining algorithm generates strong rules by extracting rules that satisfy both minimum support and minimum confidence. The minimum support level is guaranteed by generating frequent itemsets. Thus, we need only to prune those rules that do not satisfy the minimum confidence (7, 11).

The confidence can be defined based on the corresponding support values as follows: Confidence (A⇒B) = P(B/A) = support-count (AnB)/support-count (A). Where support-count (AnB) is the number of transactions in D containing the itemset AnB, and support-count (A) is the number of transactions in D containing the itemset A (7).

2.5.1.3. Measures of Association Rule Interestingness

Association rule mining algorithms are concerned with extracting dependencies between any attribute, not just relationship or dependency with specific number of classes which gives them the ability to predict combinations of attributes too. All the rules extracted are not intended to be used together as a single set as in the case of classification and numeric-prediction rules, rather association rules express different regularities that underlie the dataset, and they generally predict different things. Therefore, interest is restricted to those that apply to a reasonably large number of instances (i.e. those rules with high support) and have a reasonably high accuracy (confidence) on the instances to which they apply to (2, 8). But later on, numerous alternative measures of interestingness were developed due to the realization that association rule discovery techniques which require satisfying only user specified constraints such as support and confidence resulted in the identification of many thousands of association rules. Clearly, it is desirable to reduce the numbers of rules so that only the most interesting rules are identified. Thus, measures such as lift or leverage can be used to further constrain the set of associations discovered by setting a minimum value below which associations are discarded (50). The measures of interestingness discussed here are considered as objective measures since they can be applied independently of a particular application. Interestingness of a rule can also be measured subjectively i.e., based on the specific information needs of a user in a specific context (8, 51).

Support: which is also called as coverage (2) or frequency, support count, or count of the itemset (8) of an association rule is the proportion or percentage of transactions that contain the antecedent (8, 49, 52). In other words, support tells how many examples

(instances) from a dataset that was used to generate the rule that include items from LHS (12). It is defined as ratio of the number of transactions containing A to the total number of transactions (the probability of A in D) (7).

$$\text{Support } (A \Rightarrow B) = P(A \cup B) = \frac{\text{Number of transactions containing A}}{\text{Total number of transactions}}$$

Confidence: which is also called as accuracy (2) indicates the strength of implication in the rule (7), how often the consequent is true when the antecedent is true. Confidence is the percentage of transactions in D containing A that also contain B (8, 48, 11). It is defined as ratio of the number of transactions containing A and B to the number of transactions containing A (conditional probability of B given A) (7).

$$\text{Confidence } (A \Rightarrow B) = \frac{P(A \cap B)}{P(A)}$$

2.5.2. Predictive Methods

Predictive techniques or methods focus on building a model that will permit the value of one attribute to be predicted from the known values of other attributes. It was observed that these methods could make use of two types of techniques on the bases of the type of values the designated attribute will assume. The first of these techniques used in predictive methods is classification which is appropriate when designated attribute is categorical. Numerical prediction (often called regression) is another method in which a model is built to predict a numeric value (52). Here, the term "prediction" in predictive methods is used in a general sense and no notion of a time continuum is implied.

In classification or numerical prediction model building process, a given dataset is divided into training and test sets. First, the training set is analyzed by a classification/numeric prediction algorithms and the classifier or learner model is built. Then, test set is used in estimating the accuracy of the model built. Finally, the learner

model is represented in the form of classification rules, decision trees or mathematical formulae together with various performance measures showing its ability to correctly classify new instances (7, 8).

An alternative way to splitting the dataset to train and test set for the purpose of estimating the predictive accuracy of a classifier is the K-fold cross-validation which splits the data D in m approximately equal parts $D_1, \ldots, D_m$. Training set D_{ti} is obtained by removing part D_i from D. Typical values for m are 5, 10 and 20 (47). Stratified sampling could also be used to make each part D_i a more representative of the full dataset creating stratified K-fold cross validation. If K is 10 it will be stratified 10-fold cross validation. In stratified 10-fold cross-validation, the data is divided randomly into 10 parts in which the class is represented in approximately the same proportions as in the full dataset. Each part is held out in turn and the learning algorithm is trained on the remaining nine-tenths; then its error rate is calculated on the holdout set. Thus the learning procedure is executed a total of 10 times on different training sets. Finally, the 10 error estimates are averaged to yield an overall error estimate (2).

2.5.2.1. Classification

Classification methods aim to identify the characteristics that indicate the group to which each case belongs. This pattern can be used both to understand the existing data and to predict how new instances will behave. Classification techniques create classification models by examining already classified instances and inductively finding a predictive pattern (12). Only four classification techniques selected by the researcher based on different benefits are described here under.

2.5.2.1.1. Classification by Decision Tree Induction

Decision tree induction is the learning of decision trees or decision tree classifiers from class-labeled training instances. A decision tree is a flowchart-like tree structure, where each internal node (non leaf node) represents a test on an attribute, each branch represents

an outcome of the test, and each leaf node (or terminal node) holds a class label. The topmost node in a tree is the root node (8).

Decision tree classifier is popular and especially attractive due to the following multiple benefits. First, the construction of decision tree classifiers does not require any domain knowledge or providing input parameters, and therefore is appropriate for exploratory knowledge discovery. Second, decision trees can handle high dimensional data. Third, their representation of acquired knowledge in tree form is intuitive and generally easy to comprehend. Fourth, the learning and classification steps of decision tree induction are simple and fast (8).

Decision Tree Construction.

Decision tree construction algorithms consist of two successive stages: tree building and pruning. First, most decision tree construction algorithms grow the tree top down in a greedy way. Starting with the root node, the database is examined by "split selection method" for selecting the split condition at each node. The database is then partitioned and the procedure applied recursively. In the pruning stage, the tree constructed in the tree building phase is pruned to control its size, and sophisticated pruning methods select the tree in a way that minimizes prediction errors (11).

Once the decision tree is built and pruned, it is used for classification purposes as follows. Given an instance, X, for which the associated class label is unknown, the attribute values of the instance are tested against the non leaf nodes of the decision tree until we reach the leaf node or the class label. In other words, a path is traced from the root to a leaf node, which holds the class prediction for that instance (8).

Split selection method

Decision tree induction algorithms such as ID3 (Iterative Dichotomiser 3), C4.5, and CART (Classification And Regression Tree) construct decision trees in a top-down

approach by repeatedly splitting on the values of attributes. The top down approach to decision tree construction starts at selecting an attribute for the root node. This attribute must be the one which splits the dataset effectively. The process of choosing the best attribute for splitting continues at each non-leaf node giving an opportunity to every attributes that is not yet chosen in the same branch (52).

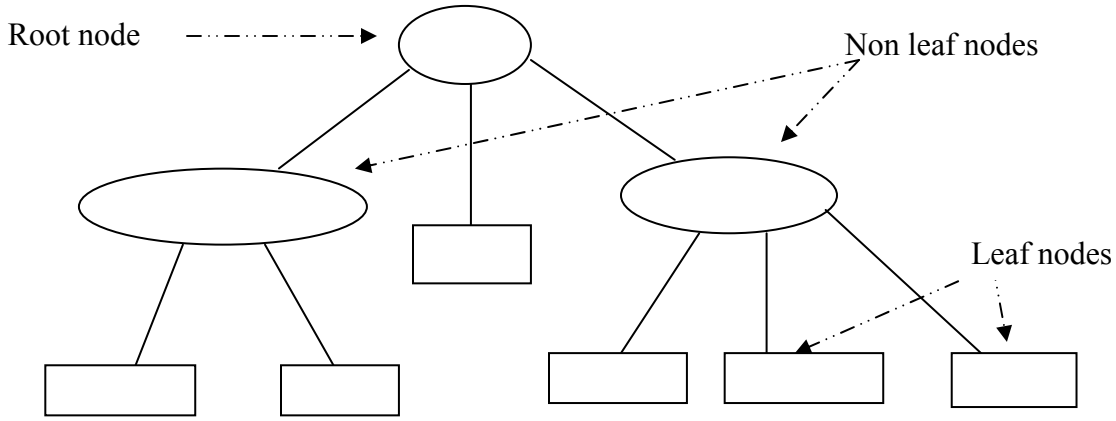

Figure 2.2: A simple decision tree

Decision tree induction algorithms differ in the measure they use for the selection of the best attribute for split. Information gain, gain ratio, and gini index are the three popular attribute selection measures used in ID3, C4.5, and CART algorithms respectively. These measures provide a ranking for each attribute describing the instances in a given training dataset. Thus, attribute having the best score for the measure is chosen as the splitting attribute at a given node. This is equivalent to saying that we want to partition on an attribute that would do the "best classification," so that the amount of information still required to finish classifying the instances is minimal (8).

Information gain and gain ratio attribute selection measures allow multi-way splits over categorical attribute (i.e., more than two branches to be grown from a node) while the gini index enforce the resulting tree to be binary (8). Particularly attributes in this study are having multiple values requiring multi-way split. Because of its limitation to deal with only binary split the gini index (found within CART) will not be discussed here. Despite the fact that both information gain and gain ratio (found in ID3 and C4.5) can be

used as a measure for split selection, ID3 requires an additional task of converting numeric attributes to nominal while C4.5 doesn't (8). Thus, how gain ratio which works inside the J48 algorithm (Weka's implementation of C4.5) is discussed here.

Gain ratio calculation

Gain ratio is computed from two values: information gain and split-info. Let us consider a transactional dataset with a total of T instances, let freq (C_i, T) stand for the number of instances in T that belong to class C_i (out of k possible classes), and let |T| denote the number of instances of set T. Information theory concept known as entropy is used to measure information gain (the expected information needed to classify a instance in dataset). The entropy of the set T (entropy (T)) with k possible classes is obtained from the relation (53):

$$\inf o(T) = -\sum_{i=1}^{k}((freq(Ci,T)/|T|).\log_2(freq(Ci,T)/|T|))$$

The next task in measuring information gain is to compute the entropy of each attribute found in the dataset. Here, we need to extend our assumption on the dataset T from partitioning it based on k possible classes only to a task of partitioning it (in to subsets T_1, T_2, …, T_n) in accordance with each attribute that assume n values. The expected information requirement (entropy), for example for attribute X can be found as the weighted sum of entropies over the subsets (53).

$$\inf o_x(T) = -\sum_{i=1}^{n}((|Ti|)/|T|).Info(Ti))$$

|Ti| denotes the number of instances with the same values for the attribute X which may be members of one or more classes.

The quantity Gain(X)=Info(T)–$Info_x$(T) measures the information that is gained by partitioning T using the attribute X. By the same analogy with which Info(T) is obtained, an additional parameter split-info is specified as:

$$Split - \inf o(X) = -\sum_{i=1}^{n}((|Ti|)/|T|).\log 2(Ti|/|T|))$$

The equation represents the potential information generated by dividing set T into n subsets. Finally, the gain ratio is obtained from gain and split information discussed by performing the following relation which expresses the proportion of information generated by the split (53).

Gain-ratio(X) = gain(X)/ Split-info(X)

Tree pruning

Decision trees that accurately model the classification in training set will be poor in classifying new cases, i.e., the decision tree is said to be over fit to the training set. With the goal of improving classification accuracy on unseen data, tree pruning attempts to identify and remove branches created from noise and outlier values (2, 8).

Tree pruning can be done in two ways: pre-pruning (or forward pruning) and post-pruning (or backward pruning). Pre-pruning halts the generation of non-significant branches i.e. deciding not to further split or partition the subset of training instances at a given node. Upon halting, the node becomes a leaf. The leaf may hold the most frequent class among the subset of instances or the probability distribution of those instances. Post-pruning, on the other hand, first generates the fully grown decision tree and then removes its non-significant branches (8, 52). One such algorithm which uses post-pruning is C4.5 and its successor C5. The C4.5/C5 algorithm first grows an over fit tree and then prunes it back to create a more stable model (48).

The removal of non-significant branches in post-pruning may be accomplished either by sub tree replacement or by sub tree raising. The idea in sub tree replacement is to select some sub tree and replace them with single leaf, basically reducing the number of tests along a certain path. This process starts from the leaves of the fully grown tree and works backward towards the root. In the case of sub tree raising a node may be moved upward towards the root of the tree, replacing the other nodes along the way (2).

2.5.2.1.2. Classification by Decision Rule Induction

Decision rule can be constructed from a decision tree simply by following a given path from the root node to any leaf. The complete set of decision rules generated from a class labeled dataset serve the same purpose as decision tree (48). Thus, decision rules are also called as classification rules (52), indicating that the rules can be used to predict the class of an unseen instance.

Rule induction algorithms generate a model as a set of rules logically ANDed together to form the rule antecedent ("IF" part) and the rule consequent ("THEN" part). The antecedent consists of the attribute values from the branches taken by particular path through the tree, while the consequent consists of the class value for the target attribute given by the particular leaf node (48). According to Witten and Frank there are two industrial-strength rule induction algorithms. But the one that works by repeatedly building partial decision trees and extracting rules from them (i.e. PART) is preferred to and used in this research because of its simplicity and its ability to achieve the same level of performance with others (2).

PART algorithm combines the divide-and-conquer strategy (the top-down approach) for decision tree construction with the separate-and-conquer approach for rule learning. The separate-and-conquer strategy first builds a rule and then removes those instances that the rule covers. These consecutive activities continue recursively for the remaining instances until none are left which generates sets of rules called 'decision lists' or ordered set of rules. On the other hand, in the partial decision tree, a pruned decision tree is built for part of the training instances, the leaf with the largest coverage is made into a rule, and the tree is discarded. Using partial decision trees in conjunction with the separate-and-conquer methodology adds flexibility and speed. A partial decision tree is an ordinary decision tree that contains branches to undefined sub trees. During the generation of such a tree, construction and pruning operations are integrated in order to find a "stable" sub tree that cannot be simplified further. Once this sub tree has been found, tree building ceases and a single rule is read off (2).

The rule sets that PART produces are as accurate as those generated by C4.5 and more accurate than other fast rule-induction methods. However, its main advantage over other schemes is simplicity. The close similarity in accuracy with C4.5 is due to the use of the C4.5 algorithm itself for building the partial decision tree whose "best" leaves are later converted into a rule (2).

Finally, both decision trees and decision rules follow same approach to deal with attributes having numeric values. First, numeric values are sorted in descending order and a binary less-than/greater-than test is considered and evaluated in exactly the same way that a binary attribute would be. A threshold i.e. a value that divides the sorted attributes into two equal parts is used to split a non-leaf node (2).

2.5.2.1.3. Bayesian Methods

Bayesian approaches employ probabilistic concept representations. The basic assumption of Bayesian reasoning is that the relation between attributes can be represented as a probability distribution. They are statistical, meaning that they predict the class membership probability rather than predicting the actual class of a particular instance (8).

Let us see here how Bayes theorem is used to compute the probability that a particular instance X belong to a specific class C. As usual, a instance is described by values of a set of n attributes. Let X be a particular instance and H be some hypothesis about its membership to specific class C. For classification objectives, we want to determine P(H/X) which is also called posteriori probability of H conditioned on X. In other words, we are looking for the probability that instance X belongs to class C, given that we know the attribute values of the instance X. In contrast, P (H) is priori probability of H. More explicitly, this is the probability that any given instance will fall into a class, regardless of any information about the values of its other attributes. The posterior probability, P(H/X), is based on more information (e.g., values of other attributes of the instance) than the prior probability, P (H), which is independent of X. Similarly, P (X/H) is the posterior probability of X conditioned on H. P(X) is the prior probability of X. P (H), P (X/H), and

P (X) may be estimated from the given data. Bayes' theorem is useful in that it provides a way of calculating the posterior probability P (H/X), from P (H), P (X/H), and P (X) (8).

Bayes' theorem is

$$P(H/X) = \frac{P(X/H)\,P(H)}{P(X)}$$

Finally, the classification of a new instance can be decided by estimating the probabilities of the classes (C_i) given the new instance (X). The class with the highest probability is the predicted class of the new instance (4).

As it was tried to show here, the goal of classification, based on Bayesian decision theory, is to classify instances based on statistical information about attributes in such a way as to minimize the probability of misclassification. The classification capability when dealing with new instances depends on prior statistical level of accuracy (or misclassification level). The accuracy with which we can predict the class of a new incoming object depends on the amount of statistical measurements gathered from the training instances. The statistical characteristics of previously seen instances can be expressed in terms of probabilities concerning the instances and their features (7).

Naive Bayes

Naive Bayesian classifier uses the Bayes' rule to compute the probability of each possible value of the target attribute given the instance, assuming the input attributes are conditionally independent given the target attribute i.e. class conditional independence. Due to the fact that this method is based on the simplistic, and rather unrealistic assumption that the causes are conditionally independent given the effect, this method is well known as Naive Bayes (2, 8). But despite the disparaging name, Naive Bayes works very well particularly when combined with some attribute selection procedure is applied to eliminate redundant (nonindependent attributes) (2).

According to Han and Kamber (8), the naive Bayesian classifier works as follows:

1. Let D be a training set of instances and their associated class labels. As usual, each instance is represented by an n-dimensional attribute vector, $X = (x_1, x_2, \ldots, x_n)$, depicting n measurements made on the instance from n attributes, respectively, $A_1, A_2, \ldots, A_n$.
2. Suppose that there are m classes, $C_1, C_2, \ldots, C_m$. Given an instance, X, the classifier will predict that X belongs to the class having the highest posterior probability, conditioned on X. That is, the naive Bayesian classifier predicts that instance X belongs to the class C_i if and only if

   $$P(C_i/X) > P(C_j/X) \text{ for } 1 \leq j \leq m;\ j \neq i$$

   Thus probability is obtained for $P(C_i/X)$. The class C_i for which $P(C_i/X)$ is maximized is called the maximum posteriori hypothesis. By Bayes' theorem

   $$P(C_i/X) = \frac{P(X/C_i)\,P(C_i)}{P(X)}$$

3. As P(X) is constant for all classes, only $P(X/C_i)\,P(C_i)$ needs to be maximized. If the class prior probabilities are not known, then it is commonly assumed that the classes are equally likely, that is, $P(C_1) = P(C_2) = \ldots = P(C_m)$, and we would therefore maximize $P(X/C_i)$. Otherwise, we maximize $P(X/C_i)\,P(C_i)$. Note that the class prior probabilities may be estimated by $P(C_i)=|C_{i,D}|/|D|$, where $|C_{i,D}|$ is the number of training instances of class C_i in D.
4. Given datasets with many attributes, it would be extremely computationally expensive to compute $P(X/C_i)$. In order to reduce computation in evaluating $P(X/C_i)$, the naive assumption of class conditional independence is made. This presumes that the values of the attributes are conditionally independent of one another, given the class label of the instance (i.e., that there is no dependence relationships among the attributes). Thus,

   $$P(X/C_i) = \prod_{k=1}^{n} p(Xk/Ci)$$
   $$= P(x_1/C_i) * P(x_2/C_i) * \ \ldots * P(x_n/C_i)$$

We can easily estimate the probabilities $P(x_1/C_i)$, $P(x_2/C_i)$, … , $P(x_n/C_i)$ from the training instances. Recall that here x_k refers to the value of attribute A_k for instance X.

5. In order to predict the class label of X, $P(X/C_i)\ P(C_i)$ is evaluated for each class C_i. The classifier predicts that the class label of instance X is the class C_i if and only if

$$P(X/C_i)\ P(C_i) > P(X/C_j)\ P(C_j) \text{ for } 1 \leq j \leq m, j \neq i$$

In other words, the predicted class label is the class C_i for which $P(X/C_i)\ P(C_i)$ is the maximum.

The Naive Bayes algorithm gives us a way of combining the prior probability and conditional probabilities in a single formula, which can be used to calculate the probability of each of the possible classifications in turn. Having done this, the class with the largest value will be selected as the class of the new instance (52).

2.5.2.1.4. Multinomial Logistic Regression

In the simplest case, numeric prediction uses standard statistical techniques such as linear regression. The linear regression method is used for modeling continuous response. Unfortunately, many real world problems are not simply linear projections of predictor values and their response values may not always be continuous rather categorical i.e. two or more. Therefore, more complex techniques, such as logistic regression and multinomial logistic regression may be necessary to predict categorical values (12).

Logistic regression models the probability of some event occurring as a linear function of a set of predictor attributes. Rather than predicting the value of the dependent attribute, the logistic regression method tries to estimate the probability P that the dependent attribute will have a given value. The actual state of the dependent attribute is determined by looking at the estimated probability. Therefore, in logistic regression, the probability P is called the success probability while 1-P denotes the failure (8, 53).

