Sudan University of Science and Technology
College of Computer Science and Information Technology
Department of Information Systems and Technologies

Extracting Associations from Kidney Transplantations Dataset

إستخراج قواعد الإرتباط من بيانات زارعي الكلى

Project submitted in partial fulfillment of the requirement for the degree of BS.c (honors) in Information Systems and Technologies

August 2014

بسم الله الرحمن الرحيم

Sudan University of Science and Technology
College of Computer Science and Information Technology
Department of Information Systems and Technologies

Extracting Associations from Kidney Transplantations Dataset

إستخراج قواعد الإرتباط من بيانات زارعي الكلى

Prepared by:

Mudathir Abdallah Abdallah.
Mohammed Hasan Altayeb Alshreef.

Project submitted in partial fulfillment of the requirement for the degree of BS.c (honors) in Information Systems and Technologies

Supervisor Signature:	**Date:**
Mrs. Wafaa Faisal Mukhtar.	**27/08/3014**

الآيـة

بسم الله الرحمن الرحيم

قال تعالى:

(وَقُلِ اعْمَلُوا فَسَيَرَى اللَّهُ عَمَلَكُمْ وَرَسُولُهُ وَالْمُؤْمِنُونَ ◌َ)

صدق الله العظيم

سورة التوبة الأية (105)

الإهداء

إلى من كلت أنامله ليقدم لنا لحظة سعادة إلى من علم وعان على الصعاب

والدي العزيز

إلى الينبوع الذي لا يمل العطاء إلى من حاكت سعادتي بخيوط منسوجة من قلبها

أمي العزيزة

إلى من تذوقت معهم أجمل اللحظات إلى ملاذي المأمون

إخوتي الأعزاء

إلى من أبدلو الألم بالأمل إلى رفقاء الدرب

اصدقائي الإعزاء

إلى تلك السواعد البيضاء التي كستنا بريق العلم والمعرفة

أساتذتي الأجلاء

شكر وعرفان

ولو أنني أوتيت كل بلاغة ****** وأفنيت بحر النطق في النظم والنثر
لما كنت بعد القول إلا مقصرا ***** ومعترفا بالعجز عن واجب الشكر

يسرنا أن نقدم خالص الشكر لمن يستحق الشكر والتقدير ,إلى جميع منوجه وارشد وعلم

إلي جميع الأساتذة الأجلاء

إلى الشموع التي ذابت في كبرياء.......
لتنير كل خطوة في دربنا.......
لتذلل كل عائق أمامنا........
فكانوا رسلاً للعلم والأخلاق.......
شكراً لكم جميعاً........

ونخص بالشكر

أ / حسن

د/ سارة عثمان سليمان

Abstract

Data Mining is one of the most motivating area of research which become increasingly popular in healthcare. Studies in heath scope used data mining techniques for diagnosing diseases and much less in prognosis stage of diseases.

The records of the members of The Sudanese Kidney Transplanted Associations has been collected and their follow-up data from Cardiac Surgery & Renal Transplantations Center at Ahmed Qassim hospital in Khartoum North.

We used Clementine, ARMADA association mining tool and MatLab to develop a program based on Apriori algorithm for extracting association rules. There are some differences in the results, but they all shows the essentially factors that impacts on renal functions.

المستخلص

التنقيب في البيانات واحد من اكثر المجالات المحفزة للبحوث الذي اصبح ذا شعبية في مجال الرعاية الصحية . استخدمت الدراسات السابقة في مجال الصحة تقنيات تنقيب البيانات لتشخيص الأمراض، وأقل بكثير في مرحلة ما بعد التشخيص .

تم جمع بيانات أعضاء الجمعية السودانية لزارعي الكلى و بيانات المتابعة للزارعين من مركز جراحة القلب وزراعة الكلى في مستشفى أحمد قاسم في بحري شمال الخرطوم.

تم إستخدام أداة الـ Clementine و ARMADA و هي أداه لإستخراج قواعد الإرتباط من البيانات في برنامج الـ Mat Lab وقمنا بواستطه بتطوير برنامج يقوم علي أساس خوارزمية APRIORI لإكتشاف قواعد الإرتباط في البيانات المجموعة. كانت هنالك بعض الإختلاف في نتائج الإدوات ولكنها أظهرت العوامل الأساسية التي تؤثر علي وظائف الكلي .

Glossary

Abbreviation	Description
KDD	Knowledge Discovery in Data Base
ARMADA	Association Rule Mining
Matlab	Matrix Laboratory
FORTRAN	Formula Translating System
MuPAD	Computer Algebra System developed by MuPAD research group

Table of Figure

Table of Tables

Table of Content

CHAPTER ONE
INTRODUCTION

1 Introduction

Data mining is defined as a step in the knowledge discovery in databases (KDD) process that consists of applying data analysis and discovery algorithms that, under acceptable computational efficiency limitations, produce a particular enumeration of patterns (or models) over the data. KDD is also defined as the nontrivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data.

1.1 Mining Health Data

Data Mining is one of the most vital and motivating area of research with the objective of finding meaningful information from huge data sets. In present era, Data Mining is becoming popular in healthcare field because there is a need of efficient analytical methodology for detecting unknown and valuable information in health data.[1]

In present time, various public and private healthcare institutes are producing enormous amounts of data which are difficult to be handled. This information is very valuable for healthcare specialist to understand the cause of diseases and for providing better and cost effective treatment to patients. [1]

The data generated by the health organizations is very vast and complex due to which it is difficult to analyze the data, in order to make important decision regarding patient health. This data contains details regarding hospitals, patients, medical claims, treatment cost etc. The analysis of health data improves the healthcare by enhancing the performance of patient management tasks. The outcome of Data Mining technologies are to provide benefits to healthcare organization for grouping the patients having similar type of diseases or health issues so that healthcare organization provides effective treatments. It can also be useful for predicting the length of stay of patients in hospital, for medical diagnosis and making plan for effective information system management.

Recent technologies are used in medical field to enhance the medical services in cost effective manner. Data mining techniques are also used to analyze the various factors that are responsible for diseases for example type of food, different working environment, education level, living conditions, availability of pure water, health care services, cultural ,environmental and agricultural factors .[1]

Data mining can significantly improve decision making by discovering patterns and trends in large amounts of complex data. Several studies employed data mining approaches to discover the knowledge of relation between the measured parameters and prevention of Arteriovenous Fistula (AVF) failure. Temporal data mining techniques are studied for dialysis failure prediction. Data mining is considered in the medical settings of treatment and provided a brief review of state-of-the-art methods for predicting patient risk and survival of dialysis patients.

Kidney disease is a serious disease to people's lives and health. Renal failure is the end of kidney disease, and a serious threat to physical and mental health. Kidney Transplanted is an effective treatment means for renal failure patients. At present, kidney disease experts usually use statistical analysis method to study renal failure treatment rules, and make some progress. But mathematical statistical methods have not meet the current development of medical treatment very well, and have not given satisfactory solution for clinical treatment data analysis, renal failure clinical syndrome related factors analysis, dialysis and drug clinical efficacy evaluation, etc.

Therefore, we propose the use of data mining techniques for mining the patients' records after transplanting of kidney and seek for relationships and associations in the Sudanese Kidney Transplanted Association, which has about 2,700 members, records to help the patients during their prognosing stage.

1.2 Objectives

- Design National Kidney Transplantation Dataset.
- Investigate the tools that use Association Rule Mining.
- Mine for any probable special effects and diseases after kidney transplantation.