The logistic regression is used when the output attribute of the model is defined as a binary categorical. However, logistic regression supports a more general input dataset having both quantitative and categorical attributes. Suppose that output Y has two possible categorical values "Yes" and "No" coded as 1 and 0. If P (Y_i = Yes) = P_i then, P (Yi = No) = 1 – Pi (54).

Fitting the linear regression model $Y= \beta_0 + \beta_1 * X_1$ with one attribute X_1 to P will result in equation of the form:

$$P= \beta_0 + \beta_1 * X_1$$

But the equation stated above has a chance of resulting in output values of P greater than 1 and less than 0 which are not required as our effort is to get probabilities associated with getting either of the two values. To limit the output value of P between 0 and 1, first, we can eliminate the chance of getting negative values by fitting the formula to

$$P= e^{\beta 0+ \beta 1 * X1}$$

However, there is still a chance of getting output values greater than 1 from the above equation. In order to solve this final problem we can fit a model of the form

$$P= \frac{e^{\beta 0+ \beta 1 * X1}}{1+ e^{\beta 0+ \beta 1 * X1}}$$

Here, the expression on the right is also called a logistic function that cannot yield a value that is either negative or greater than 1; consequently, it restricts the estimated value of P in between 0 - 1.

The logistic regression can easily be expanded to accommodate multiple attributes so that multiple logistic regression equation is formulated as:

$$P= \frac{e^{\beta 0+ \beta 1 * X1+B2 * X2+B3 * X3+\ldots+ Bn * Xn}}{1+ e^{\beta 0+ \beta 1 * X1+B2 * X2+B3 * X3+\ldots+ Bn * Xn}}$$

Multinomial logistic regression

Multinomial logistic regression is the extension for the (binary) logistic regression when the categorical dependent outcome takes more than two values. Also referred to as logit regression, multinomial logistic regression has very similar results to binary logistic regression (48). But, unlike the binary logistic model, in which a dependent attribute has

only a binary choice (e.g., presence/ absence of a characteristic), the dependent attribute in a multinomial logistic regression model can have more than two choices that are coded categorically, and one of the categories is taken as the reference category. Let us take a dataset whose class attribute takes five values or categories coded as ‘‘0’’ the reference category, ‘‘1’’, ‘‘2’’, ‘‘3’’, ‘‘4’’. Suppose Y_i is the class attribute for a particular instance I, and the probability of being in category S (S=‘‘1’’, ‘‘2’’, ‘‘3’’, and ‘‘4’’) can be denoted by $\prod_i^{(s)} = P(Y_i=S)$ with the chosen reference category $\prod_i^{(0)}$. Then, for a simple model with one attribute X_i, a multinomial logistic regression model can be represented as (55):

$$\log\left(\frac{\prod_i^{(s)}}{\prod_i^{(0)}}\right) = \beta_0^{(s)} + \beta_1^{(s)} X_i \quad S=1,2,3,4$$

In this model, the same attribute appears in each of S categories, and a separate intercept $\beta_0^{(s)}$ and slope parameter $\beta_1^{(s)}$ are usually estimated for each contrast. The parameter $\beta_1^{(s)}$ represents the additive effect of a one-unit increase in the attribute X on the log-odds of being in category S rather than the reference category. It is more meaningful to interpret exp $\beta_1^{(s)}$ which is the multiplicative effect of a one-unit increase in X on the odds of being in category S rather than the reference category. An alternative way to interpret the effect of an attribute X is to use predicted probabilities $\prod_i^{(s)}$ for different values of X (55):

$$\prod_i^{(s)} = \frac{\exp(\beta_0^{(s)} + \beta_1^{(s)} X_i)}{1 + \sum_{k=1}^{4} \exp(\beta_0^{(s)} + \beta_1^{(s)} X_i)}$$

Then, the probability of being in the reference category ‘‘0’’ can be calculated by subtracting the sum of the probabilities of each of the other categories from one:

$$\prod_i^{(0)} = 1 - \sum_{k=1}^{4} \prod_i^{(k)}$$

2.5.2.2. Classifier Accuracy Measures

Using the same dataset to derive a classifier or predictor and then to estimate the accuracy of the resulting learned model results in misleading overoptimistic estimates due to over specialization of the learning algorithm to the data. Instead, accuracy is better

measured on a test set consisting of class-labeled instances that were not used to train the model. Then, the classifier is applied on the test set and the number of instances that were assigned to their actual classes and the number of instances that were assigned to different class by the classifier are counted, a process whose result is effectively represented by confusion matrix (8).

2.5.2.2.1. Confusion Matrix

Confusion matrix is a useful tool for analyzing how well a learned model can recognize instances of different classes. A confusion matrix for two mutually exclusive classes is shown in the figure 2.5. If there are m classes, a confusion matrix will be a table of size m by m. An entry, $CM_{i,j}$ indicates the number of instances of class i that were labeled by the learned model as class j. For a learned model to have good accuracy, ideally most of the instances would be represented along the diagonal of the confusion matrix, from entry $CM_{1,1}$ to entry $CM_{m,m}$, with the rest of the entries being close to zero (8, 56).

Let C_1 be actual class label is positive, and C_2 actual class label is negative. P+ is predicted class label is positive, and P- is predicted class label is negative. Then the matrix shown in figure 2.5 provides four entries as a result of combination of the actual class label and class label provided to instances by the classifier (8, 56).

		Predicted Class	
		P+	P-
Actual Class	C_1 +	True Positive(TP)	False Negative(FN)
	C_2 -	False Positive(FP)	True Negative(TN)

Figure 2.3: A confusion matrix for two mutually exclusive classes

The number of true positives refers to the positive instances that were correctly labeled by the learned model while true negatives are the negative instances that were correctly labeled by the learned model. The number of false positives is the negative instances that were incorrectly labeled. Similarly, false negatives are the positive instances that were incorrectly labeled (8, 56). The number of false positive and false negative are summed

up to give number of errors and help calculate the learned models/classifiers error rate represented by the equation below in (56)

$$\text{Error rate} = \frac{\text{Number of errors}}{\text{Number of Instances}} = \frac{FP + FN}{TP + TN + FP + FN}$$

On the other hand, the overall success rate is the number of correct classifications divided by the total number of classifications (2) which is commonly called as accuracy of a classifier. The accuracy of a classifier on a given test set indicates the percentage of test set instances that are correctly classified by the classifier (8).

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

A relationship between the above two equations can be constructed as Accuracy + Error rate = 1 which ultimately leads to the conclusion that the one can be obtained by subtracting the value of the other from 1. For example, Error rate =1 - accuracy (8).

Learned model performance measures that ignore correctly predicted negative instances give additional information about the ability of the model more than the information obtained from the models predictive accuracy. Measures used for the purpose discussed here are precision, recall, and F measure (56).

Precision indicates the percentage of instances classified as positives by the learned model and that are actually positives.

Precision = TP/(TP+FP)

Recall shows the percentage of actual positives which the learned model has classified as positives.

Recall = TP/(TP+FN)

Another measure called the F measure combines precision and recall with the formula (2)

$$\text{F measure} = \frac{2}{\frac{1}{\text{precision}} + \frac{1}{\text{recall}}} = \frac{2 \times \text{recall} \times \text{precision}}{\text{recall} + \text{precision}} = \frac{2 \times TP}{2 \times TP + FP + FN}$$

2.5.2.2.2. Alternative accuracy measures

Sensitivity and specificity

With slight difference, these four notations in figure 2.3 are used in medicine and health care for the purpose of characterizing the performance of diagnostic tests. For example, if a certain diagnostic test shows a positive test result for people with a disease, this is referred to as sensitivity. On the other hand, specificity refers to the proportion of people without disease who have a negative test result, which is 1 - FP. Sensitivity is also referred to as the true positive recognition rate (that is, the proportion of positive instances that are correctly identified), while specificity is the true negative rate (that is, the proportion of negative instances that are correctly identified) (2, 8, 56).

Sensitivity = TP/(TP+FN),

Specificity = TN/(TN+FP).

ROC curve

According to Altman and Bland cited in (47) the critical step before any data mining model can be used in routine clinical practice is to compare its performance with equivalent statistical methods like sensitivity and specificity. ROC (receiver operating characteristics) curves that originated from signal detection theory has added more value to these two measures by creating trade-off (11). AUC (Area Under Curve) is a measure of the area under the ROC curve.

ROC curve is a two-dimensional graph to select possibly optimal models based on the TP rate and FP rate. It also represents trade-of between benefits (TP) and costs (FP). In the ROC curve, the sensitivity (TP) rate is represented on the Y-axis and the 1-specificity (FP) rate on the X-axis. Each prediction result or one instance of a confusion matrix represents one point in the ROC space. Several points on a ROC graph should be noted. The lower left point (0, 0) represents that the classifier labeled all instances out of their actual class. The upper right point (1, 1) is the case where all instances are classified in

their actual class. The point (0, 1) represents perfect classification and the line y = x defines the strategy of randomly guessing the class. In order to assess the overall performance of a classifier, the fraction of the total area that falls under the ROC curve is considered. AUC varies between 0 and 1. Larger AUC values indicate generally better classifier performance (8).

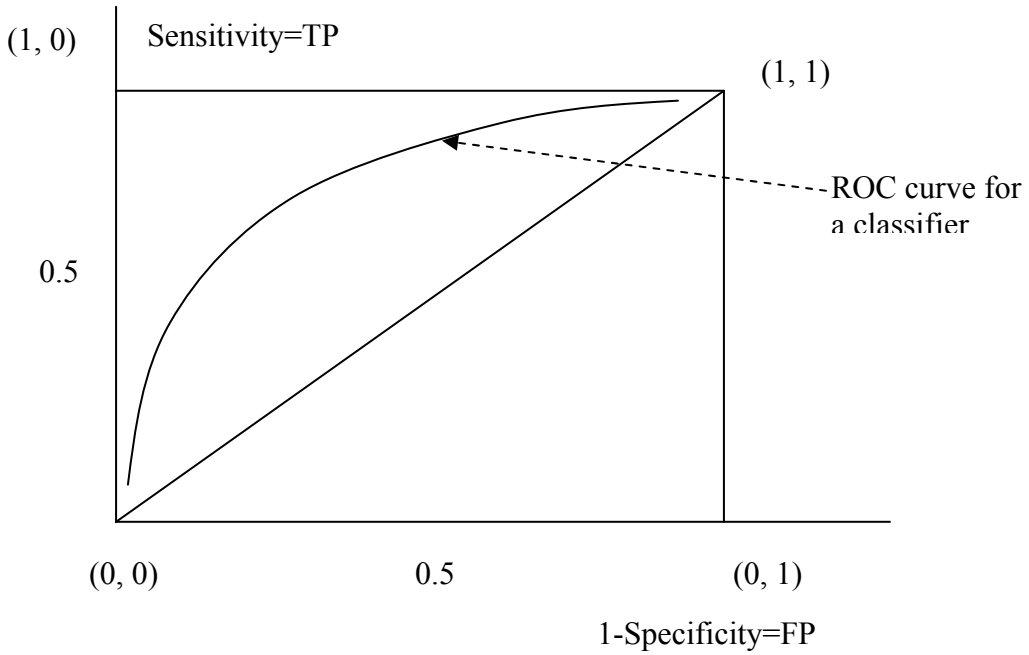

Figure 2.4: A ROC curve for a particular classifier

2.6. Related Works

2.6.1. Health Care and Medical Data Mining

According to Shortliffe and Blois (2001) cited in (45), the availability of health/medical data and information, coupled with the need to increase our knowledge and understanding of the biological, biochemical, pathological, psychosocial, and environmental processes by which health and disease are mediated, has made medicine/health particularly suitable for data mining.

Data mining methods are important in assisting clinicians to make informed decisions, thereby improving health services by extracting regularities, trends, and surprising events

which otherwise will remain hidden in these data (53). Various other authors have described similar uses of data mining in patient treatment. For example, data mining techniques have been used to discover various biological, drug discovery and patient care knowledge and patterns using selected statistical analysis, machine learning and neural networks methods (12). TCC (12) also has stated that medical applications are amongst those areas where data mining can be fruitfully used to predict the effectiveness of surgical procedures, medical tests and medications. The predictions obtained then will be used for the purpose of making decisions with regards to patient treatment.

2.6.2. Practical Application of Data Mining on Health Care Datasets

Medicine and health care are among the wide variety of fields where decision is frequently made. In addition to the frequency and urgency needed in decision making, the quality of decision in these areas affect the quality of life directly. Therefore, accurate diagnosis of disease and providing efficient treatment in a timely manner is directly associated with the decision's quality. Towards these human life saving issues, Sellappan and Awang (57) asserts that appropriate computer-based information and/or decision support systems can aid in achieving accurate clinical test results. In order to pass on quality decision, data mining has a vast potential that can be applied on large volume of data stored in databases.

Different areas in medicine and health have different decision seeking behavior and different data formats are used in these cases, the dimensionality of attributes or attributes recorded differ in type and magnitude while we move on from one sector to another. Till now, management of healthcare; customer relationship management; identification of adverse drug effect and comparison of effectiveness of drug regimens or medications; effectiveness of medical tests; prediction of health problems; effectiveness of surgical procedures are some of the areas where data mining is applied.

In health care management data mining is applied for variety of issues associated with management. Young and Pita have indicated that United Health Care has mined its treatment instance data with the objective of exploring ways to cut costs and deliver

better medicine. The same organization is mentioned as having developed clinical profiles of data mining results to give physicians' information about their practice patterns and to compare it with practices of other physicians and peer-reviewed industry standards (58). Another assurance of data mining applicability in health care management is the clinical best practices initiative of Florida Hospital launched in 1999. The goals of applying data mining in this initiative were developing a standard path of care across all campuses, clinicians, and patient admissions (59).

In relation to drugs, data mining is also found useful to identify adverse drug effects so quickly before they are identified by ordinary methods. Retrospective study conducted by the US regulatory agency, the Food and Drug Administration (FDA), found that a Bayesian statistical analysis (i.e. one algorithm in data mining) of their adverse drug event reporting database would have identified 20 out of 30 known classes of adverse drug events 1-5 years before their detection by existing methods (60). A research by Kincade (61) has shown that data mining could be utilized to compare the efficiency of different drug regimens for treatment of a particular disease and their cost effectiveness.

Prediction of future health problems was also effectively done with data mining. According to the research findings of Shantakumar and Kumaraswamy (62) that shows the extraction of significant patterns, prediction of heart disease is possible with the use of feature selection algorithms in data mining. Data mining was used in treatment comparison by analyzing which courses of action or procedure is effective so as to determine proactive steps that can reduce the risk of affliction (13, 14).

2.6.3. Running Multiple Classification Algorithms

Classification is the most widely used technique in medical data mining, and it has multiple algorithms which can do the same purpose following different internal processes. Thus, it is difficult to select in advance which algorithms will be best for a particular study; and making a "competitive evaluation" of data mining algorithms is recommended to see which performs best (5).

Different studies make use of multiple algorithms and compare performances to select the algorithm with best predictive ability. For example, a study by Gracia and Ramani utilized many classification algorithms for breast cancer detection from the breast tissue characteristics among which the Random trees (Rnd Tree) classification algorithm has given maximum predictive accuracy (63). Karpagavalli, Jamuna and Vijaya has implemented C4.5 Decision tree classifier, Naive Bayes and Multilayer Perceptrons on preoperative assessment dataset and reported that the Mutilayer perceptron classifier outperformed in predicting anesthetic risk than decision tree and Naïve Bayes methods (64).

A study by Soni, Ansari, Sharma and Soni has applied ANN, Naïve Bayes, decision tree, and KNN where the Decision Tree outperformed. The research is conducted with the objective of developing a model of higher accuracy of prediction for heart disease. This research has compared four different supervised machine learning algorithms such as: Naive Bayes, K-NN, Decision Tree algorithm, Neural network and reported that Decision Tree outperformed and some time Bayesian classification is having similar accuracy as of decision tree (65).

The above initial reviews of related literatures reveal that a wide variety of issues in the health sector are making use of the potentials of data mining. The issues for which data mining contributed can be summarized as ranging from industry and institutional level to specific cases such as treatment at individual patient level. This current research applies data mining for the improvement of decision making in the area of urinary fistula surgical repair which is performed to restore the continence of victims of obstetric fistula.

Due to the same stated reason, this study also experiments algorithms such as J48, PART, Naïve Bayes, and Multinomial Logistic Regression. However, the capability of the algorithms to work in multiclass situation and understandability of models they provide are among the additional reasons for choosing these algorithms.

CHAPTER THREE

UNDERSTANDING TREATMENT TO OBSTETRIC FISTULA VICTIMS

The process model used as a framework to guide the effort of the researcher so as to achieve the data mining objectives of this study is the six-step hybrid knowledge discovery process model. This process model is better in guiding research in academic area than the models that contributed to its development. The tasks in the first step of the process model, which is concerned with understanding the business domain was half completed during the preparation of chapter one. The current chapter describes additional tasks used for deeper understanding of the processes carried out during the treatment of obstetric fistula victims with the objective of restoring urinary continence. It is also concerned with describing the methods used in selecting the data mining tool and providing the general description of how to use the tool (7).

3.1. Methods Used in Understanding the Treatment provided to Obstetric Fistula Victims

In order to understand the problem domain the researcher has done a thorough investigation on issues that were researched so as to see a gap where the potentials of data mining could be used to fill. The researcher has reviewed studies conducted to identify most significant factors that affect surgical outcome of urinary fistula and to predict risk factors for urinary incontinence after successful surgical repair with the use of logistic regression.

In addition to what is learned from reviewing the literatures, key peoples were identified and contacted in order to further increase the knowledge about what is researched and what is not. Moreover, the efforts exerted to understand the problem domain has given an opportunity to the researcher to learn many obstetric fistula treatment related terminologies. The knowledge obtained through the above mentioned ways were used in

stating the business objectives of this study. Specifically, predicting surgical repair outcome of urinary fistula to assist the care and treatment provided to obstetric fistula victims is identified as the business objective of this study, which is then translated to the data mining objective shown in the first chapter of this thesis.

3.1.1. Obstetric Fistula Victim Treatment Process in Addis Ababa Fistula Hospital

Addis Ababa Fistula Hospital was founded in 1974 by Drs. Reginald and Catherine Hamlin both obstetrician/gynecologists with the objective of providing compassionate holistic service to women with childbirth injuries and has done so since. In addition to treatment and long term care for victims, providing short-term training to doctors, HEWs (Health Extension Workers), nurse-aide training; research, prevention of HIV/AIDS, prevention of female genital mutilation or cutting are among the areas where the hospital takes many undertakings which cumulatively improve maternal health.

The Hospital is dedicated exclusively to women with obstetric fistula and has treated over 35,000 women, free of charge, with a success/cure rate of over 95%. A detailed description of part of these cases together with the holistic service provided to the victims treated in the hospital is found stored in Microsoft's access database developed in house. The database is used to store the social and demographic data, obstetric and medical history, preoperative care data, operation date data, postoperative care related data and outcome of the treatment. From technical point of view, the actual number of attributes under these categories is too large to process with the use of traditional statistical tools which calls for another tool which has a capability of extracting knowledge that support the treatment provided by the hospital from its own database.

The process of holistic treatment provided to obstetric fistula patient in the hospital starts at registering the victim at arrival. Most of the victims arriving at the hospital have endured multiple injuries as a result of obstructed labour. Therefore, multidimensional preoperative cares, psychological counseling, treatments, medications are provided to restore the psychological and physical health of the victim which was lost as a result of

the fistula and other co-morbidities. After treating all its consequences, operation is performed in order to repair the fistula. The last course in treatment of obstetric fistula patient is intensive nursing care that follows surgical repair. At the end of this postoperative course, the woman is asked about how she feels after the surgical repair in addition to careful objective clinical assessment which both enables to get sure of the success of the surgical repair.

Under each course in the treatment process there is parallel recording of information. The information captured for each victim includes information about treatments and medications provided based on assessments taken and examinations done. Assessments taken and examinations are basic in the sense that they provide information about the type of intervention (treatment, medication, surgical approach) each specific victim's situation needs. For example, clinical assessment enables to identify the cause for incontinence for a particular victim from among the causes such as: urinary fistula (VVF), rectal fistula (RVF), both RVF and VVF, 3rd degree tear, opening of closed vagina, urinary and 3rd degree tear repair and stress thereby pointing the type of surgical repair procedure needed for the particular victim under assessment. To elaborate a bit more, for instance if an injury is a rectal fistula alone, the type of procedure will be rectal fistula repair and outcome is assessed for successful repair of rectal fistula. In the same way, if the fistula is VVF only, urinary fistula repair will be done and outcome is assessed for successful repair of urinary fistula. Measures used to assess one type of fistula cannot be applied to another type, and surgical outcomes also differ based on the type of repair needed. This is what exactly seen from the forms used for information capturing, it provides two columns with headings "surgery outcome bowel" and "surgery outcome urinary" one for each of these two types of fistulas discussed above.

VVF is the most common of all clinical presentations in different studies and case series (34, 66). Here also, a close look at of the data collected during treatment of victims helped the researcher to understand the proportion of the different types of causes for incontinence, surgically repaired in the hospital until now. Generally, majority of the cases have undergone urinary surgical repair. Based on the understanding of the need for

prioritizing and scope limitation, the current study has focused on data about VVF victims managed by urinary fistula repair.

3.1.2. Selecting and Describing the Data Mining Tool

List of criteria developed by Berry and Linhoff (46) is used to evaluate data mining tool used in the study. Table 3.1 shows the set of criterions used together with value assigned to Weka 3.6.2. Weka manual 3.6.2 (67) and Witten and Frank (2) state some of the strengths of the software which can be used to answer these questions. Therefore, after matching the identified strengths of Weka with the set of criteria, it was found that Weka 3.6.4 has passed most of them.

Table 3.1: Questions Used in Choosing the Data Mining Software Weka

No	Criteria	Value
1	What is the range of data mining techniques offered by the vendor?	Preprocessing, Association, Classification, Clustering, attribute selection
2	How scalable is the product in terms of the size of the data, the number of users, the number of fields in the data, and its use of the hardware?	Scalable
3	Does the product provide transparent access to databases and files?	Yes
4	Does the product provide multiple levels of user interfaces?	Yes
5	Does the product generate comprehensible explanations of the models it generates?	Yes
6	Does the product support graphics, visualization, and reporting tools?	Yes
7	Does the product interact well with other software in the environment, such as reporting packages, databases, and so on?	Yes
8	Can the product handle diverse data types?	Yes
9	Is the product well documented and easy to use?	Yes
10	Is support, training, and consulting there available for the product?	Yes, but not sure about training.
11	How well will the product fit into the existing computing environment?	Fit
12	Is the vendor/supplier credible?	Yes

Other freely available tools like Tanagra 1.4 and Rapid Miner 5 were also evaluated for use in the research process. The reasons that the researcher hasn't used these software are: first, Tanagra 1.4 lacks features for preprocessing (easy replacement of missing values and conversion of nominal values to binary needed during logistic regression experimentation). Second, despite the fact that Rapid Miner 5 has abundant features for data mining researches the researcher hasn't found it easy for use (No 9 of Table 3.1)

Weka (Waikato Environment for Knowledge Analysis)

Weka is a comprehensive set of advanced data mining and analysis tools. It provides a quick and easy way to explore and analyze data. Weka version 3.6.4, which is latest and stable version, was used in this research. Weka includes two executable options: command line interface (CLI) and graphical user interface (GUI) which is simpler. Weka contains tools for data pre-processing, classification, regression, clustering, association rules, attribute selection and visualization.

Weka was developed at the University of Waikato in New Zealand, and the name stands for Waikato Environment for Knowledge Analysis. The software is freely available at: http://www.cs.waikato.ac.nz/ml/weka. It is open source software released with general public license (GPL). This means that Weka is not only free to download and use but its source code is also available to be freely modified and used. The only restriction to all GPL software is that no one has the right to commercially redistribute them which works for Weka also.

The Weka Graphical User Interface (GUI) Chooser

The Weka GUI Chooser provides a starting point for launching Weka's main GUI applications and supporting tools. It includes access to the four Weka's main applications: Explorer, Experimenter, KnowledgeFlow and SimpleCLI.

A single click of the explorer button opens Weka's explorer window. The window provides graphical representations for activities like preprocessing, attribute selection, learning (association, classification, clustering), visualization. But before activities listed above are performed, the dataset needs to be imported to Weka from a file which was previously saved in one of the Weka's understandable file formats. Then the filter property has "choose" button which lists several algorithms for preprocessing data. The term filter here is used to refer to algorithms utilized for extracting information about a particular quantity from a set of unclean data (47).

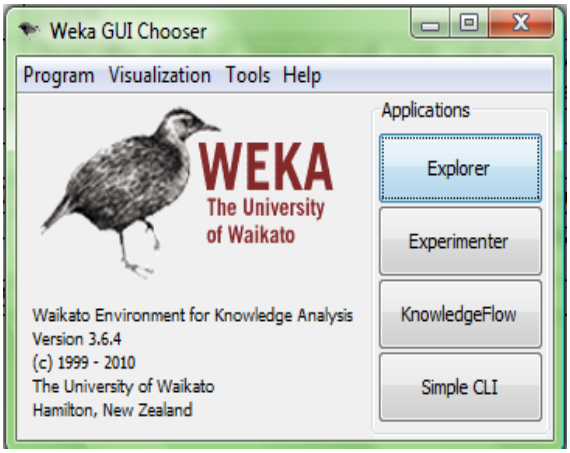

Figure 3.1: Weka GUI Chooser

Figure 3.2 shows Weka's explorer window. The window under choose button is partitioned for showing the current relation (its total number of instances and attributes); descriptive statistics for a selected attribute; listing of all attributes found on the dataset and activities that can be performed on them; and a part of window dedicated to visualize the histogram of selected attribute or a button to visualize all attributes with a new window.

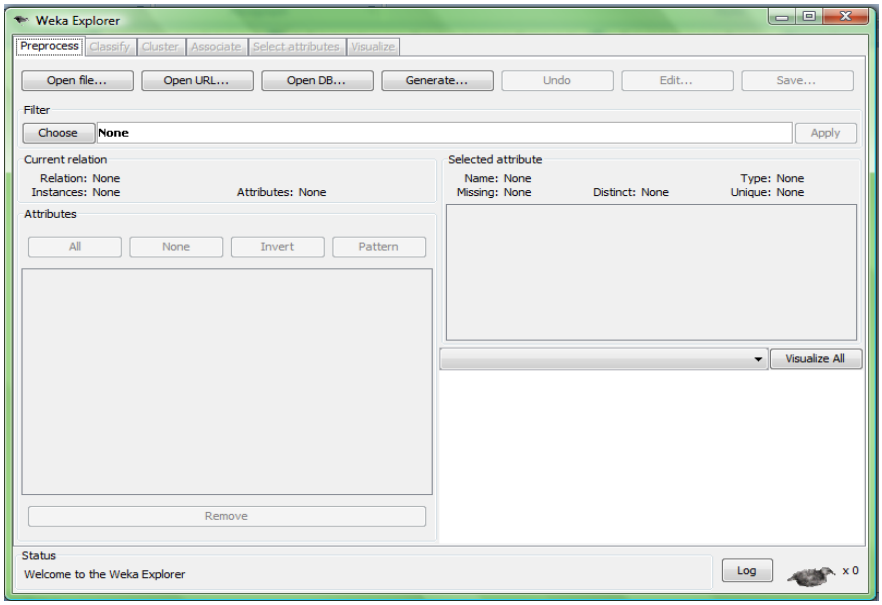

Figure 3.2: Weka's explorer window

Weka Understandable File Formats

Previous versions of Weka require the data presented in a spreadsheet or database to be converted into a Weka understandable format i.e. ARFF (Attribute Relationship File Format), which otherwise will not be opened. The current version of Weka understands many other file formats including ARFF and CSV. Thus, in this research spreadsheet files were first converted to CSV file format (less effort demanding task) and then opened by using Weka to perform tasks found in data understanding, data preparation and run the data mining algorithms.

Data understanding, data preparation/preprocessing, and data mining were performed mainly using Weka 3.6.4. Importing data from Excell to Weka is performed to make it ready for processing and discover new knowledge from the dataset with the use of different algorithms proposed to be used for the study.

CHAPTER FOUR

UNDERSTANDING AND PREPARING OBSTETRIC FISTULA VICTIMS TREATMENT DATA

4.1. Data Understanding

4.1.1. Description of the Process of Accessing the Dataset

The source of data about obstetric fistula victims' treatment was obtained from an internal application. As part of administrative tasks, the hospital has staff responsible for recording, administrating and maintaining the database. The database has attributes designed to store information on the social and demographic background of the victim who comes to the hospital seeking treatment. In addition, information on obstetric and medical history, preoperative care, and information on operation date i.e. assessment results and various interventions, and information during the post operative course are stored in access files of the database. The datasets in the access files are exported to excel files whose size amounted to 10.5 MB before any processing activity is done on it. Data found in electronic format is preferred to the manual records found on more than 35,000 victims that the hospital has treated for the past 38 years. Therefore, because of the short period of time given for the study, the study has considered only the 19,929 instances found in the access database. Removing those identifying variables such as name, religion, region, address, contact phones and surgeons list was done by the database administrator. Finally, access was obtained to analyze the dataset for the objectives specified in the thesis.

4.1.2. Data Selection

Patient identifying information such as: name of the patient, card number, region, religion; and the attribute "surgeon list" with names of surgeons who performed the repair are removed for the purpose of protecting privacy. Other attributes removed as they do not pertain to this study include: death, cause of death, was she given clothes,

was she given transport money. The 63 attributes left in the dataset were organized under five general headings such as; social and demographic variables, medical and obstetric history, preoperative care, operation date, postoperative course.

Socio demographic attributes indicate the social and demographic back ground of these women with child birth injuries. Attributes found under this general heading are serial number, age at marriage, age at causative delivery, current age, height (cm), weight (kg), parity, number of living children, days to AAFH (Addis Ababa Fistula Hospital) on foot, days to AAFH by transport, educational status, marital status, accompanying person, distance to the nearest health facility, source of information, how many days before the woman could walk.

The second groups of attributes are found under the medical and obstetric history. Values of attributes such as: antenatal care, duration of incontinence months, no of previous repairs done at other hospital, cause of fistula, other illness, duration of labour, place of delivery, mode of delivery, fetal outcome, other major illness, menstruation history, are recorded for each case.

The third groups of attributes found under preoperative care are; pre-operative stay days, antibiotic given pre-operatively, type of antibiotics, Pre-operative care provided, nerve and musculoskeletal injury.

The fourth groups of attributes are those attributes whose values are recorded during operation date. These attributes include: anesthesia, approach for urinary fistula repair, circumcision, type of procedure (repair), number of urinary fistula, type of urinary fistula (site), VVF length, VVF width, scarring, bladder size, Status of bladder neck, status of urethra, status of ureters, ureteric cateters, bladder fistula closure, graft, flaps, RVF location (rectal-injury type), RVF length, RVF width, rectal fistula closure (layers), sphincter status, intra operative complications, duration of surgery, and surgery outcome urinary, surgery outcome bowel.

Finally, information is captured during the post operative course. The attributes for recording the values during this course are: transfusion, antibiotics post operative type, pack in (days), post- operative complications, duration of bladder urethral catheters (days) and total length of stay.

4.1.2.1. Attribute Subset Selection

The major criterion for selecting an attribute set at this initial stage is to check whether each attribute is relevant to the data mining objective. Two crows corporation also ascertain that usefulness to the data mining objective is the major criteria in selecting attributes at the initial stage (12). Therefore, the literatures consulted and communications with domain experts has given the researcher the knowledge of attributes (significant risk factors) that affect surgical outcome of urinary fistula repair. Developing a model for prediction of surgical outcome of urinary fistula repair in this study emphasizes on the clinical assessments and examinations collected for VVF victims i.e. previous repairs at other hospital, type of urinary fistula, VVF length, VVF width, scarring, bladder size, status of bladder neck, status of urethra, number of fistula and status of ureters. The first eight of these aforementioned attributes are having both literature and domain expert support.

The ChiSquaredAttributeEval also ranks the attributes based on their chi-square statistics. The ability of chi-square to deal with categorical variables makes it the choice of this study because the selected attributes are all nominal valued. Appendix A shows the attributes together with the chi-square value ranked by this feature selection algorithm when the instance is urinary fistula with no other additional type. Before this feature selection algorithm is applied on the dataset all the socio demographic attributes; attributes from medical and obstetric history, preoperative care, operation, and postoperative course are removed. The study focuses on clinical assessments collected on fistula characteristics rather than taking the whole treatment process for the prediction of surgical repair outcome of urinary fistula. According to output of the feature selection algorithm used, the attributes such as: type of urinary fistula, status of bladder neck, scarring, status of urethra, bladder size, length, width, status of ureters, number of

previous repair other hospital, and number of fistula are ranked in descending order based on their chi-square value. As all the attributes are having high chi-square values indicating existence of association with the class attribute, they are all selected for analysis.

The description of selected attributes, data types they take, the unit of measure used, list of values or range of values of these attribute, are given together with statistical summaries of these attributes in data description and exploratory data analysis section. The main objective of analyzing statistical summaries of these selected attributes is to see the distribution of each value of attributes in the dataset to identify errors (noises) and discern whether there exist missing values or not.

4.1.2.2. Selection of Instances

Building a predictive model for victims of urinary fistula requires selection of instances with no additional type of fistula is identified. Thus, in addition to the removal of irrelevant attributes which was done based on the attributes irrelevance to the prediction of surgical outcome of urinary fistula repair, instances that have undergone urinary fistula surgical repair alone were selected from the database. Out of the 19929 victims, 15961 victims were affected by urinary fistula (VVF) and have undergone urinary fistula repair. Even from this number building a predictive model requires to give the learner algorithm with a training set that have all instance whose outcome or dependent attribute (class label) is not missing. Instance with missing values for outcome class are not useful for predictive model building in data mining because classification algorithms of data mining learn how instance were classified under the different classes. The classes are not existing means the algorithm learns nothing from these instance. Han and Kamber also state that records without class labels (missing or not entered) should be ignored, provided that the data mining task involves classification (8). As this study uses classification algorithms for the purpose of predictive model building, the 220 records without class information are removed from subsequent analyses. The remaining dataset will then have 15741 records whose outcomes are distributed in one of the outcome categories that are shown

in table 4.11. Thus, the statistical summaries of attributes (in tables 4.1 to 4.11) relevant to the data mining objectives are on these 15741 records.

4.1.3. Exploratory Data Analysis

In this section efforts were made to present the description of the selected attribute together with the exploratory data analysis performed with the use of frequency tables. The attribute's description, data type, unit of measure and list of values or range of values are described. With the use of frequency tables, the exploratory data analysis is performed to detect bad data i.e. attributes with the missing values and wrong entries or noises and inconsistency in values of attributes. The frequency tables for the selected attributes show the original distribution of values of attributes in instances of the dataset before any preprocessing is done on the dataset.

Number of previous repairs at other hospital: is an attribute used to show the number of previous repairs done at other hospital. It is nominal valued attribute and includes values such as 1, 2, 3, >3, not applicable, no information. Table 4.1 shows the statistical summary of these values the attribute has assumed in the dataset.

Table 4.1: Statistical summary for the number of previous repairs at other hospital as presented in AAFH database

Number of previous repairs		Frequency	Percent
Valid	>3	43	0.27
	1	465	2.95
	2	78	0.50
	3	20	0.13
	No Information	1615	10.26
	Not Applicable	12920	82.08
Missing values		600	3.81
Errors/Noises		0	0
Total		15741	100

Almost for the majority of the instances who have undergone surgical repair for urinary fistula, this attribute is not applicable. With regards to the number of missing values, 3.81% of the records have no values entered for this attribute. The attribute column is

free from errors, meaning no meaningless values are entered to the fields of these 15741 records under this attribute.

Type of urinary fistula: this attribute mainly indicates the site at which the fistula has occurred. Like the other attributes it assumes valid nominal values such as Urethral, Circumferential, Combined, Juxta-urethral, Mid Vaginal, juxta-cervical, vesico uterine, vault, uretheric, Torn urethra, Absent Urethra, No bladder, Other and no information. Table 4.2 shows the statistical summary of these values the attribute as they exist in the dataset.

Table 4.2: Statistical summary for type of urinary fistula as presented in AAFH database

Type of urinary fistula		Frequency	Percent
Valid values together with inconsistencies as a result of discrepancy in data representations	Circumferential	574	3.65
	Urethral	845	5.37
	Mid Vaginal	4968	31.56
	Juxta-urethral	2109	13.40
	Combined	2466	15.67
	Juxta-cervical	2934	18.64
	Vault	517	3.28
	torn urethra	60	0.38
	Absent urethra	94	0.60
	Vesico uterine	123	0.78
	Ureteric	125	0.79
	No bladder	2	0.01
	Torn urethra	34	0.22
	Juxta-uretral	774	4.92
	Other	9	0.06
	Visico vaginal	1	0.01
	Juxta-Urethral	13	0.08
	Juxtra-cervical	12	0.08
	Absent Urethra	3	0.02
	Mid vaginal	9	0.06
	Vesico vaginal	1	0.01

	no bladder	2	0.01
	No information	2	0.01
Missing		63	0.40
Error (>)		1	0.01
Total		15741	100.00

Despite the fact that the valid entries are listed immediately after the definition of the attribute, there are entries which mean the same but written differently. Only one error is entered mistakenly as a valid value under this attribute. Therefore, efforts are made to manually correct these values. The frequencies of the wrong values, the method and the accurate values entered to correct the mistakes, are detailed under the section that deals on the activities performed (managing missing values and errors) during data preparation.

VVF length: is a measure of fistula size which indicates the length of fistula in centi meters, takes only limited and pre-specified number of values which makes the attribute to be considered as nominal. The values of this attribute are 1, 2, 3, 4, 5, >5. Table 4.4 shows the statistical summary of these values the attribute has assumed in the dataset.

Table 4.3: Statistical summary for the distribution of VVF length as presented in AAFH database

VVF Length		**Frequency**	**Percent**
Valid	>5	651	4.14
	1	3915	24.87
	2	4606	29.27
	3	3128	19.87
	4	2027	12.88
	5	1141	7.25
Missing values		269	1.70
Errors/noises (<, 11, 22, 6)		4	0.02
Total		15741	100

VVF length measured while doing surgical repair is almost normally distributed in the treated cases. The missing values are also for small proportion of the instances that have undergone urinary fistula repair. Only four errors (0.02%) are found under the values of this attribute. Thus it required additional effort for manual correction.

VVF width: is the second measure of fistula size which indicates the width of fistula in centi meters, it takes only limited and pre-specified number of values which makes the attribute to be considered as nominal. The values of this attribute are 1, 2, 3, 4, 5, >5. Table 4.5 shows the statistical summary of these values the attribute has assumed in the dataset.

Table 4.4: Statistical summary for the distribution of VVF width as presented in AAFH database

VVF Width		Frequency	Percent
Valid	>5	624	3.96
	1	3309	21.02
	2	3727	23.68
	3	3351	21.29
	4	2436	15.48
	5	1971	12.52
Missing values		281	1.79
Errors/noises (>,>=6)		42	0.26
Total		15741	100

VVF width in cent meters is almost normally distributed in the instances. Values are missing from 281 (1.79%) instances, which is very small and cannot distort the information that the attribute provides for algorithm during model building. Only forty two errors (0.26%) are found under the values of this attribute. Thus it required additional effort for manual correction.

Scarring: is an attribute used to rank the amount of the scarring around the fistula. The values are nominal and they can be severe, mild, moderate, none, obliterated vagina. Table 4.5 shows the statistical summary of these values the attribute has assumed in the dataset.

Like most of the attributes selected for urinary fistula repair outcome predictive model building, the number (frequency) of instances are unevenly distributed for the values of scarring. The other two issues that this description should include are the number of missing values and the errors or the noises present under the attribute.

Table 4.5: Statistical summary for the distribution of the type of scaring as presented in AAFH database

Scarring		Frequency	Percent
Valid	Mild	5172	32.86
	Moderate	2661	16.90
	None	6053	38.45
	Obliterated vagina	502	3.19
	Severe	1196	7.60
Missing values		157	1.00
Errors/noises		0	0.00
Total		15741	100

The fields with missing values are only 157 (1%) of the total number of instances. Furthermore, it can be observed from table 4.5, that noises are not found under the fields of this attribute.

Bladder size: indicates the size of bladder in terms of its volume expressed with nominal values such as small, good, fair, none, no information. Table 4.6 shows the statistical summary of these values the attribute has assumed in the dataset.