1.3 Methodology

1. Collecting the data of the members of the Sudanese Kidney Transplanted Association and the hospitals where its members are treated.
2. Follow the knowledge discovery process as structured in its various stages as in fig 1.1
    a. Data selection where data is collected from various sources,
    b. Preprocess the selected data.
    c. Transformation of the data into appropriate format for further processing.
3. Apply Association Rules Mining to the dataset.
4. Analyze and evaluate the generated rules with a physician

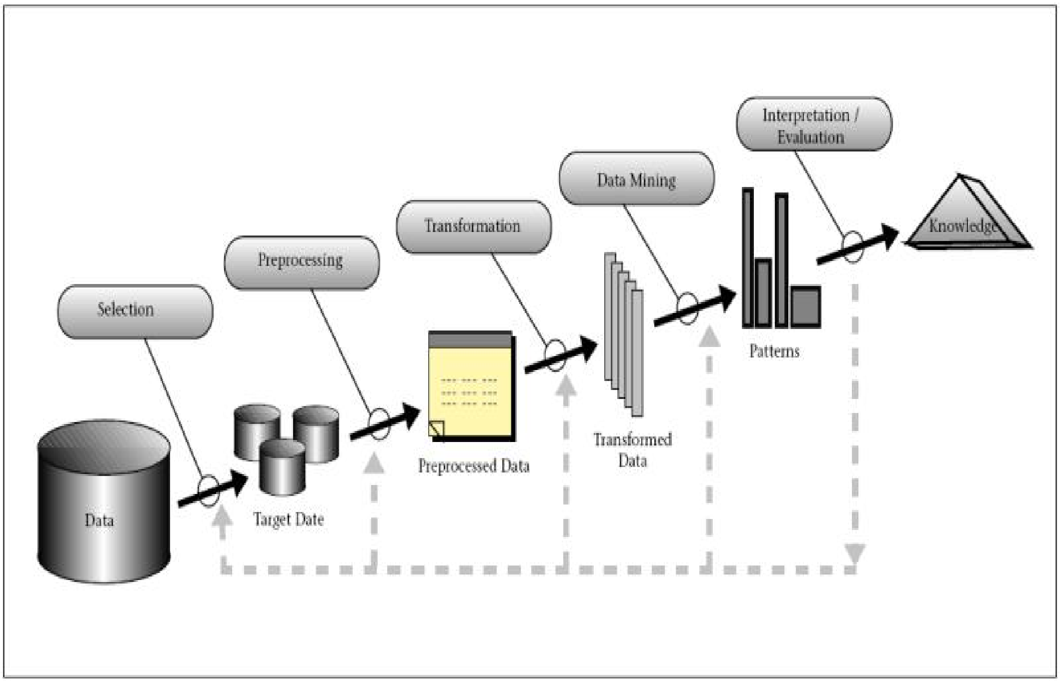


Figure 1-1: Knowledge Discovery in Database

CHAPTER TWO
DATA MINING

2 Data Mining Techniques

Data mining could be defined from several aspects. One is the process of managing data from databases for example from different perspectives and generating some useful information. Another is analyzing and establishing hidden relations and patterns in a group of data sets, yet another is knowledge discovery. In the case of Knowledge discovery, data mining digs through vast amounts of data, analyzes it, and tries to predict future trends and behavior.

2.1 Data Mining

Data Mining came into existence in the middle of 1990’s and appeared as a powerful tool that is suitable for fetching previously unknown pattern and useful information from huge dataset. Various studies highlighted that Data Mining techniques help the data holder to analyze and discover unsuspected relationship among their data which in turn helpful for making decision. In general, Data Mining and Knowledge Discovery in Databases (KDD) are related terms and are used interchangeably but many researchers assume that both terms are different as Data Mining is one of the most important stages of the KDD process.[2]

2.2 Data Mining Techniques

Data mining can be used in any organization that needs to find patterns or relationships in their data and it is increasing rapidly. Basically we have two data mining techniques; Predictive data mining it uses historical data to predict future theories. It merges database analysis with artificial intelligence. Predictive data mining is further categorized into Classification and Regression. Descriptive data mining it is to find patterns in data. They are generally used to create meaningful subgroups. Descriptive data mining is further classified into Clustering, Association and Sequential analysis.

2.2.1 Classification

Classification divides data samples into target classes. The classification technique predicts the target class for each data points. For example, patient can be classified as high risk or low risk patient on the basis of their disease pattern using data classification approach. It is a supervised learning approach having known class categories.

Binary and multilevel are the two methods of classification. In binary classification, only two possible classes such as, high or low risk patient may be considered while the multiclass approach has more than two targets for example, high, medium and low risk patient. Data set is partitioned as training and testing dataset. Using training dataset we train the classifier. Correctness of the classifier could be tested using test dataset. Classification is one of the most widely used methods of Data Mining in Healthcare organization. This information is very valuable for healthcare specialist to understand the cause of diseases and for providing better and cost effective treatment to patients. [3]

Classifications has many methods such as:

2.2.1.1 Decision Trees (DT)

Decision Trees is a classifier that use tree-like graph. The most common use of Decision Tree is in operations research analysis for calculating conditional probabilities. Using Decision Tree, decision makers can choose best alternative and traversal from root to leaf indicates unique class separation based on maximum information gain. Decision Tree is widely used by many researchers in healthcare field. A decision tree is a predictive model that, as its name implies, can be viewed as a tree. Specifically each branch of the tree is a classification question and the leaves of the tree are partitions of the dataset with their classification. [2]

2.2.1.2 Artificial Neural Networks(ANN)

True neural networks are biological systems that detect patterns, make predictions and learn. The artificial ones are computer programs implementing sophisticated pattern detection and machine learning algorithms on a computer to build predictive models from large historical databases. Artificial neural networks derive their name from their historical development which started off with the premise that machines could be made to think if scientists found ways to mimic the structure and functioning of the human brain on the computer. Thus historically neural networks grew out of the community of Artificial Intelligence rather than from the discipline of statistics. [2]

2.2.1.3 Support Vector Machine (SVM)

The concept of SVM is given by Vapnik, [2] which is based on statistical learning theory. SVMs were initially developed for binary classification but it could be efficiently extended for multiclass problems. The support vector machine classifier creates a hyper plane or multiple hyper planes in high dimensional space that is useful for classification, regression and other efficient tasks. SVM have many attractive features due to this it is gaining popularity and have promising empirical performance. SVM constructs a hyper plane in original input space to separate the data points. Some time it is difficult to perform separation of data points in original input space, so to make separation easier the original finite dimensional space mapped into new higher dimensional space.

2.2.1.4 Bayesian Methods

The classification based on Bayes theory is known as Bayesian classification. It is a simple classifier which is achieved by using classification algorithm. Bayes theorem provides basis for Naive Bayesian Classification and Bayesian Belief Networks (BBN). The main problem with Naïve Bayes Classifier is that it assumes that all attributes are independent with each other while in medical domain attributes such as patient symptoms and their health state are correlated with each other. In spite of assumption of attribute independence, Naïve Bayesian classifier has shown great performance in terms of accuracy so if attributes are independent with each other then we can use it in medical field.[2]

2.2.1.5 K-Nearest Neighbor (K-NN) K-Nearest

K-Nearest Neighbor (K-NN) classifier is one of the simplest classifier that discovers the unidentified data point using the previously known data points (nearest neighbor) and classified data points according to the voting system. K-NN classifies the data points using more than one nearest neighbor. K-NN has a number of applications in different areas such as health datasets, image field, cluster analysis, pattern recognition, online marketing etc.[2]

2.2.2 Regression

Regression is used to find out functions that explain the correlation among different variables. A mathematical model is constructed using training dataset. In statistical modeling two kinds of variables are used where one is called dependent variable and another one is called independent variable. And usually represented using 'Y' and 'X'. There is always one dependent variable while independent variable may be one or more than one.