Table 4.6: Statistical summary for the type of bladder size as presented in AAFH database

Bladder Size		Frequency	Percent
Valid	Fair	2148	13.65
	Good	6693	42.52
	No Information	5657	35.94
	None	61	0.39
	Small	998	6.34
Missing		184	1.16
Error/noise		0	0.00
Total		15741	100

Table 4.6 depicts that the distribution of instances among the values of the bladder size is uneven. The frequency of missing value in the dataset amounts 184 (1.16%) of the total instances while the search for errors introduced in the dataset resulted in zero value, meaning there are no errors entered for this attribute.

Status of bladder neck: is an attribute used to indicate the level of the effect of obstruction on bladder neck. The values to this attribute are complete destruction, partially damaged, intact, no information, not applicable. Table 4.8 shows the statistical summary of these values the attribute has assumed in the dataset.

The difference among the number of instances to the values of the attribute shows uneven distribution. Only 3.36% of the total number of instances has no values entered for this attribute and no errors are committed during entering values to the fields.

Table 4.7: Statistical summary for distribution of bladder status as presented in AAFH database

Status of Bladder Neck		Frequency	Percent
Valid	Complete destruction	1929	12.25
	Intact	9669	61.43
	No Information	288	1.83
	Not Applicable	1	0.01
	Partially damaged	3325	21.12
Missing values		529	3.36
Errors/noises		0	0.00
Total		15741	100

Status of urethra is an attribute used to indicate the level of the effect of obstruction on the urethra. The values to this attribute are complete destruction, partially damaged, intact, no information, not applicable. Table 4.9 shows the statistical summary of these values the attribute has assumed in the dataset.

Table 4.8: Statistical summary for distribution of status of urethra as presented in AAFH database

Status of Urethra		Frequency	Percent
Valid	Complete destruction	493	3.13
	Intact	10864	69.02
	Not Applicable	1	0.01
	Partial Damage	3672	23.33
Missing values		711	4.51
Errors/noises		0	0.00
Total		15134	100

As indicated in Table 4.8, there is no noise in the database entered under the attribute. Similar to the other attribute values small number of missing values and unbalanced distribution are the two characteristics of this attribute.

Number of fistula: is an attribute used to record the number of fistula at different sites. It is considered nominal because of values none and >3 cannot be taken as numeric. The values a particular record can assume are also pre-specified to include 1, 2, 3, >3, and none. Table 4.9 shows the statistical summary of these values the attribute has assumed in the dataset.

Table 4.9: Statistical summary for the number of fistula repaired as presented in AAFH database

Number of fistula		**Frequency**	**Percent**
Valid	>3	9	0.06
	1	14636	92.98
	2	884	5.62
	3	91	0.58
	None	3	0.02
Missing		118	0.74
Errors/noises		0	0
Total		15741	100

Table 4.9 depicts that the frequency of instances and its distribution is highly uneven for the values of the attribute. It is also observed that there are no errors or noises introduced in the dataset under this attribute. However, very small number of instances (0.74%) has no value for fields under this attribute.

Status of ureters is an attribute used to record the side the ureters are affected. It assumes one of the three nominal values such as one outside, both inside and both outside. Table 4.10 shows the statistical summary of these values the attribute has assumed in the dataset.

Table 4.10: Statistical summary for status of ureters as presented in AAFH database

Status of Ureters		Frequency	Percent
Valid	Both outside	149	0.95
	Both Inside	14825	94.18
	One Outside	652	4.14
Missing values		115	0.73
Errors/noises		0	0.00
Total		15741	100

The frequency distribution in table 4.10 shows that instances are unevenly distributed for the values of ureters. Noises are not identified under this attribute and the missing values amount only 0.73% of the total instances.

Surgical outcome of urinary fistula repair: indicates the restoration of urinary continence after surgical intervention. Valid values of this attribute are: cured, failed, stress, residual. Table 4.11 shows the statistical summary of the values the attribute has assumed in the dataset.

As it was observed in table 4.12, there exists numerous inconsistencies in entering the values of surgical outcome of urinary fistula. Values improved and Improved should be replaced by the same word as their difference is a matter of capitalization of the letter “I”. The same technique needs to be done for no chage and No Change; Stress and stress; Big residual, Residual, big residual. Values found under the same attribute such as: No cleer, NoInformation, no information, Other missed fistulaSpecify, Other Specify....OFS, OtherSpecify half cured, OtherSpecifyIleal conduit, Ureteric fistula not Managed, VVF cured but ureteric, VVF Cured but Ureteric, a total of 46 values should be removed together with their instances as they refer to wrongly entered values to the class attribute. The removal of instances with these mistakenly entered values to the class attribute decreases the 15741 instances for whom the previous tables of statistical summaries were built by 46, making the dataset to have 15695 instances.

Table 4.11: Statistical summary for surgical outcome of urinary fistula repair as presented in AAFH database

Surgical outcome of urinary fistula repair		Frequency	Percent
Values entered in the fields	Absicess draneige only	1	0.01
	Broken	2	0.01
	Cured	12320	78.27
	Failed	850	5.40
	improved	15	0.10
	Improved	97	0.62
	no chage	1	0.01
	NoChange	36	0.23
	No cleer	1	0.01
	NoInformation	2	0.01
	Other Missed fistulaSpecify	1	0.01
	Other Specify....OFS.	1	0.01
	OtherSpecify	35	0.22
	OtherSpecify half cured	1	0.01
	OtherSpecify big residual	1	0.01
	OtherSpecify Big residual	1	0.01
	OtherSpecifyIleal conduit	1	0.01
	Residual	186	1.18
	stress	4	0.03
	Stress	2182	13.86
	Ureteric fistula not Managed	1	0.01
	VVF cured but ureteric	1	0.01
	VVF Cured but Ureteric	1	0.01
Missing		0	0
Total		15741	100.00

Values such as improved and no change are subjective responses of the victims who have undergone urinary fistula repair surgery. Instances with these two values on the outcome class were removed as they create ambiguity because of lack of clear definition while compared with the valid values. For example, a victim's response as "improved" may actually be stress incontinence or residual. Likewise, a "no change" response may actually be stress, residual or failed. Therefore, an additional of 149 instances with these two types of outcome values were removed, in order to make the resulting predictive model to be able to classify outcomes on objective measures than subjective. The result of these subsequent reductions leaves only 15546 instances in the data set (i.e. 15546=15741- 46 - 149).

4.1.4. Data Quality Assessment

As good data is prerequisite for producing good models from any data mining project, the quality of the data is assessed. The inputs to this quality assessment are obtained from exploring the selected attributes with statistical summaries depicted as simple frequency distribution tables. Assessment of data quality is done to identify characteristics of the data that will negatively affect the quality of the resultant model.

The statistical summaries of the attributes have revealed that there are missing values, errors (incomplete value, wrong value, and invalid data types) and unbalanced occurrence of instances. Missing value is the feature of all the selected attributes and it ranges from the smallest number 62 (0.39%) for "type of urinary fistula" to the largest number 711 (4.52%) for "status of urethra". Errors due to invalid data type more specifically a symbol in place of nominal attribute is observed in "VVF width" and "type of urinary fistula". Errors due to inconsistencies in attribute values (capitalization of spellings, swapping of places of spellings) were found under the attribute "surgical repair outcome of urinary fistula (VVF)" and "type of urinary fistula". With regards to distribution of different values of the selected attributes, instances are unevenly distributed in almost all the selected attributes.

4.2. Data Preparation and Preprocessing

Proceeding to data mining step with low-quality data will lead to low-quality results (8). Thus, the dataset is prepared to improve the quality of the data and to finally result into good data mining models. Data preparation or preprocessing is done with the objective of cleaning the data from quality problems identified in data quality assessment.

4.2.1. Data Cleaning

The current dataset on which the study was conducted is having missing values and errors for many of the attributes. Therefore, data cleaning activities are done to clean the data by

filling in missing values, correcting noisy/error values, and resolving inconsistencies due to erroneous entries.

4.2.1.1. Managing Missing Values

Missing values refers to one or more fields of an attribute which have no value in it. The existence of many such cases makes datasets incomplete and building models of any type whether descriptive or predictive with incomplete data makes the resulting model non representative of the reality (7). As it was learnt from data understanding step, with the use of descriptive statistical summaries, all the attributes are having missing values. Thus, the missing values found under each attribute in the attributes selected for this study are replaced automatically by a feature called "ReplaceMissingValues" in Weka. "ReplaceMissingValues" replaces the mode of nominal valued attribute for missing values and the mean of continuous valued attribute for missing values. Replacing the mode or the mean is preferred method to removing an instance only because of a single missing value in on particular cell (8). Table 4.12 shows the attributes, percentage of missing values and the handling mechanism that "ReplaceMissingValues" implements.

Table 4.12: The percentage of missing values and their handling mechanism for the selected attributes.

No	Attributes	Missing values(%)	Handling mechanism
1	Previous repairs at other hospital	3.81	Replaced by the most frequent value.
2	Type of urinary fistula	0.39	Replaced by the most frequent value.
3	VVF length	1.71	Replaced by the most frequent value.
4	VVF width	1.78	Replaced by the most frequent value.
5	Scarring	0.99	Replaced by the most frequent value.
6	Bladder size	1.17	Replaced by the most frequent value.
7	Status of bladder neck	3.36	Replaced by the most frequent value.
8	Status of Urethra	4.52	Replaced by the most frequent value.
9	Number of Fistula	0.75	Replaced by the most frequent value.
10	Status of Ureters	0.73	Replaced by the most frequent value.
11	Surgical outcome of urinary fistula repair	0	Because all (220) instances with missing class information are ignored.

As shown in table 4.12, missing values found in all selected attributes are small, however, not insignificant. Therefore, these missing values are managed automatically by replacing them with the most frequent value.

4.2.1.2. Noise Correction

Noise refers to a random error mostly characterized by a deviation from valid values of the attribute. Errors make the learning time of algorithms very long and reduce quality resulting in unreliable output (8). Summary statistics with the use of frequency tables in data understanding step has revealed that three of the selected attributes have error entries. The errors for nominal valued attribute are resolved by methods used for handling missing values i.e. modal value for nominal attribute (8). First, the error values are removed manually, and then replaced by the modal value. Table 4.13, shows the attributes in which error values were detected together with how the researcher has managed them.

Table 4.13: Noises identified and corrected in the attributes selected for the study

Attributes	Errors/noises	Frequency	Handling mechanism (manual)
Type of urinary fistula	>	1	Replaced manually with the frequent value.
VVF length	<, 11, 22, 6	One for each	Replaced by the most frequent value

4.2.1.3. Resolving Inconsistencies

The two possible causes for the inconsistencies detected in the fields of selected attributes are human error in data entry and the design of the values of attributes of the database with no predefined values. The problem associated with existence of inconsistencies is that they reduce the quality of the final model and makes learning difficult for the algorithms (8).

Discrepancies were detected while extracting statistical summaries of attribute values. Despite the valid values of the attributes observed in the manual form used in actual data

collection (see Appendix F), there are invalid values entered in the database. Han and Kamber (8) also state that knowledge about the properties of the data can be used in detecting discrepancies that may exist in databases. With the help of the knowledge of the possible values that each attribute can take and the statistical summaries depicted as tables under section 4.1.3, the attribute, the identified inconsistencies due to discrepancy in data representations were shown in table 4.14 together with the data values used to replace them and maintain consistency.

Table 4.14: Inconsistencies identified and resolved in the attributes selected for the study

Attributes	Frequency	Identified Inconsistency and Handling Mechanism used
Type of urinary fistula	2	Vesico vaginal & Visico vaginal replaced manually with the frequent value (mid vaginal).
	13	Replace Juxta-Urethral with Juxta-urethral
	3	Replace Absent Urethra with Absent urethra
	2	Replace no bladder with No bladder
	12	Replace Juxtra-cervical with Juxta-cervical
	1	Replace Visico vaginal with Vesico vaginal
	60	Replace torn urethra with Torn urethra
	9	Replace Mid vaginal with Mid Vaginal
	774	Replace Juxta-uretral with Juxta-urethral
Surgical outcome of urinary fistula repair	1	no Change replaced by No Change
	15	improved replaced by Improved
	4	stress replaced by Stress
	1	No chage replaced by No Change
	2	No change replaced by No Change
	2	Broken replaced by Failed
	1	Other specify big residual replaced by Residual
	1	Other specify Big residual replaced by Residual
VVF width	41	>=6 replaced by the more general concept i.e. >5

4.2.1.4. Description of Preprocessed and Prepared Data

At the beginning of this chapter the dataset acquired as an information source is described. Since then, different activities were performed on the dataset with the objective of making it suitable for the data mining algorithms and producing representative model. Very large numbers of instances were removed and large numbers of attributes are removed. Different corrective measures were applied on the remaining attributed. The final summary of the dataset constructed ready for experiments with the use of algorithms is shown in table 4.14.

Table 4.15: Summary of the selected dataset

Categories	**Description**
Number of instances	15546
Number of attributes	11
Number of classes	4(Cured, Stress, Failed, Residual)
Size of the data	2MB

The dataset whose general description is given in the table 4.14 is ready to be imported to the data mining tool (Weka 3.6.4). After importing the CSV file to Weka with the use of "explorer" interface, the experiments explained in the next chapter are done to meet the objectives stated in the study.

CHAPTER FIVE

EXPERIMENTATION, ANALYSIS AND EVALUATION OF DISCOVERED KNOWLEDGE

Experimentation, in this study, represents the data mining step in the six step hybrid KDP model where five data mining algorithms (including the association algorithm) are applied on the dataset to achieve the objective of extracting association rules from attribute values of urinary fistula assessment and to build a model for predicting the outcome of urinary fistula surgical repair. Evaluation, on the other hand, is concerned with evaluating the result of each experiment with its own measuring criterion. This section of the study presents all the experiments together with objective measures and based on knowledge of the domain area, and expert evaluation.

The experiments conducted in this study can be categorized under association rule mining experiments and predictive model building experiments. Association rule mining experiments are carried out with the use of Apriori algorithm, specifically changing its three parameters such as minimum confidence and minimum support in order to discover values of urinary fistula assessment attributes which frequently co-occur with particular outcomes. Likewise, experiments which make use of different classification algorithms are intended to build urinary fistula surgical repair outcome predictive model of relatively better sensitivity and specificity as compared to others.

5.1. Experimental Design

The same clean dataset is used both for association rule mining and predictive model building. All the experiments that are discussed in the following subsequent sections are carried on 15546 instances and 11 attributes. The attribute set includes “previous repairs at other hospital”, “type of urinary fistula”, “VVF length”, “VVF width”, “bladder size”, “status of bladder neck”, “status of urethra”, “scarring”, “status of ureters”, “number of

fistula" and "surgical outcome of urinary fistula repair". The last attribute in the list represents the class attribute which is mandatory in developing predictive models i.e. the dependent variable in statistics lingo. This attribute is also used in association rule discovery experiments because it supports experiments used to develop a predictive model for urinary fistula surgical repair outcome.

As it was explained earlier, Apriori algorithm is used to mine rules that indicate the co-occurrence of values of attributes which are determinants of a particular urinary fistula repair outcome. The rules extracted from the dataset are then evaluated first with the use of objective measures of interestingness such as support and confidence.

In order to build predictive models for urinary fistula surgical repair outcome, four different algorithms were used. More specifically, J48, PART, naïve Bayes, and multinomial logistic regression are the algorithms with which predictive model building experiments are conducted. These algorithms split the dataset to learn a model and test its performance on a dataset prepared for a study. In 10 fold cross validation, one option in Weka for the purpose mentioned, the dataset is split into 10 equal parts. The algorithm is trained on nine–tenth of the dataset and then the classifier is tested on one-tenth. This way, the error of the resultant model will be the average of all the models found during each fold or iteration.

The algorithms used during both predictive model building and association rule mining experimentations are found in Weka 3.6.4. This version works on many file formats than its predecessors and it is compatible with CSV file format. Thus, no additional effort was exerted to change the dataset from excel to ".arff" file format which is necessary in the previous versions. The prepared dataset is saved using CSV file format. Then, this file is imported to Weka by clicking on "open" button of explorer window and browsing to the files destiny.

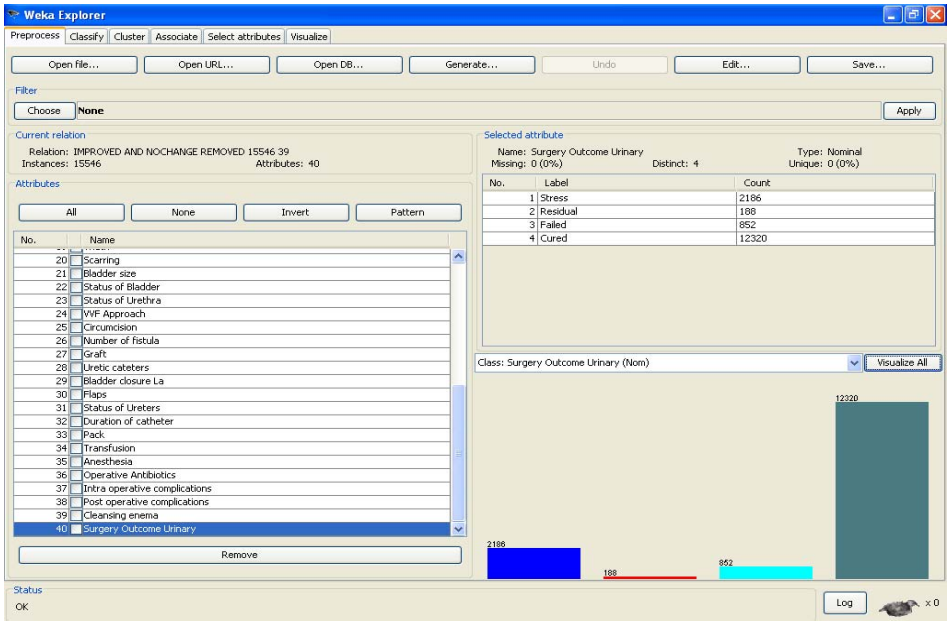

Figure 5.1: Weka 3.6.4 explorer window showing the list of attributes, the current selected attribute's statistical summaries and its graphical representation.

The "explorer window" is opened from the Weka GUI chooser "Explorer" button. The front end of Weka with obstetric fistula victims treatment dataset imported through the above activities is shown in figure 5.1. Figure 5.1 shows some important and general characteristics of the file. For example, the name of the CSV file, the number of instances, the number of attributes, the list of attributes in the dataset, the values and their frequency for a selected attribute, histogram of the selected attribute are immediately shown under "preprocess" tab.

5.2. Experimentation with Apriori Algorithm to Discover Association Rules

Association rule mining algorithm, Apriori, is used to identify attribute values co-occurring with urinary fistula surgical repair outcome. Important parameters of Apriori algorithm used by the experiments are summarized in table 5.1.

Table 5.1: Summary of Apriori Parameters.

Parameters	Description	Parameter types
CAR	If enabled, class association rules are mined instead of general association rules.	Boolean
numRules	The required number of rules	Numeric
metricType	Type of metric by which to sort rules such as confidence, lift, leverage, conviction.	Nominal
minMetric	Minimum metric score. Consider only rules with scores higher than the specified value. Minimum confidence by default is 0.9	Numeric
Delta	The delta by which the minimum support is decreased in each iteration (default: 0.05).	Numeric
lowerBoundMinSupport	The lower bound for minimum support (default: 0.1)	Numeric

The "CAR" (Classification based Association Rules) parameter is used when the rules are needed to be both classifiers and association depicters. As the current study has an objective of identifying co-occurrence of attributes with the class and with the outcome, this parameter is enabled i.e. set "True". Apriori in weka 3.6.4 starts with the upper bound support and incrementally decreases support by delta value while extracting general association rules from datasets. By default the upper bound for minimum support is set to 1.0 (100%) and the lower bound to 0.1 (10%). The algorithm stops when either the specified numbers of rules are generated, or the lower bound for minimum support is reached. Despite this automatic way of rule extraction implemented inside Apriori, the two important parameters "lower bound for minimum support" and "minMetric" is adjusted to extract as much number of rules as possible. Keeping one of these attributes and altering the other, a total of 30 experiments were done and 218 rules are extracted. These 30 experiments are conducted by increasing the minimum support from 10% to 50% and by decreasing minimum confidence from 100% to 50% on each experiment. No best rules are obtained in 8 of the experiments and the remaining 21 have resulted in 10 rules each and only one experiment resulted in less than ten rules.