Regression is a statistical method which investigates relationships between variables. By using Regression dependences of one variable upon others may be established. Based on number of independent variables regression is of two types, one is Linear and another one is Non-linear. Linear regression identifies relation of a dependent variable and one or more independent variables. It is based on a model which utilizes linear function for its construction. [2]

2.2.3 Clustering

Clustering is an unsupervised learning method that is different from classification, since it has no predefined classes. In clustering, large databases are separated into the form of small different subgroups or clusters. Clustering partitioned the data points based on the similarity measure. Clustering approach is used to identify similarities between data points. Each data points within the same cluster are having greater similarity as compare to the data points belongs to other cluster. Various clustering techniques are established and used over the last few decades. As pointed out earlier clustering need less or no information for analyzing the data. So it is mainly used for analyzing microarray data because very little details are available for genes. [2]

2.2.4 Association

Association is one of the most vital approaches of data mining that is used to find out the frequent patterns, interesting relationships among a set of data items in the data repository. It is also known as market basket analysis due to its capability of discovering the association among purchased item or unknown patterns of sales of customers in a transaction database. For example if a customer is buying a computer then the chance of buying antivirus software is high. This information helps the storekeeper to further enhance their sales. Association also has great impact in the healthcare field to detect the relationships among diseases, health state and symptoms. Healthcare organization widely used Association approach for discovering relationships between various diseases and drugs. It is also used for detecting fraud and abuse in health insurance. Association is also used with classification techniques to enhance the analysis capability of Data Mining. [2]

There are many association rule algorithms for mining frequent item sets such as

2.2.4.1 AIS

There was a real buzz in early 90s about how to emulate the biological immune system in the real world scenarios. The capacity of the immune system to proliferate cells that produce antibodies whenever it detects a high degree of matching with an antigen is, without doubts, fascinating. A series of algorithms were invented and new systems called artificial immune systems were designed. AIS algorithm uses candidate generation to detect the frequent item sets. The candidates are generated on the fly and are compared with previously found frequent item sets. The disadvantage of the algorithm is that it generates and counts too many candidate item sets that turn out to be small. AIS was the first algorithm that introduced the problem of generating association rules. [9]

2.2.4.2 Frequent Pattern - Growth

FP-Growth approach is based on divide and conquers strategy for producing the frequent item sets. FP- growth is mainly used for mining frequent item sets without candidate generation. Major steps in FP-growth is- Step1- It firstly compresses the database showing frequent item set in to FP-tree. FP-tree is built using 2 passes over the dataset. Step2: It divides the FP-tree in to a set of conditional database and mines each database separately, thus extract frequent item sets from FP-tree directly.

2.2.4.3 APRIORI

It is by far the most important data mining algorithms for mining frequent item sets and associations. It opened new doors and created new modalities to mine the data. Since its inception, many scholars have improved and optimized the Apriori algorithm and have presented new Apriori-like algorithms. The authors became living legends in the data mining communities. They both received masters and PhDs from University of Wisconsin, Madison and both worked for IBM. The IBM's Intelligent Miner was created mainly by them. Once colleagues, they now work for competing companies – Agrawal for Microsoft and Srikant for Google. Apriori uses a breadth-first search strategy to count the support of item sets and uses a candidate generation function which exploits the downward closure property of support. [9]

CHAPTER THREE

RENAL FAILURE

3 Renal Failure

Renal failure (kidney failure or renal insufficiency) is a medical condition in which the kidneys fail to adequately filter waste products from the blood. The two main forms of failure are acute kidney injury, which is often reversible with adequate treatment, and chronic kidney disease, which is often not reversible. In both cases, there is usually an underlying cause. Renal failure is mainly determined by a decrease in glomerular filtration rate, the rate at which blood is filtered in the glomeruli of the kidney. This is detected by a decrease in or absence of urine production or determination of waste products (creatinine or urea) in the blood. Depending on the cause, hematuria (blood loss in the urine) and proteinuria (protein loss in the urine) may be noted. In renal failure, there may be problems with increased fluid in the body (leading to swelling), increased acid levels, raised levels of potassium, decreased levels of calcium, increased levels of phosphate, and in later stages anemia. Bone health may also be affected.

3.1 Renal Failure Classification

Acute kidney and chronic kidney disease. The types of renal failure is determined by the trend in the serum creatinine. Other factors that may help differentiate acute kidney from chronic kidney disease include anemia and the kidney size on sonography. Chronic kidney generally leads to anemia and small kidney size.

3.1.1 Acute kidney

Acute kidney injury (AKI), previously called acute renal failure (ARF), is a rapidly progressive loss of renal function, generally characterized by oliguria (decreased urine production, quantified as less than 400 ml per day in adults, less than 0.5 mL/kg/h in children or less than 1 mL/kg/h in infants); and fluid and electrolyte imbalance. AKI can result from a variety of causes, generally classified as pre renal, intrinsic, and post renal. The underlying cause must be identified and treated to arrest the progress, and dialysis may be necessary to bridge the time gap required for treating these fundamental causes. It usually occurs when the blood supply to the kidneys is suddenly interrupted or when the kidneys become overloaded with toxins, or Drug overdoses accidental or from chemical overloads of drugs such as antibiotics or chemotherapy, and the crush syndrome, when large amounts of toxins are suddenly released in the blood circulation after a long compressed limb is suddenly relieved from the pressure obstructing the blood flow through its tissues, causing ischemia. The resulting overload can lead to the clogging and the destruction of the kidneys.

3.1.2 Chronic kidney

Chronic kidney disease (CKD), also known as chronic renal disease (CRD), is a progressive loss in renal function over a period of months or years. The symptoms of worsening kidney function are non-specific, and might include feeling generally unwell and experiencing a reduced appetite. Often, chronic kidney disease is diagnosed as a result of screening of people known to be at risk of kidney problems, such as those with high blood pressure or diabetes and those with a blood relative with it.

The most common causes of CKD are diabetes mellitus and long-term, uncontrolled hypertension. Polycystic kidney disease is another well-known cause of CKD. The majority of people afflicted with polycystic kidney disease have a family history of the disease. Other genetic illnesses affect kidney function, as well. Overuse of common drugs such as aspirin, ibuprofen, and acetaminophen (paracetamol) can also cause chronic kidney damage. Some infectious diseases, such as hantavirus, can attack the kidneys, causing kidney failure.

3.1.3 Acute-on-chronic renal failure

Acute kidney injuries can be present on top of chronic kidney disease, a condition called acute-on-chronic renal failure (AoCRF). The acute part of AoCRF may be reversible, and the goal of treatment, as with AKI, is to return the patient to baseline renal function, typically measured by serum creatinine. Like AKI, AoCRF can be difficult to distinguish from chronic kidney disease if the patient has not been monitored by a physician and no baseline (i.e., past) blood work is available for comparison.