Table 5.2: Number of rules (in each cell)

Minimum support	Minimum confidence					
	100%	90%	80%	70%	60%	50%
10%	0	10	10	10	10	10
20%	0	10	10	10	10	10
30%	0	0	10	10	10	10
40%	0	0	10	10	10	10
50%	0	0	8	10	10	10

In most cases, it is sufficient to focus on a combination of support and confidence to qualitatively measure the quality of the rules. However, the real value of a rule, in terms of usefulness and practicability is subjective and depends heavily on the particular domain and the business objective.

It was indicated above that the 30 experiments resulted in 218 rules. Many of the rules from these experiments are redundant. In order to elimination the redundant rules, all the rules are sorted by the antecedent. Then only one from each redundant rule is taken while the rest are removed. In this way, out of the 218 rules from the 22 experiments 218 rules are redundant of the 29 best rules shown in the table 5.2.

For ease of analyzing the discovered rules, the rules are grouped based on the antecedent and consequent. The association rules discovered show frequently co-occurring values of the predictive attributes than causal relationship.

5.2.1. Association rules by the number of fistula

The number of fistula is a characteristic of fistula which is identified by counting the number of fistulas occurred in different sites of the birth canal and bladder.

From the total of 29 best rules obtained by eliminating the redundant ones, the antecedent part of only two of the rules start by stating the number of fistula. Table 5.3 shows the rules which make the number of fistulas identified as an antecedent.

Table 5.3: Association rules by the number of fistula

Rule	Antecedent	Consequent	Conf	Supp
1	number of fistula=1	surgery outcome=cured	79%	93.08%
2	number of fistula=1, status of ureters=Both inside	surgery outcome=cured	80%	88.3%

The first rule shows coverage of 93.08% and 79% of which is surgically cured. If the rule has to include instances whose both ureters are affected from the inside (rule #2), the coverage of the rule will decrease to 88.3%. However, it seems that instances with both ureters affected have increased chance to get cured; care should be taken in interpreting these results. Comparison of these two rules based on confidence would have been possible only if their support was the same.

5.2.2. Association rules by the number of previous repairs at other hospitals

The number of repairs at other hospitals is one of the predictors of the outcomes of urinary fistula surgical repair. It indicates the number of repeated repair attempts that has been made but hasn't enabled the victim to regain complete continence.

Table 5.4: Association rules by the number of previous repairs at other hospitals

Rule	Antecedent	Consequent	Conf	Supp
1	number of previous repairs at other hospitals=not applicable	surgery outcome=cured	78%	82.08%
2	number of previous repairs at other hospitals=not applicable, number of fistula=1	surgery outcome=cured	79%	76.69%
3	number of previous repairs at other hospitals=not applicable, number of fistula=1, status of ureters=Both inside	surgery outcome=cured	79%	73.06%
4	number of previous repairs at other hospitals=not applicable, status of ureters=Both inside	surgery outcome=cured	79%	77.69%

The first of the association rules in table 5.4 shows that no previous repairs were observed on 82.08% of the instances that have undergone urinary fistula repair and 78%

of these were totally cured. The existence of only one fistula and/or both ureters affected from the inside decreases the number of instances that the rule covers in the dataset from 82.08% to 73.06% when the rule includes all the three values in its antecedent part. The coverage of the rules with the additional criterions ranged from 77.69% to 73.06% (Rules 4, 3 and 4), the confidence of the rules remained the same 79%.

5.2.3. Association rules by the status of ureters

It was discussed in the literature that obstruction of labour affects multiple organ systems, one of which is ureters. Obstruction of labour may affect only one ureter or both ureters which may still be from the inside or the outside.

Table 5.5: Association rules by the status of ureters

Rule	Antecedent	Consequent	Conf	Supp
1	status of ureters=Both inside	surgery outcome=cured	79%	94.38%

The rule extracted as interesting because of the large frequency of co-occurrence with cured surgical repair outcome is shown in table 5.5. The rule says that out of the total number of instances 94.38% are with both ureters affected from the inside. The confidence (Conf=79%) on the other hand shows 79% of the instances whose both ureters affected from the inside are totally cured after surgical repair. The remaining 21% is associated with other surgical repair outcomes.

5.2.4. Association rules by the status of urethra

Solbjorg Sjoveian, Siri Vangen, Denis Mukwege, Mathias Onsrud stated that published reports indicate the degree of involvement of urethra (status of urethra) as one of the main prognostic factors for surgical outcome (42).

Table 5.6: Association rules by the status of urethra

Rule	Antecedent	Consequent	Conf	Supp
1	Status of urethra=intact	surgery outcome=cured	86%	69.53%
2	Status of urethra=intact, number of fistula=1	surgery outcome=cured	86%	65.49%
3	Status of urethra=intact, number of fistula=1, status of ureters=Both inside	surgery outcome=cured	87%	62.27%
4	Status of urethra=intact, status of ureters=Both inside	surgery outcome=cured	86%	65.35%

Table 5.6 shows the four rules whose antecedent part starts with healthy urethra. The coverage of the first rule is 69.53%, meaning it applies for 69.53% of the instances in the dataset. Urinary fistula surgical repair outcome is cured given that urethra is healthy applies to 86 % of the total instances. The support or the number of instances that the rule covers decreases from 69.53% to 62.27%, indicating decrease in the number of instances with healthy urethra but with both ureters affected from the inside and/or with only one fistula. For example, rule #3 indicates that the number of instances which exhibit what is observed in the antecedent part are 62.27% and 87% of these has urinary fistula cured.

5.2.5. Association rules by status of bladder neck

Status of bladder neck ranks the degree of injury that the obstruction has resulted on bladder neck on a nominal scale. Clinical assessment at the outpatient or immediately before surgical repair reveals the status of the bladder neck.

Table 5.7 shows the rules extracted from the dataset that start the antecedent part by stating the Status of bladder neck as intact. The first rule has greater coverage (61.88%) as compared to the others because it only sees the status of the bladder alone. Rules 2-8 has less coverage than the rule on number one of the same table 5.7. The rule with smallest coverage is rule #7 with 53.21% where status of bladder neck, status of ureters, status of urethra and number of fistula assume values intact, both inside, intact, and one respectively.

Table 5.7: Association rules by the status of bladder neck

Rule	Antecedent	Consequent	Conf	supp
1	status of bladder neck=intact	surgery outcome=cured	87%	61.88%
2	status of bladder neck=intact, number of fistula=1	surgery outcome=cured	88%	58.29%
3	status of bladder neck=intact, status of ureters=Both inside, number of fistula=1	surgery outcome=cured	88%	55.53%
4	status of bladder neck=intact, status of ureters=Both inside	surgery outcome=cured	88%	58.27%
5	status of bladder neck=intact, Status of urethra=intact	surgery outcome=cured	88%	58.71%
6	status of bladder neck=intact, Status of urethra=intact, number of fistula=1	surgery outcome=cured	88%	55.90%
7	status of bladder neck=intact, status of ureters=Both inside, Status of urethra=intact, number of fistula=1	surgery outcome=cured	89%	53.21%
8	status of bladder neck=intact, status of ureters=Both inside, Status of urethra=intact	surgery outcome=cured	89%	55.60%

5.2.6. Association rules by scaring

Scaring refers to fibrosis or dead tissue around the fistula margins. If exists it may vary from minimal when the fistula margins are soft and mobile to extreme when the fistula margins are rigid and fixed. For fresh fistula scaring will be none.

In Table 5.8, instances with no scarring and with healthy urethra amount to 31.23%, 91% of which have got cured. In all the subsequent rules, the addition of mentioned assessments and values decreases the rules coverage. The rule with the smallest coverage (27.14%) is with no scarring and intact urethra and bladder which is shown on rule #5. However, all the rules that start the antecedent part with no scarring cover very small number of instances, most of the instances got cured after the repair for urinary fistula.

Table 5.8: Association rules by the scarring around the fistula

Rule	Antecedent	Consequent	Conf	supp
1	scaring=none, status of bladder neck=intact	surgery outcome=cured	92%	29.09%
2	scaring=none, status of bladder neck=intact, number of fistula=1	surgery outcome=cured	92%	27.88%
3	scaring=none, status of bladder neck=intact, status of ureters=Both inside,	surgery outcome=cured	92%	28.07%
4	scaring=none, status of bladder neck=intact, Status of urethra=intact	surgery outcome=cured	92%	28.17%
5	scaring=none, status of bladder neck=intact, Status of urethra=intact	surgery outcome=cured	92%	27.14%
6	scaring=none, status of bladder neck=intact, status of ureters=Both inside, Status of urethra=intact	surgery outcome=cured	92%	27.19%
7	scaring=none, Status of urethra=intact	surgery outcome=cured	91%	31.23%
8	scaring=none, Status of urethra=intact, number of fistula=1	surgery outcome=cured	91%	29.98%
9	scaring=none, status of ureters=Both inside, Status of urethra=intact, number of fistula=1	surgery outcome=cured	91%	29.08%
10	scaring=none, status of ureters=Both inside, Status of urethra=intact	surgery outcome=cured	91%	30.18%

5.3. Experimentation for Predictive Model Building

Developing a predictive model in datasets with high class imbalance and multiple classes requires some kind of countering the imbalance. Otherwise, simple comparison of models with accuracy alone may result in high predictive accuracy but low sensitivity and specificity. Figure 5.2 shows the imbalance among the outcome classes.

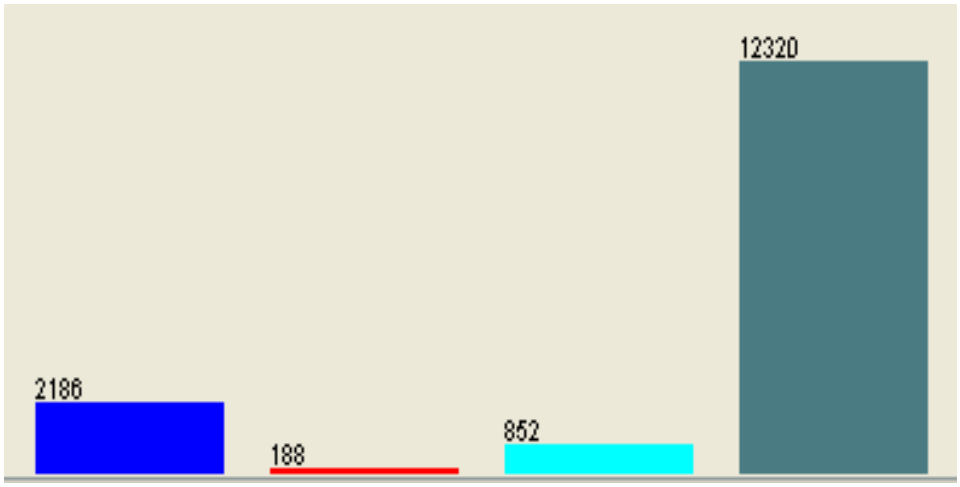

Figure 5.2: Outcome classes (Cured=12320, Stress=2186, Residual=188, Failed=852) before SMOTE is applied

A class imbalance occurs when very large numbers of instances are the members of one class i.e. majority class and very few instances are found in the other classes commonly called as the minority classes. Class imbalance provides the learner algorithm unbalanced instances, and algorithms incline to develop a model more appropriate to the majority class than the minorities. SMOTE (synthetic minority oversampling technique) is a technique to deal with this problem. As the name implies synthetic instances are generated and added to the instances in the minority class. The application of SMOTE requires close follow up of the histogram or the distribution of instances in the classes before and after SMOTE in order to decide on when to stop the SMOTE percentage (68).

5.3.1. Experimentation with J48 Algorithm

J48 is Weka's implementation of the C4.5 algorithm which can work on multiple valued attributes. As it was observed from the data description the attributes that affect surgical repair outcome of urinary fistula are multi valued. In addition to using the default parameter settings of the algorithm to build predictive model with J48, an attempt was made to find better classifier by varying its important parameters.

Table 5.7: Summary of the J48 classifier parameters

Parameters	Description	Types
binarySplits	Whether to use binary splits on nominal attributes when building trees	Boolean
confidenceFactor	The confidence factor used for pruning (smaller values incur more pruning)	Numeric
minNumObj	The minimum number of instances per leaf	Numeric
subtreeRaising	Whether to consider the subtree raising operation	Boolean
unpruned	Whether pruning is performed	Boolean

BinarySplits parameter by default is set to "False". If this value is changed to "True", it enforces the model generated to be binary decision tree rather than generalized decision tree. The confidence factor helps to set a limit so that the algorithm makes more or less pruning. The default value for confidence factor is 0.25. As shown in the table 5.7, smaller values of confidence factor enforce more pruning. The working of confidence factor requires the unpruned parameter to be set to "False". The subtreeRaising parameter is by default set to "True" to replace the nodes in a decision tree with a leaf during pruning. The fourth experiment (Exp #4) in Table 5.8 is done by setting unpruned parameter value to "True". The tree generated will represent un pruned decision tree. The first three experiments of Table 5.8 were done by setting the unprunned parameter value to "False" and decreasing the confidence factor from 0.25 to 0.10 and 0.05.

Table 5.8: Experimentation with J48 by modifying its parameters before SMOTE

Exp	Schemes	Accuracy	WTPR	WFPR	WROC
1	J48-C0.25-M2	79.16%	79.2%	77.4%	0.568
2	J48-C0.1-M2	79.24%	79.2%	79.2%	0.500
3	J48-C0.05-M2	79.24%	79.2%	79.2%	0.500
4	J48-U-M2	75.87%	75.9%	58.4%	0.665

Key: Exp=Experiment Number, Accuracy=Correctly Classified Instances, WTPR = Weighted Average TP Rate, WFPR = Weighted Average FP Rate, WROC=Weighted Average ROC Area, M = Minimum number of instances per leaf, C = Confidence factor, U=Unpruned.

Table 5.8 depicts the four experiments done by changing the important parameters of the algorithm. Each experiment is accurately expressed by a scheme that includes the name of classifier algorithm used and its corresponding parameters. Accuracy and WTPR indicate the average performance of the model in accurately classifying instances in classes of urinary fistula surgical repair outcomes. WFPR shows the average percentage of instances that the classifier classifies as positives while they are actually negative. WROC indicates the tradeoff between WTPR and WFPR.

After building four predictive models by modifying the parameters of J48, it has been observed that the performances of the models are not the same. Thus, as indicated in the methodology part based on measures of performance, an evaluation is made by comparing these models.

The first comparison is made between experiments 1, 2, and 3. The common feature of these experiments is that they all return trees by pruning. The second and the third experiments has resulted in predictive accuracy of 79.24% with 0.50 WROC shows that this experiment has very low sensitivity and specificity. Greater sensitivity and specificity among these experiments is observed in experiment one with 0.568 WROC.

The second comparison is between the unpruned model from the fourth experiment and the model from the first experiment. The model from unpruned J48 scheme has resulted in 75.87% accuracy and WROC area of 0.665. This model is better in WROC, however, not of good accuracy as compared to the model from the first experiment.

The J48 unpruned has shown better performance based on area under the ROC curve from the previous experiments. Experimentation is done using the J48 unpruned after SMOTE is applied.

Table 5.9: Experimentation with J48-U-M2 after successive SMOTEs

Exp	Schemes	SMOTE	Accuracy	WTPR	WFPR	WROC
1	J48-U-M2	100%	75.14%	75.1%	57%	0.668
2	J48-U-M2	200%	74.13%	74.1%	54.7%	0.68
3	J48-U-M2	300%	74.12%	74.1%	52.2%	0.688
4	J48-U-M2	400%	73.96%	74.09%	50.1%	0.699
5	J48-U-M2	500%	73.66%	73.7%	46.6%	0.714

As sensitivity and specificity has greater importance than general accuracy of the classifier in clinical and medical fields, models are better compared based on WROC area. But another challenge with the use of SMOTE is the question where to set the threshold. Here, the researcher has taken 300% SMOTE as the threshold because after the third experiment oversampling the minorities will lead to under sampling of previously majority classes, despite the continuous decrease in accuracy and continuous increase in WROC area. Figure 5.3 and 5.4 show class after 300% and 400% SMOTE respectively. The same increase in WROC area and decrease in accuracy is observed in all the schemes of table 5.9 after applying SMOTE (Appendix B shows the summary output the third experiment i.e after 300% SMOTE).

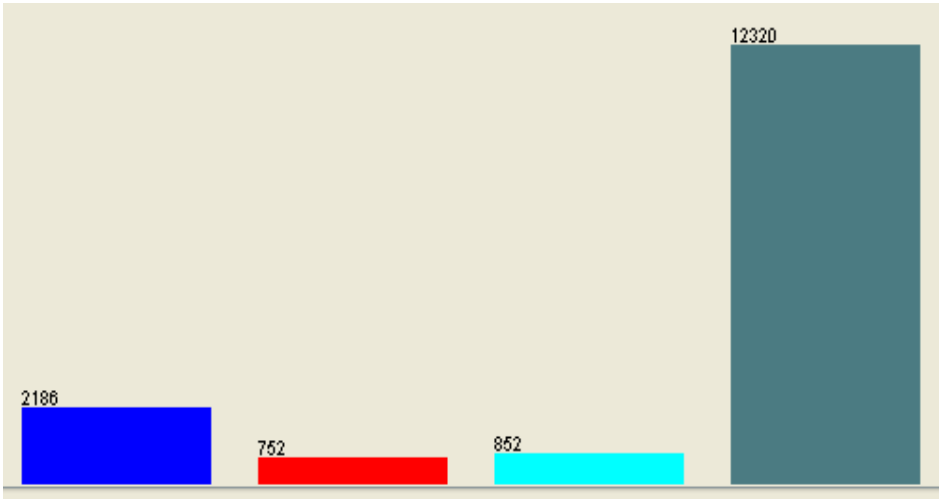

Figure 5.3 Classes after 300 SMOTE (Cured=12320, Stress=2186, Residual=752, Failed=852)

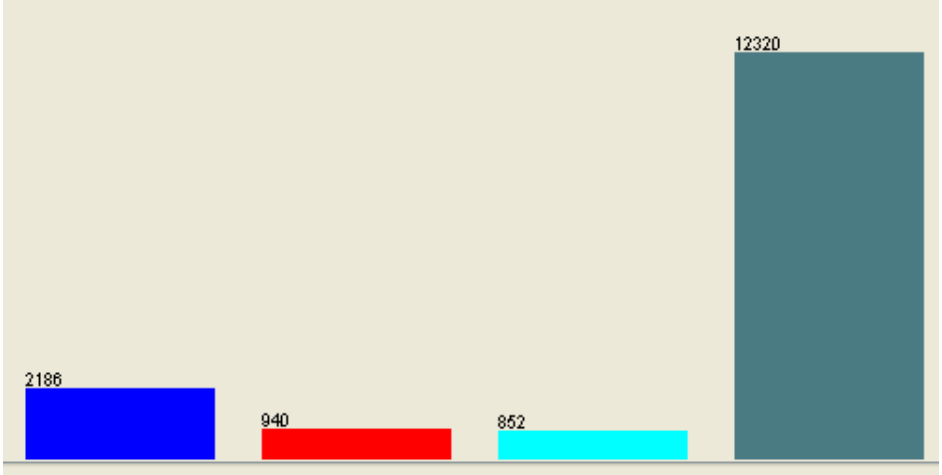

Figure 5.4 Classes after 400 SMOTE (Cured=12320, Stress=2186, Residual=940, Failed=852)

An accuracy of the model selected from experimenting J48 by applying SMOTE and altering its parameters has 74.12%. The weighted area under the ROC curve (0.688) shows the models' ability to identify the actual surgical repair outcome of a specific victim, and recognize those that are out of a specific outcome category. Therefore, experiments with the use of other algorithms are conducted to get the best model to drive relevant classification rules. The next section shows the experiments done with PART algorithm.

5.3.2. Experimentation with PART Algorithm

PART algorithm extracts rules. Due to this reason the algorithm is categorized under classification by rule induction. The detailed procedure of this algorithm in extracting rules is explained in chapter two. The algorithm builds partial decision trees and reads a path from the root of the tree to the leaf to read of a rule. The rules are ANDed together to give a complete set of rules. PART has almost a similar set of parameters with J48 algorithm that can be adjusted to build better model from datasets. Table 5.10 shows some of PART rule learner parameters.

Table 5.10: Summary of the PART rule learner parameters

Parameters	Description	Types
binarySplits	Whether to use binary splits on nominal attributes when building the partial trees	Boolean
confidenceFactor	The confidence factor used for pruning (smaller values incur more pruning)	Numeric
minNumObj	The minimum number of instances per rule.	Numeric
reducedErrorPruning	Whether reduced-error pruning is used instead of C4.5 pruning.	Boolean
unpruned	Whether pruning is performed	Boolean

The parameter "binarySplits" is left as it is by default, since many of the attributes, including the class, in the current study are multi valued. The default value of binarySplits is "False". Parameters of the algorithm adjusted by the researcher to experiment on the dataset include confidence factor, unpruned and reduced error pruning. Other parameters of the algorithm are left as they are by default.