3.2 Signs and Symptoms

Symptoms can vary from person to person. Someone in early stage kidney disease may not feel sick or notice symptoms as they occur. When kidneys fail to filter properly, waste accumulates in the blood and the body, a condition called azotemia. Very low levels of a zotaemia may produce few, if any, symptoms. If the disease progresses, symptoms become noticeable (if the failure is of sufficient degree to cause symptoms), the most famous are:

- High levels of urea in the blood, which can result in:
    - Vomiting and/or diarrhea, which may lead to dehydration
    - Nausea
    - Weight loss
    - Nocturnal urination
    - More frequent urination, or in greater amounts than usual, with pale urine
    - Less frequent urination, or in smaller amounts than usual, with dark colored urine
    - Blood in the urine
    - Pressure, or difficulty urinating
    - Unusual amounts of urination, usually in large quantities
- Buildup of phosphates in the blood that diseased kidneys cannot filter out may cause:
    - Itching
    - Bone damage
    - Nonunion in broken bones
    - Muscle cramps (caused by low levels of calcium which can be associated with hyperphosphatemia)
- Buildup of potassium in the blood that diseased kidneys cannot filter out (called hyperkalemia) may cause:
    - Abnormal heart rhythms
    - Muscle paralysis
- Failure of kidneys to remove excess fluid may cause:

  - Swelling of the legs, ankles, feet, face and/or hands
  - Shortness of breath due to extra fluid on the lungs (may also be caused by anemia)
- Polycystic kidney disease, which causes large, fluid-filled cysts on the kidneys and sometimes the liver, can cause:
  - Pain in the back or side
- Healthy kidneys produce the hormone erythropoietin that stimulates the bone marrow to make oxygen-carrying red blood cells. As the kidneys fail, they produce less erythropoietin, resulting in decreased production of red blood cells to replace the natural breakdown of old red blood cells. As a result, the blood carries less hemoglobin, a condition known as anemia. This can result in:
  - Feeling tired and/or weak
  - Memory problems
  - Difficulty concentrating
  - Dizziness
  - Low blood pressure
- Normally, proteins are too large to pass through the kidneys, however, they are able to pass through when the glomeruli are damaged. This does not cause symptoms until extensive kidney damage has occurred, after which symptoms include:
  - Foamy or bubbly urine
  - Swelling in the hands, feet, abdomen, or face
- Other symptoms include:
  - Appetite loss, a bad taste in the mouth
  - Difficulty sleeping
  - Darkening of the skin
  - Excess protein in the blood
  - With high dose penicillin, renal failure patients may experience seizures.

CHAPTER FOUR

LITERATURE REVIEW

4 Literature review

Data Mining is one of the most exciting area of research that is become increasingly popular in health organization. Data Mining plays an important role for uncovering new trends in healthcare organization which in turn n helpful for all the parties associated with this field. Table 4-1 explores the utility of various Data Mining techniques such as classification, clustering, association, regression in health domain. The previous studies about applying data mining techniques on kidney failure dataset are discussed afterwards.

Table 4-1: Previous Studies

Technique	Author	Field	Method
Classification	Hatice (2012)	Skin diseases	K-Nearest Neighbour (K-NN)
	Bestsimas (2008)	Cost of healthcare	Classification tree
	Jen (2012)	Chronic disease	K-Nearest Neighbour (K-NN) Linear Discriminate Analysis (LDA)
	Er (2010)	Chest diseases	Artificial Neural network (ANN)
	Liu (2012)	Decision support system for analyzing risks that are associated with health	Bayesian Belief Network (BBN)
	Potter (2007)	Breast cancer	Weka tool
	Moon (2012)	Exemplify the patterns of smoking in adults	Decision tree
	Huang (2008)	A predictive model for breast cancer diagnosis	Hybrid SVM based strategy
Clustering	Lenert	Health services	K-means Clustering

	Chipman (2009)	Analyzing microarray data	Hierarchical Clustering
	Research work	Extracts the useful and interesting patterns from biomedical images	Density Based Clustering
Association	Patil (2010) Ilayaraja (2013)	Classify the patients suffering from type-2 diabetes Discover frequent diseases in medical data	Apriori algorithm
	Noma (2012)	Identifying interesting patterns in medical audiology data	Frequent pattern tree algorithm

4.1 Classifications of Heart Disease

The study was performed in order to help doctors in diagnosing the heart diseases, they used Weka techniques explore interface experimenter interface knowledge flow interface and Clementine tool to analyse the data which is from UCI open repository for data sets and apply Decision tree and support vector machine and Naïve Bayes Classifier. The result shows that the Naïve Bayes Classifier is the best in diagnosing the heart disease data. [14]

4.2 Incidence of Renal Failure in Cardiovascular Surgery Patients

The study was occurred in Medical College and Hospital in Pimpri, Pune, Maharashtra, India for studying the incidence of real failure in cardiovascular surgery patients .[4]

The work was carried out of at a General Hospital over a period of 12 months. A total of 110 patient sunder went cardiovascular surgery over a period of 12 months of these 40 patients with renal failure were studied in details of history of hypertension, diabetes mellitus, previous renal insufficiency, ischemic heart disease, congestive cardiac failure and other co morbid illnesses were recorded.

Highest serum creatinine value in postoperative period was noted. They had chosen a serum creatinine value of greater than 50% over the baseline value after cardiovascular surgery as level for definition of renal failure in postoperative cardiovascular surgery.

The result showed that when age increases incidence of renal failure goes on increasing. Incidence of renal failure in present study was 36.36%. The researchers come up with a conclusion that older age, hypertension, diabetes mellitus, chronic heart failure and previous renal insufficiency were significant predisposing risk factors for development of postoperative renal failure after cardiovascular surgeries.

4.3 Early kidney failure prediction

The study was done in Iran University of Medical Sciences, Tehran. The goals of study was to extract pattern of early Arteriovenous Fistula (AVF)failure, predict this issue, and determine the high-risk factors on it, using data mining approaches. [5]

They merged two datasets and found eight similar parameters of them, where each patient is characterized by seven attributes. The last column thrombosis (yes: failure or no: survival) is the designated class attribute. They obtained a final dataset with 8 parameters of 193 records (patients), which contain 106 cases of failures and 87 cases of survivals.

They used Weka operators and performed many types of them and consulted the surgeon about obtained decision trees (DT). The result was Existence of either diabetes mellitus or smoking in HD patients increases early AVF failure in their surgery. They designed two applied methods and at first predicted risk factors of this complication with accuracy rates of 61.66%– 74.61%. Then they added data of side of AVF (location in hand) to the data and predicted this complication with accuracy rates between 67.91% and 75.13%. Results support the impressive roles of risk factors in AVF failures. They found that “diabetes mellitus,” “smoking,” and “hypertension” have important roles in early AVF failure, which are more effective roles than other factors such as age.

CHAPTER FIVE

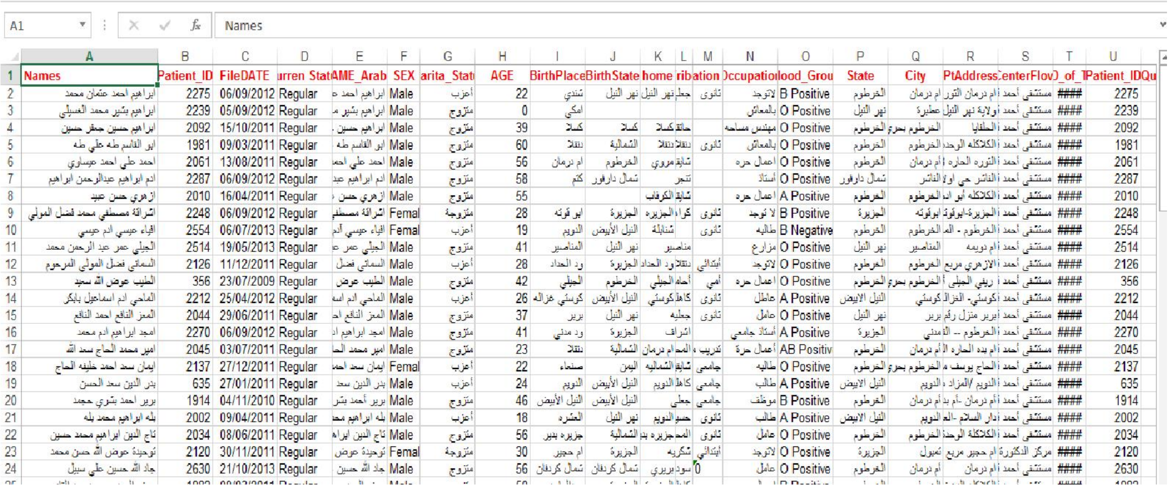
METHODOLOGY

5 Methodology

This part describe how data is collected, through processing following the KDD steps. The tools used and the details of the Apriori algorithm.

5.1 Dataset Collection

The data set was collected from Sudanese Kidney Transplanted Association which keeps records about the patients who transplanted a new kidney. Their database contains 2,780 records that include the basic information about the patient, such as name, age, sex and where they live. It also contains information about the patient's family history with the disease and other medical information about patients' health records, and if they have other chronic diseases such as hypertension or diabetes diagnosed that might lead to kidney failure or other genetic causes that may lead to kidney failure. The following figure shows the excel sheet that contains the patient's record.