Table 5.11: Experimentation with PART rule learner by modifying its parameters

Exp	Schemes	Accuracy	WTPR	WFPR	WROC
1	PART-M2-C0.25-Q1	78.12%	78.1%	63.9%	0.714
2	PART-M2-C0.1-Q1	78.59 %	78.6%	66.4%	0.725
3	PART-M2-C0.05-Q1	78.66 %	78.7%	68%	0.728
4	PART-U-M2-C0.25-Q1	73.65%	73.7%	59.9%	0.632
5	PART-R -M 2-N3-Q1	78.37 %	78.4%	66%	0.721

Explanations for column labels of table 5.11 can be obtained from under table 5.8 for J48. NR represents the number of rules that a particular scheme has resulted. Here, the experiments on table 5.11 are discussed and analyzed. The first experiment, Exp #1, is performed by taking the default values of all the parameters. The default value for the minimum number of instances per rule denoted by M is 2. On the other hand, the default value for confidence factor is 0.25.

The second and the third experiments were done by decreasing the confidence factor to 0.1 and 0.05. Decreasing the confidence factor enforces more pruning. The fourth experiment shows the results of setting the unprune parameter to "True" and taking the

default values of the other parameters. The last experiment is done by applying reduced error pruning i.e. setting the value of this parameter to "True".

The value assumed by each experiment to all the performance measures is different. This difference, observed among the schemes, makes comparison necessary to select the one with relatively better measures of performance. Performance measures such as accuracy, WROC and the number of rules are better in the third experiment. The third experiment is better both in accuracy and WROC area than the other algorithms. Therefore, the model from the third experiment i.e. PART-M2-C0.05-Q1 has an accuracy of 78.66%, and WROC of 0.728 which is better than the others.

Schemes discussed in Table 5.11 are experiments performed before applying SMOTE. Additional comparison among the performance measure of the classifiers from the best schemes after SMOTE has been applied shows a continuous decrease in accuracy and a continuous increase in area under the ROC curve. The results of PART-M2-C0.05-Q1 after successive SMOTEs are shown in Table 5.12.

Table 5.12: Experimentation with PART-M2-C0.05-Q1 after successive SMOTEs

Exp	Schemes	SMOTE	Accuracy	WTPR	WFPR	WROC
1	PART-M2-C0.05-Q1	100%	77.87%	77.9%	66.4%	0.729
2	PART-M2-C0.05-Q1	200%	76.95%	77%	63.8%	0.732
3	PART-M2-C0.05-Q1	300%	76.81%	76.8%	61%	0.742
4	PART-M2-C0.05-Q1	400%	76.41%	76.4%	59.3%	0.742
5	PART-M2-C0.05-Q1	500%	76.21%	76.2%	56.3%	0.75

In order to reach on final decision of selecting the best model from PART, one final comparison is made between the performances of models from among experiments after SMOTE was applied. By the same analogy applied in selecting the threshold for stopping SMOTE incrementing in experimenting the J48 unpruned, the third experiment with an accuracy of 76.81% and WROC area of 0.742 is selected from among the models built using PART algorithm. The summary statistics of PART-M2-C0.05-Q1 after 300%

SMOTE is shown on figure 5.5 and some portion of rules from this algorithm are shown on appendix E.

5.3.3. Experimentation with Naïve Bayes Algorithm

Bayesian methods are based on assumptions of probability. The naïve bayes algorithm assumes the attributes are independent. The probability of co-occurrence of an attribute value together with a particular outcome value is computed. Then, the class of a new instance will be computed by multiplying the probabilities of values the instance has assumed under each attribute. Section 2.5.2.1.3 of chapter two discusses the general procedure that the naïve Bayes algorithm follows in identifying the probabilities of the attribute values together with how the probability of a class is computed in order to predict the class of a new instance.

Table 5.13: Summary of the Naïve Bayes classifier parameter

Parameter	Description	Types
displayModelInOldFormat	Use old format for model output. The old format is better when there are many class values. The new format is better when there are fewer classes and many attributes.	Boolean

The most important parameter in relation to this study is displayModelInOldFormat. However, there are also other parameters which can be adjusted according to needs of data used in different research areas. Table 5.13 shows the description of the parameter and type of values it takes. The default value to this parameter is "False". The researcher has altered this value to "True" as displaying the model in old format is recommended to output the classifier's result for multi-valued class classification.

Table 5.14: Experimentation with Naïve Bayes classifier by modifying its parameter

Exp	Schemes	Accuracy	WTPR	WFPR	WROC
1	NaiveBayes	74.32%	73.3%	44.8%	0.753
2	NaiveBayes-O	74.32%	73.3%	44.8%	0.753

Table 5.14 shows some of the performance measures for Naïve Bayes with default values for its parameters (Exp #1) and effect of altering the value of displayModelInOldFormat to "True" on the models performance (Exp #2). Irrespective of the schemes applied, Naïve Bayes resulted in accuracy of 74.32%, and WROC area of 0.753 for both experiments.

Table 5.15: Experimentation with NaiveBayes-O after successive SMOTEs

Exp	Schemes	SMOTE	Accuracy	WTPR	WFPR	WROC
1	NaiveBayes-O	100%	74.49%	73.5%	44.7%	0.75
2	NaiveBayes-O	200%	74.32%	72.5%	44.9%	0.748
3	NaiveBayes-O	300%	71.82%	71.8%	44.8%	0.747
4	NaiveBayes-O	400%	70.8%	70.8%	44.7%	0.746
5	NaiveBayes-O	500%	69.91%	69.9%	44.3%	0.745

Table 5.15 shows experimentation extended in order to see the performance of the model from NaiveBayes-O in successive SMOTEs (100-500%). In contrary to the effect of SMOTE on models from schemes of J48 and PART, successive increase in percentage of SMOTE decreases the performance measures of model from NaiveBayes-O. After 300% SMOTE NaiveBayes-O resulted in a model having predictive accuracy of 71.82% and WROC area of 0.747. Comparing the experiments after SMOTE and before SMOTE shows that the scheme without applying SMOTE has built model with better predictive accuracy and area under the WROC curve. Thus, for the purpose of comparing the performance of this algorithm with the others, the scheme before SMOTE (NaiveBayes-O before SMOTE) is taken. (See Appendix C: for output of the summary measures of NaiveBayes-O without applying SMOTE).

5.3.4. Experimentation with Logistic Regression

In traditional statistics logistic regression is applicable only in cases where the outcome attribute is binary. In Weka, logistic regression can perform learning on a dataset with multiple outcome classes. As urinary fistula surgical repair intervention can result into more than two outcome classes, experiments were done with multinomial logistic

regression. In cases of much co-linearity in the attributes of datasets ridge estimator is used to limit the range of values that the coefficient of regression function assumes. Experiments were done as to see the effect of SMOTE on the logistic regression model that is built from the dataset. As the other algorithms used in this study, logistic regression also has parameters called ridge estimator by which the user can adjust the ridge value in the log likely hood; and the "maxIts" used to set the number of iterations that the tool performs in order to output the model.

Table 5.16: Experimentation with logistic regression by modifying its ridge parameter

Exp	Schemes	Accuracy	WTPR	WFPR	WROC
1	Logistic-R1.0E-4-M-1	79.42%	79.4%	68.4%	0.762
2	Logistic-R1.0E-5-M-1	79.42%	79.4%	68.4%	0.762
3	Logistic-R1.0E-6-M-1	79.42%	79.4%	68.4%	0.762
4	Logistic-R1.0E-7-M-1	79.42%	79.4%	68.4%	0.762
5	Logistic-R1.0E-8-M-1	79.42%	79.4%	68.4%	0.762
6	Logistic-R1.0E-9-M-1	79.42%	79.4%	68.4%	0.762
7	Logistic-R1.0E-10-M-1	79.42%	79.4%	68.4%	0.762

The experiments shown in table 5.16 were performed to develop model with a higher performance measures by incrementing the ridge parameter value from 10^{-8} up to 10^{-10} and decrementing it up to 10^{-4}. The default value for ridge parameter in logistic regression is 10^{-8}. In times of much co-linearity the very small ridge value enables to detect the coefficients of the values of each attribute. As it could be inferred from the table incrementing and decrementing the ridge parameter has never altered the performance of the model from logistic regression. All the models from logistic regression have shown 79.4% accuracy and area under the WROC curve of 0.762. Comparison among these experiments can be concluded by selecting the default scheme (Logistic-R1.0E-8-M-1).

Table 5.17: Experimentation with Logistic-R1.0E-8-M-1after successive SMOTEs

Exp	Schemes	SMOTE	Accuracy	WTPR	WFPR	WROC
1	Logistic-R1.0E-8-M-1	100%	78.50%	78.5%	67.7%	0.758
2	Logistic-R1.0E-8-M-1	200%	77.57%	77.6%	66.9%	0.753
3	Logistic-R1.0E-8-M-1	300%	76.80%	76.8%	66.3%	0.752
4	Logistic-R1.0E-8-M-1	400%	75.75%	75.8%	65.6%	0.749
5	Logistic-R1.0E-8-M-1	500%	75.02%	75%	64.9%	0.749

Like the effect of successive SMOTE observed in NaïveBayes-O, table 5.17 shows decrease in performance of the model from logistic regression when SMOTE is increased successively from 100-500%. After 300% SMOTE, model from Logistic-R1.0E-8-M-1 is having as accuracy of 76.8% and area under the WROC curve of 0.752. Comparison of measures of performances of models before and after SMOTE shows that the models before SMOTE are better in both predictive accuracy and area under the WROC curve. Thus, model from Logistic-R1.0E-8-M-1 before SMOTE is selected for further comparison with models from best schemes of the other algorithms (See Appendix D for weka explorer output window for this scheme).

5.3.5. Findings from the Classification Algorithms

The researcher has tried to experiment four algorithms namely: J48, PART, NaiveBayes, and logistic regression with the purpose of developing a model for urinary fistula surgical repair outcome. Under each algorithm multiple schemes are tested for their ability in predicting outcomes at better sensitivity and specificity which is expressed in WROC. This measure is selected as a base for comparing performances of schemes because accuracy alone is not a good measure of selecting models in medical areas. The last activity is to compare the best schemes from each algorithm with other best schemes found from other algorithms. Table 5.18 shows the comparison of best schemes from each of the algorithm used in the study.

Table 5.18: Measures of performance of models from best schemes of the different algorithms based on area under the WROC curve

Exp	Schemes	SMOTE	Accuracy	WTPR	WFPR	WROC
1	J48-U-M2	300%	74.12%	74.1%	52.2%	0.688
2	PART-M2-C0.05-Q1	300%	76.81%	76.8%	61%	0.742
3	NaiveBayes-O	No	74.32%	73.3%	44.8%	0.753
4	Logistic-R1.0E-8-M-1	No	79.42%	79.4%	68.4%	0.762

At first glance of table 5.18, it seems that logistic regression is better than the others in area under the WROC curve. Close investigation of the models based on area under the ROC curve for each outcome class as shown in table 5.19 depicts that the logistic regression is relatively insensitive to "residual" outcome for urinary fistula repair ($ROC_{Residual} = 0.669$). The same drawback is observed in Naïve Bayes-O ($ROC_{Residual} = 0.677$). However, high compromise is made in the ROC area for failed outcomes in PART-M2-C0.05-Q1 as compared to logistic regression and Naïve Bayes models, PART-M2-C0.05-Q1 with no SMOTE is highly sensitive to residual outcome than the models from logistic and Naïve Bayes. Additional comparison based on each outcome's ROC area with J48-U-M2 after 300% SMOTE shows that PART-M2-C0.05-Q1 with no SMOTE is better in all the ROC areas for the outcomes except ROC area for residual outcome. Based on these multiple reasons it could be inferred that PART-M2-C0.05-Q1 scheme after 300% SMOTE is relatively better than models from the other schemes.

Table 5.19: Area under the ROC curve for each outcome in the models which have greater weighted area under the ROC curve (WROC)

	J48-U-M2	**PART-M2-C0.05-Q1**	**NaiveBayes-O**	**Logistic-R1.0E-8-M-1**
ROC_{Stress}	0.625	0.729	0.751	0.761
$ROC_{Residual}$	0.872	0.822	0.677	0.669
ROC_{Failed}	0.548	0.656	0.725	0.726
ROC_{Cured}	0.698	0.745	0.757	0.766

Scheme: weka.classifiers.rules.PART-M2 -C0.05 -Q 1
Relation: IMPROVED AND NOCHANGE REMOVED 15546
weka.filters.unsupervised.attribute.ReplaceMissingValues-
weka.filters.supervised.instance.SMOTE-C0-K5-P300.0-S1
Instances: 16110
Attributes: 11
Number of Rules : 262
=== Summary ===
Correctly Classified Instances 12375 76.8156 %
Incorrectly Classified Instances 3735 23.1844 %
=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
	0.137	0.033	0.395	0.137	0.203	0.729	Stress
	0.295	0.011	0.565	0.295	0.388	0.822	Residual
	0.031	0.007	0.194	0.031	0.053	0.656	Failed
	0.96	0.791	0.798	0.96	0.871	0.745	Cured
Weighted Avg.	0.768	0.61	0.7	0.768	0.715	0.742	

=== Confusion Matrix ===

a	b	c	d	<-- classified as
299	54	43	1790	a = Stress
19	222	9	502	b = Residual
102	18	26	706	c = Failed
337	99	56	11828	d = Cured

Figure 5.5: Summary statistics of PART-M2-C0.05-Q1 after 300 SMOTE

Figure 5.5 shows the summary statistics of the performance of the best model the study has resulted. Therefore, model from PART-M2-C0.05-Q1 after 300% SMOTE is selected as best model and the researcher used this model to drive some relevant classification rules.

Classifier's Error

In classification or prediction tasks, the accuracy of the resulting model is measured either in terms of the percentage of instances correctly classified or in terms of "error rate" i.e. the percentage of records classified incorrectly. Classification error rate on pre classified test set is commonly used as an estimate of the expected error rate when classifying new records (46). To make the procedure valid, the 10- fold cross validation is used, so that model is built and tested 10 times. Errors during each test are averaged to

give the average error rate of the model. The classification error rate for the selected model is 23.8 %, which means the model has incorrectly classified about around 23.8% instances out of their actual classes each time when the model is tested on the test set.

The percentage of incorrectly classified instances (23.8%) indicates the chance with which the developed model misclassifies a new victim out of the actual class her future surgical repair outcome. Several reasons may be attributed for increased error rate from the models. First, algorithms differ in their capability as observed from comparisons of performance measures. Second, attributes in preoperative, operative and postoperative course that are not included in the study might have influenced it. In fact, a particular victim regains her continence not because of clinical examination and assessments rather because of the treatments and the surgical repair.

5.3.5.1. Analysis of Classification Rules from PART-M2-C0.05-Q1

PART rule learner with the specified scheme has resulted in 262 rules. Listing all the rules here will be quite cumbersome, thus, the rules which are highly predictive are selected and discussed as the finding of this study based on success ratio. The success ratio of a rule is found in parenthesis just at the end of the predictive rules. The numbers in parenthesis at the end of each rule tells the number of instances in the rule. If one or more of the rules were not pure (that is all in the same class), the number of misclassified cases also are given after slash (/). The researcher has converted these numbers into percents to compare the chance of the rule to be correct with that of its chance to be incorrect. The greater the number before the parenthesis the greater the chance of the rule to predict the class indicated by that particular rule.

5.3.5.1.1. Classification Rules Predicting Cure after Surgical Repair

The same way of interpretation of the rules can be used for the classification rules that the researcher has selected and presented in the tables hereunder. For example, rule number one in Table 5.20 shows that a new instance with (Status of Urethra = Intact AND Status of bladder neck Neck= Intact AND Scarring = None AND Length = 1) has 93.82%

chance of being cured after surgical repair and 6.17% chance of not being cured. The second rule shows that if the length increases by one, keeping the other measures the likely hood of being cured after surgical repair decreases 93.35%.

Table 5.20: Classification rules predicting cure for a surgical repair

Rule No	"IF" Part	"Then" part	Success ratio	%
1	Status of Urethra = Intact AND Status of bladder neck = Intact AND Scarring = None AND Length = 1	Cures	(1746.0/115.0)	93.82
2	Status of Urethra = Intact AND Status of bladder neck = Intact AND Scarring = None AND Length = 2	Cures	(1656.0/118.0)	93.35
3	Status of Urethra = Intact AND Scarring = Mild AND Type of urinary fistula = Juxta-cervical AND Length = 2	Cures	(312.0/10.0)	96.89
4	Status of bladder neck = Intact AND Scarring = None AND Type of urinary fistula = Vault	Cures	(45.0/1.0)	97.83
5	Status of bladder neck = Partially Damaged AND Type of urinary fistula = Juxta-cervical AND Scarring = Mild	Cures	(40.0/3.0)	93.02
6	Status of bladder neck = Intact AND Bladder size = No information AND Scarring = Moderate AND No of Prev Repair Other Hospital = No Information	Cures	(35.0/2.0)	94.59
7	Status of bladder neck = Partially Damaged AND Bladder size = Fair AND Scarring = None	Cures	(18.0/1.0)	94.74

5.3.5.1.2. Classification Rules for Predicting Stress Incontinence after Surgical Repair

Each rule in Table 5.21 should be taken independently and no form of relationship can be created among these rules. The rules can be used to situations in which a new instance assumes attributes values indicated by the rule. All the rules shown in the table work for smaller number of instances in the dataset, however, stress incontinence is observed in large number of instances for whom the rules apply.

Table 5.21: Classification rules for predicting stress incontinence after a surgical repair

Rule No	"IF" Part	"Then" part	Success ratio	%
1	Type of urinary fistula = Circumferential AND Status of Ureters = Both Inside AND Length = >5 AND Status of Urethra = Partial Damage	Stress	(7.0/1.0)	87.5
2	Length = 2 AND Type of urinary fistula = Juxta-urethral AND Scarring = Moderate AND Width = 2 AND No of Prev Repair Other Hospital = Not applicable AND Status of bladder neck = Complete Destruction	Stress	(7.0/1.0)	87.5
3	Status of bladder neck = Partially Damaged AND Type of urinary fistula = Circumferential AND Status of Ureters = Both Inside AND Length = >5 AND Status of Urethra = Partial Damage	Stress	(7.0/1.0)	87.50
4	Length = 3 AND Status of Urethra = Intact AND Width = 4	Stress	(16.0/2.0)	88.89

5.3.5.1.3. Classification Rules for Predicting Residual Incontinence Stress Incontinence after Surgical Repair

Table 5.22: Classification rules for predicting residual incontinence after a surgical repair

Rule No	"IF" Part	"Then" part	Success ratio	%
1	Type of urinary fistula = Combined AND Status of bladder neck = Partially Damaged AND Width = 3 AND Status of Urethra = Complete Destruction	Residual	(7.0/1.0)	87.50
2	Type of urinary fistula = Combined AND Status of bladder neck = Partially Damaged AND Number of fistula = 2 AND Status of Urethra = Partial Damage AND Length = 4	Residual	(8.0/1.0)	88.89

Table 5.22 shows the likelihood of occurrence of residual incontinence while the attributes mentioned by the rules take the associated values. The predictive abilities of these rules are high, despite the fact that these rules apply for small number of instances in the dataset. For example, the second rule in Table 5.22 has smaller, 11.11%, chance of misclassifying a new instance with the same values for the indicated attributes.

5.3.5.1.4. Classification rules for predicting failure after a surgical repair

Table 5.23: Classification rules for predicting failure after a surgical repair

Rule No	"IF" Part	"Then" part	Success ratio	%
1	Type of urinary fistula = Absent urethra AND Bladder size = Small	Fails	(12.0/2.0)	85.71
2	Type of urinary fistula = Combined AND Status of bladder neck = Partially Damaged AND Bladder size = Small AND Length = 5.0 AND Status of Urethra = Partial Damage	Fails	(5.0/1.0)	83.33

Table 5.23 presents only two of the rules PART after 300% SMOTE has learnt from the dataset. Like the rules for stress and residual incontinence, the total number of instances on which rules for predicting failure of surgical repair applies is small. For example, the second rule applies only for 6 instances in the dataset and the first rule applies for larger number of instances with better predictive ability; however, these rules work for completely different cases except the similarity in the value of bladder size that is small.