A1 Names

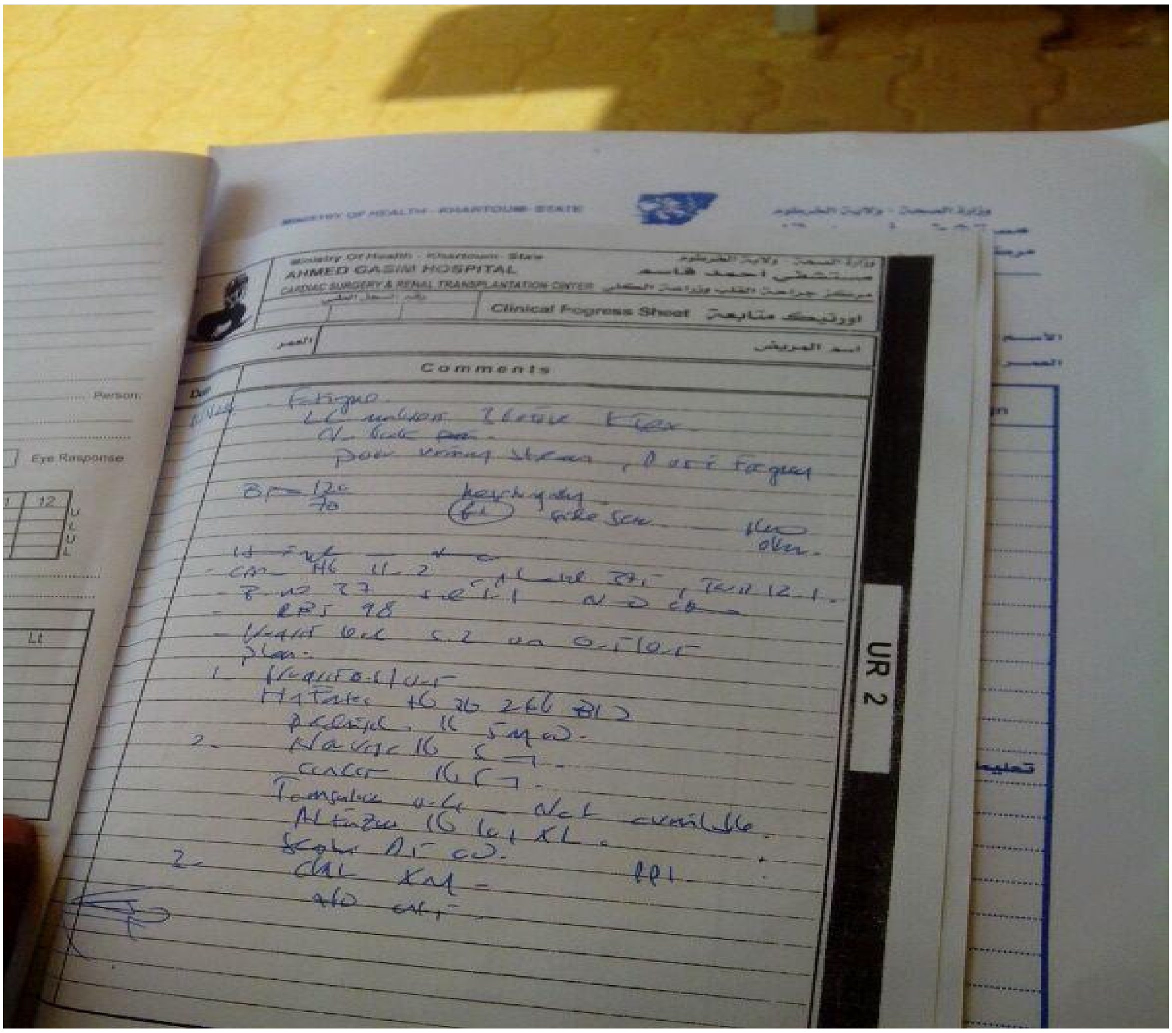
	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U
1	Names	Patient_ID	FileDATE	urren Stat	AME_Arab	SEX	arita_Stat	AGE	BirthPlace	BirthState	home	rib	ation	)ccupatio	lood_Grou	State	City	PtAddress	enterFlov	)_of_T	Patient_IDQu
2	ابراهيم احمد عثمان محمد	2275	06/09/2012	Regular	ابراهيم احمد ه	Male	أعزب	22	شندي	نهر النيل	نهر النيل	جعل	ثانوى	لاتوجد	B Positive	الخرطوم	ام درمان	أم درمان الثور	مستشفى أحمد	####	2275
3	ابراهيم بشير محمد الغسيلي	2239	05/09/2012	Regular	ابراهيم بشير م	Male	متزوج	0	امكي					بالمعاش	O Positive	نهر النيل	عطبرة	ولاية نهر النيل	مستشفى أحمد	####	2239
4	ابراهيم حسين جعفر حسين	2092	15/10/2011	Regular	ابراهيم حسين .	Male	متزوج	39	كسلا	كسلا	كسلا	حلنق		مهندس مساحه	O Positive	الخرطوم	الخرطوم بحري	الحلفايا	مستشفى أحمد	####	2092
5	ابو القاسم طه علي طه	1981	09/03/2011	Regular	ابو القاسم طه .	Male	متزوج	60	دنقلا	الشمالية	دنقلا	دنقلا	ثانوى	بالمعاش	O Positive	الخرطوم	الخرطوم	الكلاكله الوحد	مستشفى أحمد	####	1981
6	احمد علي احمد عيساوي	2061	13/08/2011	Regular	احمد علي احم	Male	متزوج	56	ام درمان	الخرطوم	مروي	شايق		اعمال حره	O Positive	الخرطوم	أم درمان	الثوره الحاره	مستشفى أحمد	####	2061
7	ادم ابراهيم عبدالرحمن ابراهيم	2287	06/09/2012	Regular	ادم ابراهيم عبد	Male	متزوج	58	كتم	شمال دارفور		تنجر		أستاذ	O Positive	شمال دارفور	الفاشر	الفاشر حي اولا	مستشفى أحمد	####	2287
8	ازهري حسن عبيد	2010	16/04/2011	Regular	ازهري حسن .	Male	متزوج	55			الكرفاب	شايق		اعمال حره	A Positive	الخرطوم	الخرطوم	الكلاكله أبو اد	مستشفى أحمد	####	2010
9	اشراقة مصطفي محمد فضل المولي	2248	06/09/2012	Regular	اشراقة مصطفي	Femal	متزوجة	28	ابو قوته	الجزيرة	الجزيره	كواه	ثانوى	لا توجد	B Positive	الجزيرة	ابوقوته	الجزيرة-ابوقوة	مستشفى أحمد	####	2248
10	افياء عيسي ادم عيسي	2554	06/07/2013	Regular	افياء عيسي آدم	Femal	أعزب	19	الدويم	النيل الأبيض	شنابلة		ثانوى	طالبه	B Negative	الخرطوم	الخرطوم	الخرطوم - الم	مستشفى أحمد	####	2554
11	الجيلي عمر عبد الرحمن محمد	2514	19/05/2013	Regular	الجيلي عمر ع	Male	متزوج	41	المناصير	نهر النيل		مناصير		مزارع	O Positive	نهر النيل	المناصير	ام دويمه	مستشفى أحمد	####	2514
12	السماني فضل المولي المرحوم	2126	11/12/2011	Regular	السماني فضل	Male	أعزب	28	ود الحداد	الجزيرة	ود الحداد	دنقلا	أبتدائي	لاتوجد	O Positive	الخرطوم	الخرطوم	الازهري مربع	مستشفى أحمد	####	2126
13	الطيب عوض الله سعيد	356	23/07/2009	Regular	الطيب عوض	Male	متزوج	42	الجيلي	الخرطوم	الجيلي	أحام	أمي	اعمال حره	O Positive	الخرطوم	الخرطوم بحري	ريفي الجيلى أ	مستشفى أحمد	####	356
14	الماحي ادم اسماعيل بابكر	2212	25/04/2012	Regular	الماحي ادم اسه	Male	أعزب	26	كوستي غزاله	النيل الأبيض	كوستي	كاهلا	ثانوى	عاطل	A Positive	النيل الابيض	كوستي	كوستي- الغزال	مستشفى أحمد	####	2212
15	المعز النافع احمد النافع	2044	29/06/2011	Regular	المعز النافع احـ	Male	متزوج	37	بربر	نهر النيل		جعليه	ثانوى	عامل	O Positive	نهر النيل	بربر	بربر منزل رقم	مستشفى أحمد	####	2044
16	امجد ابراهيم ادم محمد	2270	06/09/2012	Regular	امجد ابراهيم اد	Male	متزوج	41	ود مدني	الجزيرة		اشراف		أستاذ جامعي	A Positive	الجزيرة	مدني	الخرطوم -- ال	مستشفى أحمد	####	2270
17	امير محمد الحاج سعد الله	2045	03/07/2011	Regular	امير محمد الحا	Male	متزوج	23	دنقلا	الشمالية	ام درمان	المدا	تدريب ه	أعمال حرة	AB Positiv	الخرطوم	أم درمان	ام بده الحاره ا	مستشفى أحمد	####	2045
18	ايمان سعد احمد خليفه الحاج	2137	27/12/2011	Regular	ايمان سعد احمـ	Femal	أعزب	22	صنعاء	اليمن	الشماليه	شايق	جامعي	طالبه	O Positive	الخرطوم	الخرطوم بحري	الحاج يوسف م	مستشفى أحمد	####	2137
19	بدر الدين سعد الحسن	635	27/01/2011	Regular	بدر الدين سعد	Male	أعزب	24	الدويم	النيل الأبيض	الدويم	كاهلا	جامعي	طالب	A Positive	النيل الابيض	الدويم	الدويم /المزاد	مستشفى أحمد	####	635
20	برير احمد بشري محمد	1914	04/11/2010	Regular	برير أحمد بشر	Male	متزوج	46	النيل الأبيض	النيل الأبيض		جعلي	جامعي	موظف	B Positive	الخرطوم	أم درمان	أم درمان -أم بد	مستشفى أحمد	####	1914
21	بله ابراهيم محمد بله	2002	09/04/2011	Regular	بله ابراهيم محد	Male	أعزب	18	العشره	نهر النيل	الدويم	حسو	ثانوى	طالب	A Positive	النيل الابيض	الدويم	دار السلام -الع	مستشفى أحمد	####	2002
22	تاج الدين ابراهيم محمد حسين	2034	08/06/2011	Regular	تاج الدين ابراه	Male	متزوج	56	جزيره بدين	الشمالية	جزيره بد	المح	ثانوى	عامل	O Positive	الخرطوم	الخرطوم	الكلاكلة الوحد	مستشفى أحمد	####	2034
23	توحيدة عوض الله حسن محمد	2120	30/11/2011	Regular	توحيدة عوض	Femal	متزوجة	30	ام حجير	الجزيرة		شكريه	أبتدائي	لاتوجد	O Positive	الجزيرة	تمبول	ام حجير مربع	مركز الدكتورة	####	2120
24	جاد الله حسين علي سبيل	2630	21/10/2013	Regular	جاد الله حسين	Male	متزوج	56	شمال كردفان	شمال كردفان	بربري	سود	0	عامل	O Positive	الخرطوم	أم درمان	أم درمان	مستشفى أحمد	####	2630

Figure 5-1: SKTA Patient’s Records

We choose from the last five years data 1,116 records, which turns to be 326 records after preprocessing. Then we collect the follow up data of these patients from Ahmed Qassim Hospital where the treatment after the transplantation of kidney were occurred in Cardiac Surgery & Renal Transplantation Center.

Follow-up data for renal transplant Include patient visits and tests carried out at each visit and stored in a file called patient's file which is all patient's information after the operation and doctor's comments about the current state of his health.

The patient's files are secure only the doctors can read and access them, because this files are not electronically stored, we came at the same days of patient visits and doctors write the new tests on the files and read the old tests for us. The following figures illustrates the patient's file first pages:



MINISTRY OF HEALTH - KHARTOUM STATE وزارة الصحة - ولاية الخرطوم

Ministry Of Health - Khartoum State وزارة الصحة - ولاية الخرطوم
AHMED GASIM HOSPITAL مستشفى احمد قاسم
CARDIAC SURGERY & RENAL TRANSPLANTATION CENTER مركز جراحة القلب وزراعة الكلى

Clinical Fogress Sheet اورنيك متابعة

العمر | اسم المريض

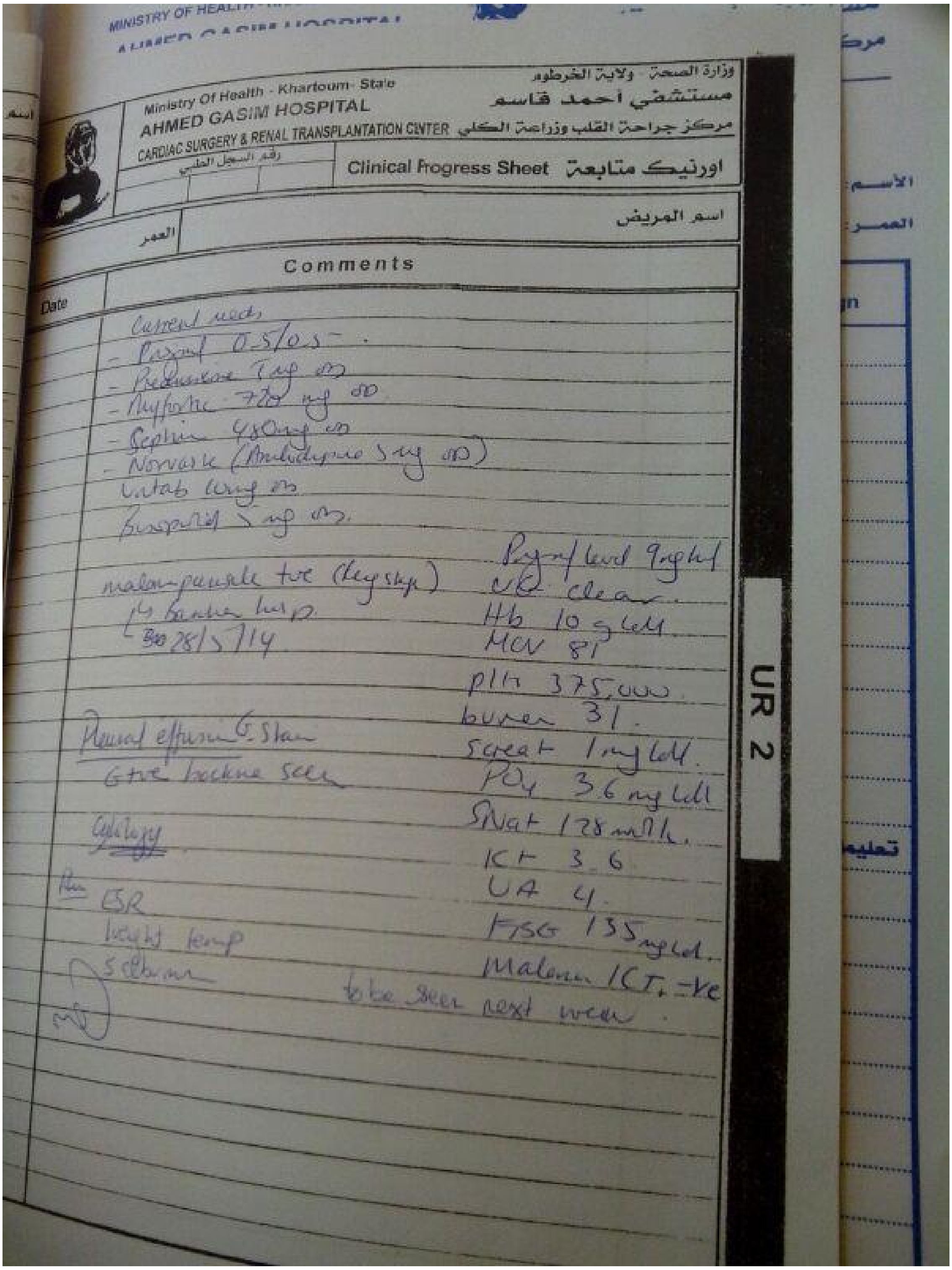
Comments

Date

Fatigue.
[illegible]
poor [illegible], [illegible] fatigue
BP 120/70
[illegible]
- CA HG 11.2 [illegible]
- RBS 98
[illegible]
plan.
1. [illegible]
2. [illegible]
Tamsulox 0.4 not available.
[illegible]
2. [illegible]

UR 2

Figure 5-2: Patient's File, page 1



وزارة الصحة - ولاية الخرطوم
مستشفي أحمد قاسم
مركز جراحة القلب وزراعة الكلي

Ministry Of Health - Khartoum- State
AHMED GASIM HOSPITAL
CARDIAC SURGERY & RENAL TRANSPLANTATION CENTER