CHAPTER SIX

CONCLUSION AND RECOMMENDATION

6.1. Conclusion

Prediction of outcomes of urinary fistula surgical repair intervention is of paramount importance for both during surgical decision making and for special post-operative care that particular victims may require. Browning A. (43) has indicated the purpose of predicting victims who are more likely to suffer post-repair complication because of residual outcome. According to him identifying these victims can enable to tailor surgical techniques to try and decrease complication rate and to make the surgery be done by more experienced fistula surgeon. The results from predictive models could also be used in post-operative consultations with the victim who has undergone repair surgery.

Association rules are extracted from the clean dataset with the use of Apriori algorithm which showed attribute values that frequently co-occur together with specific classes. All of the rules showed that less severity of injury co-occurring more with "cured" outcome than any other outcome. The reverse of which indicates stress, residual, and failed surgical outcomes may occur in cases of higher severity of an injury. Moreover, the addition of an attribute value decreases the coverage rules indicating cured surgical outcome, which means that instances with additional injury have a decreased chance of cure than a victim with only one injury of same type.

The study has shown the necessity to experiment as many classification algorithms as possible before picking and using a single algorithm for prediction. On the way to the major objective i.e. developing predictive model, performances of models from best schemes of J48, PART, Naïve Bayes algorithms were compared with the performance of the best scheme from logistic regression. The comparison has revealed that PART-M2-C0.05-Q1 after 300% SMOTE has performed prediction better than logistic regression in $ROC_{Residual}$. The model that PART-M2-C0.05-Q1 scheme after 300% SMOTE learns is

better in area under the ROC curve for residual outcome than Naïve Bayes and logistic and better than J48 in the ROC area for the other outcome classes.

PART-M2-C0.05-Q1 after 300% SMOTE resulted in 76.81% accuracy and with a weighted area under the ROC curve of 0.742 was used to build the predictive model (see appendix E for some portion of rules the algorithm has resulted in). At first scene these performance measures seem very low as compared to the very high accuracy, sensitivity and specificity needed in surgical decision making. But, predicting surgical outcomes disregarding the preoperative care provided, intra-operative complexities that may occur during surgery, the post-operative care and complexities at this level of accuracy and ROC area are encouraging.

To sum up, consultation with domain experts on the rules and models that were left after objective evaluations also confirms that the increase in the severity of fistula diminishes the chance of being cured after surgical repair. Less severity, on the other hand, is a positive ground for "cure" as an outcome. This shows that the finding of this research agrees with the previously existing knowledge in urinary fistula surgical repair outcome.

6.2. Recommendation

Before the data has been used for the purpose of predictive model building and association rule mining, a number of preprocessing and preparation steps were carried out on the data. Those activities which resulted in clean data are: cleaning for errors, and handling missing values. As indicated in summary statistics during data preparation, the dataset has some error entries that could be prevented by predefining the values a particular attribute can take. This is because of the holistic treatments that the hospital provides to victims of obstetric fistula and injuries in birth tract, so that the database was made to include all the variables to all the different types of injuries. Thus, variables that apply to a particular injury type will be non applicable to the other (see appendix F for the treatment form). These attribute values create difficulties to the extraction of meaningful knowledge from the database. The solution to this problem, for example,

could be to create different forms and tables to record victims based on the type of surgical repair performed. Some important benefits that this solution can provide are, ease in generating reports in simple statistical tools and decrease the task of filling non applicable attribute values if the case is only of a specific type.

The predictive model can assist urinary fistula surgical repair outcome prediction with the given levels of accuracy and weighted area under the ROC curve. The model can also be used to provide post operative advice and make consultation with a victim who has already undergone surgical repair. With the development of small knowledge base system the usability of the model can go further to the time of actual surgery, by making the system available on hand held small portable computers. But before moving to the construction of knowledge base system (KBS) that contains knowledge of the domain area as depicted by the model obtained, the researcher would like to give some recommendation about the data attribute values captured. First, the entry of errors to columns of the database should be protected by predefining the valid values the attribute can take. Second, to eradicate some inapplicable values for a particular case it would be better to capture data based on the type of surgical intervention that are needed by the situation of victims who came for treatment.

Finally, it has been observed that classification algorithms differ based on the performance of the model they build. With the short period of time given to this research, it was found impossible to experiment more than four algorithms for predictive model building. Therefore, to come up with a model that may show better performance even from the model used to extract predictive rules, classification algorithms such as support vector machine (SVM), multi layer perceptrones (MLP) and many others can be experimented. This will help to compare the performances of the models with the model from this research, and to move onto the level of deployment.

REFERENCES

1. SIGKDD: Special Interest Group (SIG) on Knowledge Discovery & Data Mining Curriculum Committee: Data Mining Curriculum [internet]. [cited 2011 Dec 2]. Available from: http://www.sigkdd.org/curriculum.php.

2. Witten Ian H, Frank Eibe. Data Mining: Practical Machine Learning Tools and Techniques. Second edition. USA: Elsevier inc; 2005.

3. Encyclopaedia Britannica Online dictionary. [internet]. Christopher Clifton, editor. Data mining: Definition. [cited 2011 Dec 2]. Available from: http://www.britannica.com/EBchecked/topic/1056150/data-mining.

4. Hsu Hui-Hwang. editor. Advanced Data Mining Technologies in Bioinformatics. Hershey, USA: Idea Group Inc; 2006.

5. Nisbet R, Elder J, Miner G. Handbook of Statistical Analysis and Data Mining Applications. Canada: Elsevier Inc; 2009.

6. Hand D, Mannila H, Smyth P. Principles of Data Mining. London, UK: MIT Press; 2001.

7. Cios Krzysztof J, Pedrycz Witold, Swiniarski Roman W, Kurgan Lukasz A. Data Mining: A Knowledge Discovery Approach. New York, USA: Springer Science Business Media LLC; 2007.

8. Han Jiawei, Kamber Micheline. Data Mining: Concepts and Techniques. New York. USA: Morgan Kaufmann Publishers; 2001.

9. Fayyad Usama, Piatetsky-shapiro Gregory, Smyth Padharic. From Data Mining to Knowledge Discovery in Databases. [internet] 1996. [cited 2011 April 10]. Available from: http://citeseer.nj.nec.com.fayyad96from.html

10. Pieter Adrians, Dolf Zantinge. Data Mining. New York: Addison Wesley; 1996.

11. Sumathi S, Sivanandam SN. Introduction to Data: Mining and its Applications. Berlin, German: Springer-Verlag inc; 2006.

12. Two Crows Corporation. Introduction to Data Mining and Knowledge Discovery. Third Edition. USA: Two Crows Corporation; 2005.

13. Milley A. Healthcare and Data Mining. Health Management Technology. 2000; 21(8): 44-47.

14. Tzung-I Tang, Gang Zheng, Yalou Huang, Guangfu Shu, Pengtao Wang. "A Comparative Study of Medical Data Classification Methods Based on Decision Tree and System Reconstruction Analysis". IEMS. 2005 June; 4(1): 102-108.

15. Reinbolt J A, Melanie D. Fox, Michael H. Schwartz, Scott L. Delp. Predicting outcomes of rectus femoris transfer surgery. Journal of Gait & Posture [internet]. 2009 [cited 2011 Dec 2]; 30: 100–105. Available from: http://www.elsevier.com/locate/gaitpost

16. WHO. Trends in Maternal Mortality: 1990 to 2008 Estimates developed by WHO, UNICEF, UNFPA and The World Bank: World Health Organization; 2010. [internet]. [cited 2012 Jan 19]. Available from: whqlibdoc.who.int/publications/2010/9789241500265_eng.pdf

17. WHO. International Statistical Classification of Diseases and Related Health Problems. Tenth Revision, vol. 2, Instruction Manual. 2010 edition. Malta 2010. [internet]. [cited 2012 Jan 19]. Available from: www.who.int/entity/classifications/icd/ICD10Volume2_en_2010.pdf

18. Lori Ashford. Hidden Suffering: Disabilities from Pregnancy and Childbirth in Less Developed Countries. Population Reference Bureau. USA; 2002 Aug: 1-5.

19. Neilson JP, Lavender T, Quenby S, Wray S.Obstructed labour: Reducing maternal death and disability during pregnancy. Br Med Bull. 2003; 67:191-204.

20. Carla AbouZahr. Global burden of maternal death and disability. World Health Organization, Geneva, Switzerland. British Medical Bulletin 2003; 67: 1–11. [internet]. [cited 2012 Jan 23]. Available from: http://bmb.oxfordjournals.org/DOI: 10.1093/bmb/ldg015.

21. Wall L Lewis. Obstetric Vesico Vaginal Fistula as an International Public-Health Problem. Lancet. 2006 September 30; 368: 1201–09.

22. Hinrichsen D. Obstetric Fistula: Ending the Silence, Easing the Suffering. INFO Reports, No. 2. Baltimore, Johns Hopkins Bloomberg School of Public Health, the INFO Project, Sept. 2004. [internet]. [Cited 2011 Nov 23]. Available from: www.youthwg.org/system/files/ObstetricFistula.pdf

23. WHO. Obstetric Fistula: Guiding Principles for Clinical Management and Programme Development: 2006. Gwyneth Lewis and Luc de Bernis, editors. [internet]. [cited 2011 Dec 2]: Available from: http://www.who.int/reproductivehealth/publications/maternal_perinatal_health/9241593679/en/

24. Johnson Kiersten, Peterman Amber. Incontinence Data from the Demographic and Health Surveys: Comparative Analysis of a Proxy Measurement of Vaginal Fistula and Recommendations for Future Population-Based Data Collection. DHS Analytical Studies No. 17. Macro International Inc. Calverton, MD USA; November 2008 [internet]. [cited 2011 Nov 25]. Available from: pdf.usaid.gov/pdf_docs/PNADN516.pdf

25. Udoma EJ, Umoh MS, Udosen EO. Recto-Vaginal Fistula Following Coitus: An Aftermath of Vaginal Douching with Aluminium Potassium Sulphate Dodecahydrate (potassium alum). International Journal of Gynecology and Obstetrics. 1999; 66: 299-300.

26. Melah GS, Massa AA, Yahaya UR, Bukar M, Kizaya DD, El- Nafaty AU. Risk Factors for Obstetric Fistulae in North-Eastern Nigeria. J Obstet Gynaecol. 2007 Nov; 27(8): 819-23.

27. PT Haspels-Kenter C.W.M. Postoperative Urine and Fecal Incontinence after surgical VVF/RVF repair. Report on the meeting for the prevention & treatment of obstetric fistula. London. July 2001. [Cited 2011 Nov 23]. Available from: www.unfpa.org/webdav/site/global/.../2001/fistula_report_2001.pdf

28. Goh Judith T.W. Sloane Kate M., Krause Hannah G. Browning Andrew, Akhter Sayeba. Mental health screening in women with genital tract fistulae. BJOG: an International Journal of Obstetrics and Gynaecology. 2005 Sept; 112: 1328–1330.

29. Goh Judith T.W. Genital Tract Fistula Repair on 116 Women. Aust. NZ J Obstet Gynaecol. 1998; 38(2): 158-161.

30. Arrowsmith SD, Hamlin EC, Wall LL. Obstructed Labour Injury Complex: Obstetric Fistula Formation and the Multifaceted Morbidity of Maternal Birth Trauma in the Developing World. Obstet Gynecol Surv. 1996 Sep; 51(9): 568-74.

31. Goh Judith T.W, Stanford EJ, Genadry R. Classification of female genito-urinary tract fistula: a comprehensive review. Int Urogynecol J Pelvic Floor Dysfunct. 2009 Jan 30.

32. Creanga AA, Genadry RR. Obstetric Fistulas: A Clinical Review. Int J Gynaecol Obstet. 2007 Nov; 99 Suppl 1: 40-6.

33. Arrowsmith SD. The Classification of Obstetric Vesico-Vaginal Fistulas: A Call for an Evidence-Based Approach. Int J Gynaecol Obstet. 2007 Nov; 99 Suppl 1: 25-7.

34. Wall L Lewis, Arrowsmith SD, Briggs ND, Browning A, Lassey A. The obstetric Vesicovaginal Fistula in the Developing World. Obstet Gynecol Surv. Committee 20. 2005 Jul; 60 (7 Suppl 1): 3-51.

35. Rakesh Kapoor, M. S. Ansari, Pratipal Singh, Parag Gupta, Naval Khurana, Anil Mandhani, Deepak Dubey, Aneesh Srivastava, and Anant Kumar. Management of vesicovaginal fistula: An experience of 52 cases with a rationalized algorithm for choosing the transvaginal or transabdominal approach. Indian J Urol. 2007 Oct-Dec; 23(4): 372–376.

36. Morhason-Bello I. O, Ojengbede O. Adedokun A., B. O, Okunlola M. A., and Oladokun A. Annals of Ibadan Postgraduate Medicine. 2008 Dec; 6 (2): 39-43.

37. Kelly J. Vesico-Vaginal and Recto-Vaginal Fistulae. Journal of the Royal Society of Medicine. 1992 May; 85: 257-258.

38. Pierre Marie Tebeu, Joseph Nelson Fomulu, Achille Aurelien Mbassi, Jean Marie Tcheliebou, Anderson Sama Doh, Charles Henry Rochat. Quality Care in Vesico-Vaginal Obstetric Fistula: Case Series Report From The Regional Hospital of Maroua-Cameroon.
The Pan African Medical Journal. 2010; 5: 6.

39. Wall L Lewis, Arrowsmith SD. The "Continence Gap": A Critical Concept in Obstetric Fistula Repair. Int Urogynecol J Pelvic Floor Dysfunct. 2007 Aug; 18(8): 843-4.

40. Wall L Lewis, Karshima JA, Kirschner C, Arrowsmith SD. The Obstetric Vesicovaginal Fistula: Characteristics of 899 Patients from Jos, Nigeria. Am J Obstet Gynecol. 2004 Apr; 190 (4): 1011-9.

41. Genadry RR, Creanga AA, Roenneburg ML, Wheeless CR. Complex Obstetric Fistulas. Int J Gynaecol Obstet. 2007 Nov; 99 Suppl 1: S51-6.

42. Solbjorg Sjoveian, Siri Vangen, Denis Mukwege, Mathias Onsrud. Surgical Outcome of Obstetric Fistula: A Retrospective Analysis of 595 Victims. Nordic Federation of Societies of Obstetrics and Gynecology. 2011; 90: 753–760.

43. Browning A. Risk Factors for Developing Residual Urinary Incontinence after Obstetric Fistula Repair. BJOG An International Journal of Obstetrics and Gynaecology. 2006 Feb 20; 113: 482–485.

44. Goh Judith T. W, Browning Andrew, Birhanu Berhan, Chang Allan. Predicting the Risk of Failure of Closure of Obstetric Fistula and Residual Urinary Incontinence Using a Classification System. Int Urogynecol J. 2008; 19: 1659–1662.

45. Cerrito Patricia, editor. Cases on Health Outcomes and Clinical Data Mining: Studies and Frameworks. Introduction to Data Mining Methodology to Investigate Health Outcomes. USA: IGI Global; 2010.

46. Berry Michael J.A., Linoff Gordon S. Data Mining Techniques: For Marketing, Sales, and Customer Relationship Management. Second Edition. Wiley Publishing, Inc., United States of America. 2004.

47. Bath Peter A. Data Mining in Health and Medical Information. In: Blaise Cronin, editor. Annual review of Information Science and Technology. Vol 38. USA: Information Today Inc; 2004.

48. Larose Daniel T. Discovering Knowledge in Data - An Introduction to Data Mining. New Jersey, USA: John Wiley & Sons Inc; 2005.

49. Cios Krzysztof J, Kurgan Lukasz A. Trends in Data Mining and Knowledge Discovery. In: Pal Nikhil R., Jain Lakhmi, editors. Advanced Techniques in Knowledge Discovery and Data Mining. London. UK: Springer Science Business LLC; 2005.

50. Ponce Julio, Adem Karahoca, editors. Data Mining and Knowledge Discovery in Real Life Applications. Croatia: I-Tech Education and Publishing Inc; 2009.

51. Webb Geoffrey I. Association Rules. In: Ye Nong, Editor. The Handbook of Data Mining. New Jersey. USA: Lawrence Erlbaum Associates Inc; 2003.

52. Bramer Max. Principles of Data Mining. London. Springer-Verlag Limited; 2007.

53. Mehmed Kantardzic J. B. Data Mining—Concepts, Models, Methods, and Algorithms.USA: John Wiley & Sons publication Inc; 2003.

54. Pagano Marcello, Gauvreau Kimberlee. Principles of Bio Statistics. Second edition. New Delhi. India: Cengage Learning; 2010

55. Yun Wang. A Multinomial Logistic Regression Modeling Approach for Anomaly Intrusion Detection. Journal of computer and security. Yale University. USA. Elsevier. 2005. 662-674.

56. Weiss Sholom M, Zhang Tong. Performance Analysis and Evaluation. In: Ye Nong, Editor. The handbook of data mining. New Jersey. USA: Lawrence Erlbaum Associates Inc; 2003.

57. Sellappan Palaniappan, Awang Rafiah. Intelligent Heart Disease Prediction System Using Data Mining Techniques. International Journal of Computer Science and Network Security. 2008 Aug; 8 (8): 343-350.

58. Young J, Pitta J. Wal-Mart or Western Union? United Health Care Corp. Forbes. John Wiley & Sons Inc. 1997; 160(1): 244.

59. Kolar HR. Caring for Healthcare. Health Management Technology. 2001; 22(4): 48-47.

60. Tylor P. From Patient Data to Medical Knowledge: The Principles and Practices of Health Informatics. UK: Backwell Publishings ltd; 2008.

61. Kincade K. Data Mining: Digging for Healthcare Gold. Insurance & Technology, 1998; 23(2): 2-7.

62. Shantakumar B.Patil, Kumaraswamy Y.S.. Intelligent and Effective Heart Attack Prediction System: Using Data Mining and Artificial Neural Network. European Journal of Scientific Research [internet]. EuroJournals Publishing Inc. 2009; 31(4): 642-656. [cited 2011 Dec 2]: Available from: http://www.eurojournals.com/ejsr.htm.

63. Gracia Jacob Shomona, Ramani R.Geetha. Discovery of Knowledge Patterns in Clinical Data through Data Mining Algorithms: Multiclass Categorization of Breast Tissue Data. International Journal of Computer Applications. Tamilnadu, India. 32 (72011): 0975 – 8887.

64. Karpagavalli S, Jamuna KS, and Vijaya MS. Machine Learning Approach for Preoperative Anaesthetic Risk Prediction. International Journal of Recent Trends in Engineering. May 2009 India; 1(2).

65. Soni Jyoti, Ansari Ujma, Sharma Dipesh, Soni Sunita. Predictive Data Mining for Medical Diagnosis: An Overview of Heart Disease Prediction. International Journal of Computer Applications. March 2011. India. 17(8): 0975 – 8887.

66. Wall L Lewis, Arrowsmith S.D, Briggs N. D, Lassey A. Committee 12 Urinary Incontinence in the Developing World: The Obstetric Fistula. [Cited 2011 Nov 23]. Available from: http://icsoffice.org/Publications/ICI_2/chapters/Chap12.pdf

67. Bouckaert Remco R., Frank Eibe, Hall Mark, Richard Kirkby, Reutemann Peter, Seewald Alex, Scuse David. WEKA Manual for Version 3-6-2. University of Waikato, Hamilton, New Zealand. January 11, 2010.

68. Nitesh V. Chawla, Kevin W. Bowyer, Lawrence O. Hall, W. Philip Kegelmeyer. "SMOTE: Synthetic Minority Over-sampling Technique". Journal of Artificial Intelligence Research. June 2002. AI Access Foundation and Morgan Kaufmann Publishers. 16: 321-357.

APPENDICES

Appendix A: Attributes Selected as Predictors of Surgical Repair Outcome, After the Removal of Socio Demographic Variables.

Search Method:
Attribute ranking.
=== Run information ===
Evaluator: weka.attributeSelection.ChiSquaredAttributeEval
Search: weka.attributeSelection.Ranker
Relation: INSTANCES
Instances: 15546
Attributes: 11
No of Prev Repair at Other Hospital
Type of urinary fistula
Length
Width
Scarring
Bladder size
Status of Bladder neck
Status of Urethra
Number of fistula
Status of Ureters
Surgery Outcome Urinary
Evaluation mode: evaluate on all training data
=== Attribute Selection on all input data ===
Search Method:
Attribute ranking.
Attribute Evaluator (supervised, Class (nominal): 11 Surgery Outcome Urinary):
Chi-squared Ranking Filter
Ranked attributes:
1509.1938 7 Status of Bladder neck
1413.9918 2 Type of urinary fistula
1406.2786 5 Scarring
1151.1119 8 Status of Urethra
782.0362 6 Bladder size
780.7277 3 Length
749.337 4 Width
166.9573 10 Status of Ureters
74.5643 1 No of Prev Repair at Other Hospital
72.0559 9 Number of fistula
Selected attributes: 7,2,5,8,6,3,4,10,1,9 : 10

Appendix B: Output of unpruned J48 Selected Scheme

Scheme: weka.classifiers.trees.J48 -U -M 2
Relation: PREPROCESSED DATASET 15546
weka.filters.unsupervised.attribute.ReplaceMissingValues-
weka.filters.supervised.instance.SMOTE-C0-K5-P300.0-S1
Instances: 16110
Attributes: 11
=== Summary ===
Correctly Classified Instances 11941 74.1217 %
Incorrectly Classified Instances 4169 25.8783 %
=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
	0.226	0.075	0.321	0.226	0.265	0.625	Stress
	0.395	0.019	0.505	0.395	0.443	0.872	Residual
	0.073	0.02	0.167	0.073	0.101	0.548	Failed
	0.9	0.666	0.814	0.9	0.855	0.698	Cured
Weighted Avg.	0.741	0.522	0.699	0.741	0.716	0.688	

=== Confusion Matrix ===

a	b	c	d	<-- classified as
494	69	105	1518	a = Stress
27	297	4	424	b = Residual
180	26	62	584	c = Failed
836	196	200	11088	d = Cured

Appendix C: Output of the Naïve Bayes Selected Scheme

Scheme: weka.classifiers.bayes.NaiveBayes
Relation: PREPROCESSED DATASET 15546
weka.filters.unsupervised.attribute.ReplaceMissingValues
Instances: 15546
Attributes: 11
=== Summary ===
Correctly Classified Instances 11555 74.3278 %
Incorrectly Classified Instances 3991 25.6722 %
=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
	0.403	0.137	0.325	0.403	0.36	0.751	Stress
	0	0	0	0	0	0.677	Residual
	0.133	0.029	0.211	0.133	0.163	0.725	Failed
	0.857	0.538	0.859	0.857	0.858	0.757	Cured
Weighted Avg.	0.743	0.448	0.738	0.743	0.739	0.753	

=== Confusion Matrix ===

a	b	c	d	<-- classified as
881	0	148	1157	a = Stress
52	0	12	124	b = Residual
283	0	113	456	c = Failed
1496	1	262	10561	d = Cured

Appendix D: Output of the Logistic Regression Selected Scheme

Scheme weka.classifiers.functions.Logistic -R 1.0E-8 -M -1
Relation: PREPROCESSED DATASET weka.filters.unsupervised.attribute.ReplaceMissingValues
Instances: 15546
Attributes: 11
=== Summary ===
Correctly Classified Instances 12347 79.4224 %
Incorrectly Classified Instances 3199 20.5776 %
=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
	0.134	0.029	0.426	0.134	0.203	0.761	Stress
	0	0	0	0	0	0.669	Residual
	0.028	0.003	0.381	0.028	0.052	0.726	Failed
	0.977	0.858	0.813	0.977	0.887	0.766	Cured
Weighted Avg.	0.794	0.684	0.725	0.794	0.735	0.762	

=== Confusion Matrix ===

a	b	c	d	<-- classified as
292	0	17	1877	a = Stress
16	0	2	170	b = Residual
108	0	24	720	c = Failed
269	0	20	12031	d = Cured

Appendix E: Some Portion of Output from PART Algorithm (scheme: PART -M 2-C0.05-Q1)

Status of Urethra = Intact AND Status of bladder neck = Intact AND Scarring = None AND Length = 1.0: Cured (1746.0/115.0)	Status of Urethra = Intact AND Status of bladder neck = Intact AND Scarring = None AND Length = 2.0: Cured (1656.0/118.0)
Status of Urethra = Intact AND Scarring = Mild AND Type of urinary fistula = Juxta-cervical AND Length = 2.0: Cured (312.0/10.0)	Status of Urethra = Intact AND Scarring = Mild AND Type of urinary fistula = Mid Vaginal: Cured (1493.0/198.0)
Status of bladder neck = Intact AND Scarring = None AND No of Prev Repair Other Hospital = Not applicable AND Type of urinary fistula = Juxta-cervical: Cured (385.0/55.0)	Status of bladder neck = Intact AND Scarring = None AND No of Prev Repair Other Hospital = No Information: Cured (221.0/31.0)
Status of bladder neck = Intact AND Scarring = None AND Type of urinary fistula = Juxta-urethral: Cured (97.0/12.0)	Status of bladder neck = Intact AND Scarring = Mild: Cured (1780.0/306.0)
Status of bladder neck = Intact AND Scarring = None AND Type of urinary fistula = Vault: Cured (45.0/1.0)	Status of bladder neck = Intact AND Scarring = None AND Width = 4.0 AND Status of Urethra = Intact: Cured (213.0/42.0)
Status of bladder neck = Intact AND Scarring = None AND Width = 3.0: Cured (238.0/51.0)	Status of bladder neck = Intact AND Scarring = None AND Width = 5.0: Cured (147.0/28.0)
Status of bladder neck = Intact AND Scarring = None AND Width = 2.0: Cured (105.0/9.0)	Status of bladder neck = Intact AND Scarring = None AND Width = 1.0: Cured (65.0/11.0)
Status of bladder neck = Intact AND Scarring = Moderate AND Bladder size = Fair: Cured (244.0/50.0)	Status of bladder neck = Intact AND Scarring = Moderate AND Bladder size = Good: Cured (519.0/97.0)
Status of bladder neck = No Information AND Scarring = None AND Status of Urethra = Intact AND Type of urinary fistula = Mid Vaginal: Cured (50.0/1.0)	Status of bladder neck = Partially Damaged AND Type of urinary fistula = Juxta-cervical AND Scarring = Moderate AND Number of fistula = 1.0: Cured (25.0/7.0)
Status of bladder neck = Partially Damaged AND Type of urinary fistula = Juxta-cervical AND Scarring = Mild: Cured (40.0/3.0)	Status of bladder neck = Partially Damaged AND Type of urinary fistula = Mid Vaginal: Cured (206.0/54.0)

Status of bladder neck = Partially Damaged AND Type of urinary fistula = Juxta-urethral AND Scarring = None AND Status of Urethra = Intact: Cured (79.0/13.0)	Type of urinary fistula = Circumferential AND Status of Ureters = Both Inside AND Length = >5 AND Status of Urethra = Partial Damage: Stress (7.0/1.0)
Status of bladder neck = Intact AND Bladder size = No information AND Length = 1.0 AND No of Prev Repair Other Hospital = Not applicable AND Type of urinary fistula = Mid Vaginal: Cured (23.0/4.0)	Status of bladder neck = Complete Destruction AND Scarring = Mild AND Width = 5.0 AND Number of fistula = 1.0 AND Type of urinary fistula = Mid Vaginal: Stress (4.0/1.0)
Status of bladder neck = Partially Damaged AND Type of urinary fistula = Juxta-urethral AND No of Prev Repair Other Hospital = No Information AND Scarring = Mild: Cured (17.0/3.0)	Status of bladder neck = Complete Destruction AND Scarring = Mild AND Width = 5.0 AND Number of fistula = 1.0 AND Type of urinary fistula = Circumferential AND Length = 4.0: Stress (4.0/1.0)
Status of bladder neck = Intact AND Bladder size = No information AND Scarring = Moderate AND Type of urinary fistula = Juxta-urethral AND Length = 3.0: Cured (15.0/4.0)	Status of bladder neck = Complete Destruction AND Scarring = Obliterated vagina AND Number of fistula = 1.0 AND No of Prev Repair Other Hospital = Not applicable AND Type of urinary fistula = Juxta-urethral: Stress (4.0/1.0)
Status of bladder neck = Partially Damaged AND Scarring = Mild AND Number of fistula = 2.0: Cured (75.0/23.0)	Type of urinary fistula = Combined AND Status of bladder neck = Partially Damaged AND Scarring = Obliterated vagina AND Status of Urethra = Intact: Stress (4.0/1.0)
Status of bladder neck = Intact AND Bladder size = No information AND Length = 1.0 AND No of Prev Repair Other Hospital = Not applicable AND Type of urinary fistula = Mid Vaginal: Cured (23.0/4.0)	Type of urinary fistula = Combined AND Status of bladder neck = Partially Damaged AND Scarring = Severe AND Status of Ureters = Both Inside AND Number of fistula = 1.0 AND Length = >5 AND Bladder size = No information: Stress (4.0/1.0)
Status of bladder neck = Partially Damaged AND Scarring = Mild AND Width = 1.0 AND Length = 2.0: Cured (23.0/5.0)	Length = 3.0 AND Type of urinary fistula = Combined AND Status of Ureters = Both Inside AND Width = 5.0: Stress (12.0/2.0)
Status of bladder neck = No Information AND Scarring = None: Cured (51.0/10.0)	No of Prev Repair Other Hospital = 2.0: Stress (8.0/1.0)

Status of bladder neck = Partially Damaged AND Type of urinary fistula = Juxta-cervical AND Scarring = Moderate AND Number of fistula = 1.0: Cured (25.0/7.0)	Type of urinary fistula = Circumferential AND Status of Ureters = Both Inside AND Length = >5 AND Status of Urethra = Partial Damage: Stress (7.0/1.0)
Status of bladder neck = Partially Damaged AND Type of urinary fistula = Juxta-urethral AND No of Prev Repair Other Hospital = No Information AND Scarring = Mild: Cured (17.0/3.0)	Status of bladder neck = Complete Destruction AND Scarring = Mild AND Width = 5.0 AND Number of fistula = 1.0 AND Type of urinary fistula = Mid Vaginal: Stress (4.0/1.0)
Status of bladder neck = Intact AND Bladder size = No information AND Scarring = Moderate AND Type of urinary fistula = Juxta-urethral AND Length = 3.0: Cured (15.0/4.0)	Status of bladder neck = Complete Destruction AND Scarring = Mild AND Width = 5.0 AND Number of fistula = 1.0 AND Type of urinary fistula = Circumferential AND Length = 4.0: Stress (4.0/1.0)
Status of bladder neck = Partially Damaged AND Type of urinary fistula = Juxta-urethral AND Scarring = None AND Status of Urethra = Intact: Cured (79.0/13.0)	Status of bladder neck = Complete Destruction AND Scarring = Obliterated vagina AND Number of fistula = 1.0 AND No of Prev Repair Other Hospital = Not applicable AND Type of urinary fistula = Juxta-urethral: Stress (4.0/1.0)
Status of bladder neck = Partially Damaged AND Scarring = Mild AND Number of fistula = 2.0: Cured (75.0/23.0)	Type of urinary fistula = Combined AND Status of bladder neck = Partially Damaged AND Scarring = Obliterated vagina AND Status of Urethra = Intact: Stress (4.0/1.0)
Status of bladder neck = Partially Damaged AND Scarring = Mild AND Width = 1.0 AND Length = 2.0: Cured (23.0/5.0)	Type of urinary fistula = Combined AND Status of bladder neck = Partially Damaged AND Scarring = Severe AND Status of Ureters = Both Inside AND Number of fistula = 1.0 AND Length = >5 AND Bladder size = No information: Stress (4.0/1.0)
Type of urinary fistula = Combined AND Status of bladder neck = Partially Damaged AND Bladder size = Good AND Status of Ureters = Both Inside: Stress (8.0/3.0)	Length = 3.0 AND Type of urinary fistula = Combined AND Status of Ureters = Both Inside AND Width = 5.0: Stress (12.0/2.0)
Status of bladder neck = No Information AND Scarring = None: Cured (51.0/10.0)	No of Prev Repair Other Hospital = 2.0: Stress (8.0/1.0)

Length = 2.0 AND Type of urinary fistula = Juxta-urethral AND Scarring = Moderate AND Width = 2.0 AND No of Prev Repair Other Hospital = Not applicable AND Status of bladder neck= Complete Destruction: Stress (7.0/1.0)	Length = 2.0 AND Bladder size = No information: Stress (25.0/8.0)
Length = 3.0 AND Type of urinary fistula = Combined AND Status of Urethra = Partial Damage: Stress (8.0/2.0)	Length = 3.0 AND Type of urinary fistula = Juxta-cervical: Stress (6.0/3.0)
Length = 3.0 AND Status of Urethra = Intact AND Width = 4.0: Stress (16.0/2.0)	Length = 3.0 AND Type of urinary fistula = Combined AND Status of Ureters = Both Inside AND Width = 5.0: Stress (12.0/2.0
Type of urinary fistula = Combined AND Status of bladder neck = Partially Damaged ANDBladder size = Small AND Length = 5.0 AND Status of Urethra = Partial Damage: Failed (5.0/1.0)	Length = 2.0 AND Type of urinary fistula = Juxta-urethral AND Scarring = Moderate AND Width = 2.0 AND No of Prev Repair Other Hospital = Not applicable AND Status of bladder neck= Complete Destruction: Stress (7.0/1.0)
Status of bladder neck = No Information AND Scarring = Mild AND Width = 4.0: Residual (8.0/1.0)	Status of bladder neck = Partially Damaged AND Type of urinary fistula = Circumferential AND Status of Ureters = Both Inside AND Length = >5 AND Status of Urethra = Partial Damage: Stress (7.0/1.0)
Type of urinary fistula = Combined AND Status of bladder neck = Partially Damaged AND Width = 3.0 AND Status of Urethra = Complete Destruction: Residual (7.0/1.0)	Type of urinary fistula = Combined AND Status of bladder neck = Partially Damaged AND Number of fistula = 2.0 AND Status of Urethra = Partial Damage AND Length = 4.0: Residual (8.0/1.0)
Type of urinary fistula = Absent urethra AND Bladder size = Small: Failed (12.0/2.0)	Status of Ureters = Both outside: Failed (6.0/1.0)

Appendix F: Addis Ababa Fistula Hospital Treatment Form

<table>
<tr><th colspan="8">Addis Ababa Fistula Hospital</th></tr>
<tr><th colspan="8">Patient Data</th></tr>
<tr><th colspan="8">Socio Demographic Data-Date of Registration</th></tr>
<tr><td>Name</td><td>Card No</td><td colspan="2">Current Age
Age at Marriage
Age at Causative delivery</td><td>Height</td><td>Weight</td><td colspan="2">Parity
- primipara
- multipara
-GrandMultpara
-nulipara
-no information</td></tr>
<tr><td colspan="2">No of living children</td><td colspan="2">Region
-Amhara -SNNP
-Oromiya -Somalia Region
- Tigray -Afar
- Benishangul -Addis Ababa
- No information -Other</td><td colspan="2">Religion
- Christian
- Muslim
- Others
- No Information</td><td colspan="2">Days traveled (foot)
- <1
-1-2
- 2-3
- >3
- No information
- Insignificant</td></tr>
<tr><td colspan="2">Days traveled: (by transport)
- <1
-1-2
- 2-3
- >3
- No information
- Insignificant</td><td colspan="4">Educational status
- Illiterate
- Literate (Basic)
- Secondary
- Elementary
- Secondary
- Above High School
- No Information</td><td colspan="2">Marital Status
- Not maried
- Maried
- Separated
- Divorced
- Widowed
- No information</td></tr>
<tr><td colspan="2">Accompanying person:-
- Self
- Husband
- Relatives
- Husband and Relatives
- Others
- No information</td><td colspan="4">Estimated distance to nearest institution (traveled on foot):- (days)
- <1/2
- ½-1
- 1-2
- >2
- No information</td><td>Source of information:-
- Cured Patient
- Health Professional
- Mission
- Other
- No information</td><td>How many days before she could walk?</td></tr>
<tr><th colspan="8">Obstetric and Medical History</th></tr>
<tr><td colspan="2">Antenatal Care
- Yes
- No
- No information</td><td colspan="2">Duration of incontinence: (in months)</td><td colspan="2">No of previous fistula repair at other hospital:-
- 1
- 2
- 3
- >3
- Not applicable
- No information</td><td colspan="2">Cause of fistula
- Child birth
- Post Coital
- Surgery
- Trauma
- Radiation
- Other
- No information</td></tr>
</table>

Duration of labour (day) - 1 - 2 - 3 - 4 - >=5 - No information	Place of delivery - Home - Institutional - Other - No information	Mode of delivery - Vaginal - Assisted vaginal - Abdominal - No information	Fetal Outcome - Still birth - Alive - Early neonatal death - One dead and one alive - Not applicable - No information

Other medical surgical illness - Yes - No	Menstruation Hx: - Amenorrhic - Resumed before 3 months - Resumed between 3 and 6 months - Resumed by 6/12 - Resumed by >7/12 - Resumed after 1 year - Not applicable - Other - No information	Place of treatment - Addis Ababa - Bahar Dar - Harar - Mekele - Metu - Yirgalem

Preoperative Care

Preoperative stay wk - < 1/52 - 1/52-4/52 - 5/52-12/52 - 13/52-20/52 - 21/52-36/52 - >37/52 - No information	Antibiotics - Yes - No	Antibiotics - Penicillin - Aminoglycocids - Cepalo sporins - Combination - Metronidazole - Sulphonamides - Others - No information	Pre-operative care provided:- - Does not apply - Nutritional - Physiotherapy for foot contracture - Physiotherapy for foot drop - Medical - Colostomy - Local vaginal care - Bladder stone removal - Others - No information	Nerve and musculo skeletal injury:- - Unilateral foot drop - Bilateral foot drop - Unilateral knee contracture - Bilateral knee contracture - No injuries

Operation Date				
Surgeon List	Anesthesia - Spinal - Spinal-sedation - Spinal-general - General	Approach for urinary fistula repair:- - Viginal - Abdominal - Combined	Circumcision - Type I - Type II - Type III - Type IV - None - No information	Type of procedure: - Urinary fistula - Rectal - Both - Stress - Third degree tear - Urinary and 3rd degree tear repair - Diversion - Other

Number of urinary fistula: - 1 - 2 - 3 ->3	Type of urinary fistula:- - Urethral - Absent Urethral - Torn Urethra - Juxta-urethral - Mid vaginal - Combined - Juxta-cervical - Vesico-uterine - Circumferencial - Ureteric - Vault - No bladder - Other - No information	VVF length (cm):- - 1 - 2 - 3 - 4 - 5 - >5	VVF width (cm):- - 1 - 2 - 3 - 4 - 5 - >5	Scarring - None - Mild - Moderate - Severe - Obliterated vagina	Bladder Size:- - None - Small - Fair - Good - No information

Status of bladder neck:- - Intact - Partially damaged - Complete destruction - No information	Status of urethra:- - Intact - Partially damaged - Complete destruction - No information	Status of ureters:- - Both inside - Both outside - One outside	Ureteric catheters - Catheterized - Not Catheterized
Bladder fistula closure - 1 - 2 - More	Graft - Yes - No	Flap - Perennial - Labial - None - Not applicable	Rectal-injury type - High - Low - Mid - Circumferential - Combined - Other

RVF length (cm):- - 1 - 2 - 3 - 4 - 5 - >5	RVF width (cm):- - 1 - 2 - 3 - 4 - 5 - >5	Rectal fistula closure layers:- - 1 - 2	Spincter Status - Intact - Destroyed - Lax

Intra operative complications:- - Iatrogenic - Hemorrhage - Others - None	Duration of Surgery:- (hr) - < 1 -1-2 -2-3 ->3 -No information	Surgery Outcome Urinary:- - Cured - Failed - Stress - Residual - Improved - No Change - Others (specify)	Surgery Outcome Bowel - Cured - Failed - Improved - No Change - Others (specify)

Death - Pre-operation - Post-operation - Intra-operative	Cause of death - Sepsis - Bleeding - Renal failure - Thromboembolic

Post-operative course

Transfusion - Yes - No	Antibiotics - Prophylaxis - Treatment - None - No information	Pack:- (days) - 0 - 1 - 2 - 3-5 - > 5 - No information

Post operative complications - UTI - Chest infection - Wound infection - Bleeding - Others - None	Duration of Bladder Urethral catheter:- (days) - < 10 - 10-15 - 16-21 - >21 - None - No information

Was she given clothes:- - Yes - No	Was she given transport money:- - Yes - No - No information	Total Length of Stay:- - <3 - 5-8 - 9-12 - 13-24 - >24 - No information