رقم السجل الطبي

Clinical Progress Sheet اورنيك متابعة

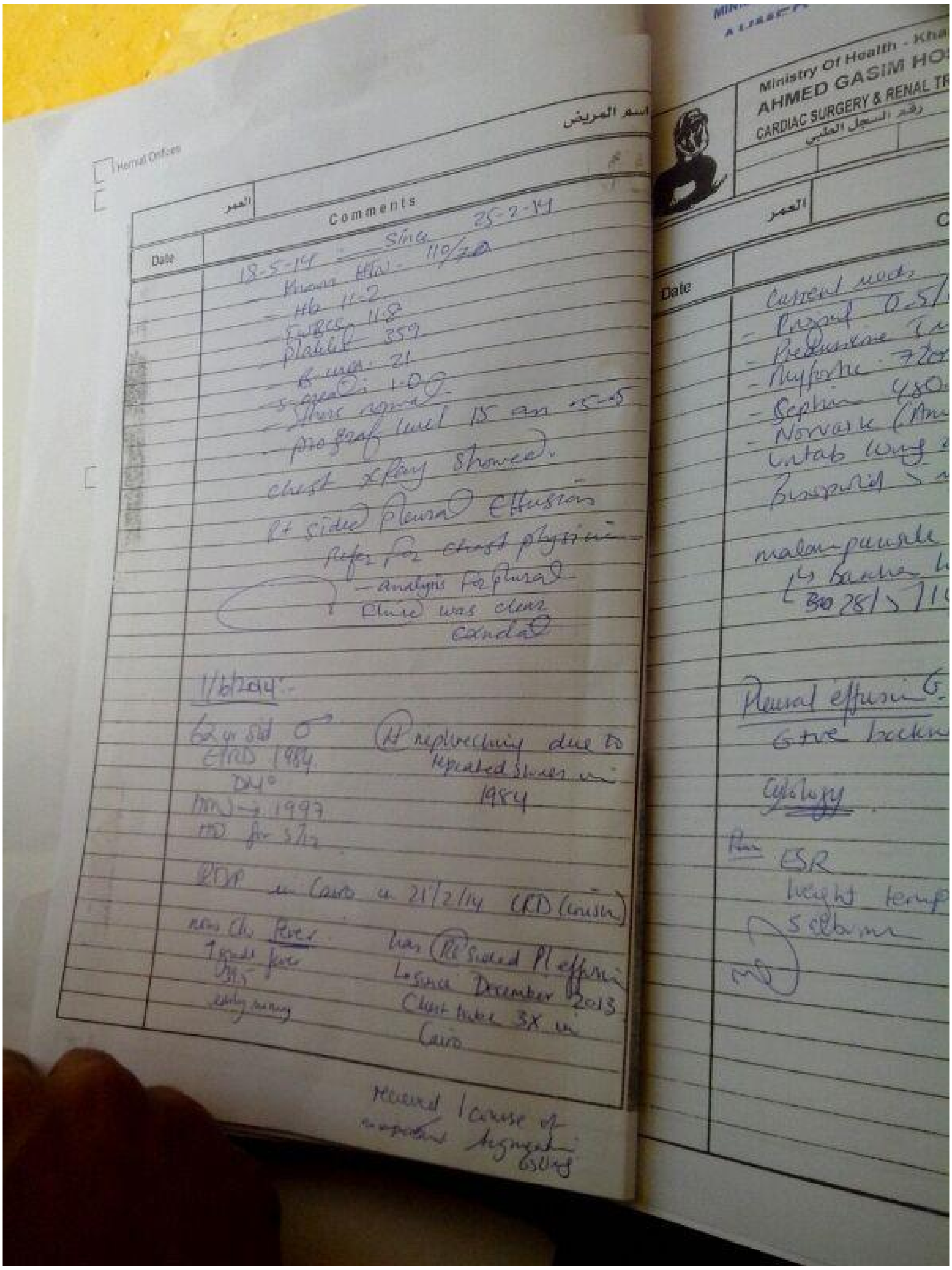
اسم المريض

العمر

UR 2

Date	Comments
	Current meds,
	- Prograf 0.5/0.5.
	- Prednisone 7mg od
	- Myfortic 720 mg BD
	- Septrin 480mg od
	- Norvask (Amlodipine 5mg od)
	Untab 10mg od
	Bisoprolol 5mg od.
	Prograf level 9ng/ml
	malaria-parasite +ve (ring stage) U/E clear.
	↳ Banker hosp. Hb 10 g/dl
	28/5/14 MCV 81
	plt 375,000.
	b.urea 31.
	Pleural effusion G. Stain S.creat 1mg/dl.
	G+ve bacteria seen PO4 3.6 mg/dl
	S.Na+ 128 mmol/L.
	Cytology K+ 3.6
	UA 4.
	Plan ESR FSG 135 mg/dl.
	weight temp
	S.albumin Malaria ICT. -ve
	to be seen next week.

Figure 5-3: Patient's File, page 2



اسم المريض

Hernal Orifices

العمر

Comments

Date

25-2-14

18-5-14 ___ Sing

Khana HTN - 110/[illegible]

- Hb 11.2
- TWBCs 11.8
- platelet 359
- B urea 21
- S creat: 1.09
- Liver normal
- prograf level 15 on 15-5

chest Xray showed.

Rt sided Pleural Effusion

Refer for chest physician

- analysis for pleural

fluid was clear

[illegible]

1/6/2014

62 yr old ♂ (Rt) nephrectomy due to repeated stones in 1984

ESRD 1984

DM°

HTN → 1997

HD for 5/12

RTx in Cairo in 21/2/14 LRD (cousin)

no c/o fever. has (Rt) sided Pl effusion

[illegible] fever since December 2013

38.5 chest tube 3X in Cairo

[illegible]

received 1 course of [illegible] [illegible]

Ministry Of Health - Kha

AHMED GASIM HO

CARDIAC SURGERY & RENAL TR

رقم السجل الطبي

العمر

Date

Current med,

- Prograf 0.5/
- Prednisone
- Myfortic 720
- Septrin 480
- Norvask (Am
- Untab [illegible]
- Bisoprolol 5

malan-parade

[illegible]

[BD 28/5] 11

Pleural effusion

Gve backw

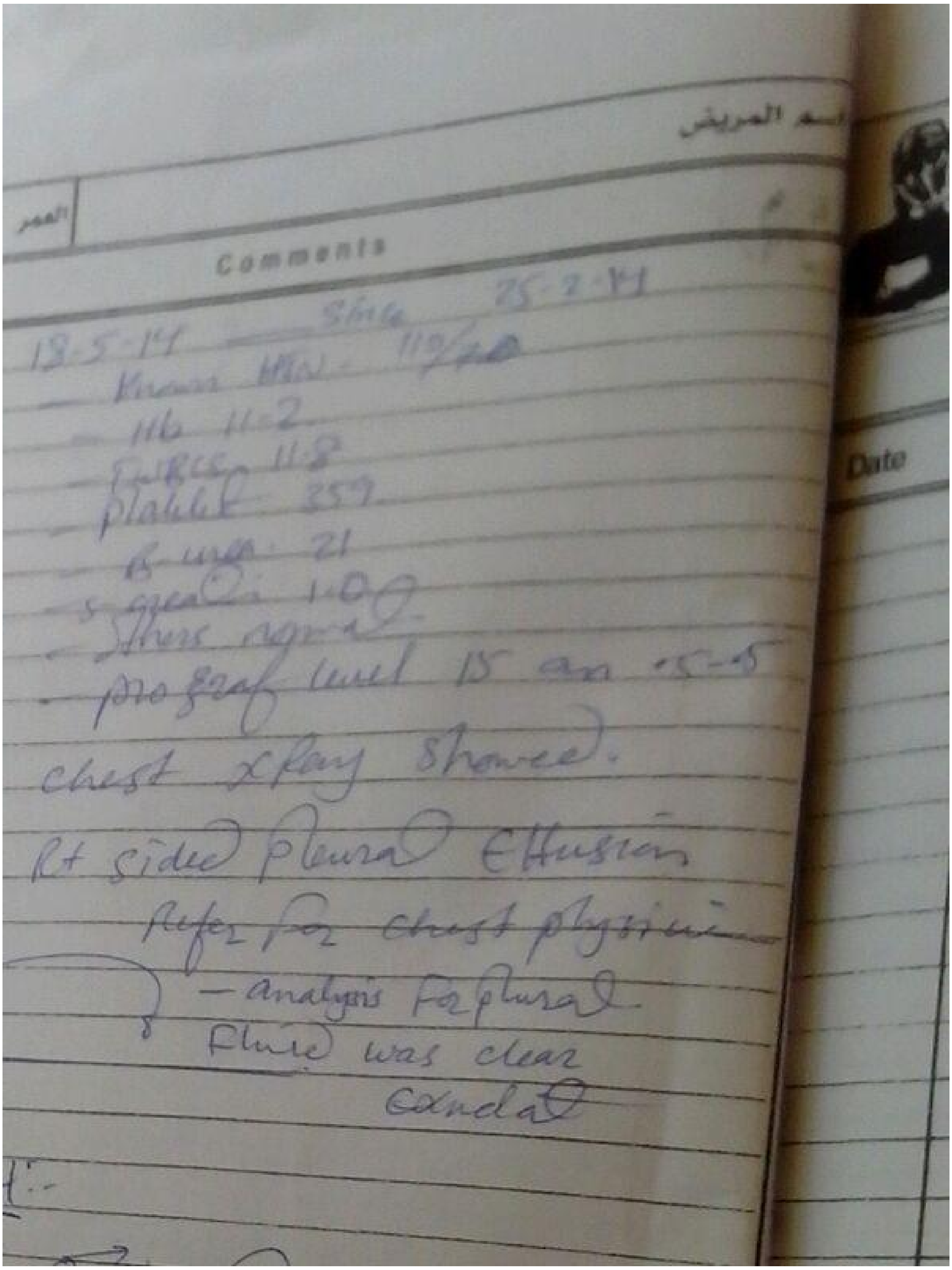
Cytology

Plan ESR

height temp

S albumin

Figure 5-4: Patient’s File, page 3



اسم المريض

العمر

Comments

Date

18-5-14 —— Sma 25-2-41

— Known HTN - 110/80

— Hb 11.2

— WBCs 11.8

— platelet 359

— B urea 21

— S creat 1.00

— Others normal

— prograf level 15 on •5-•5

chest xray showed.

Rt sided pleural Effusion

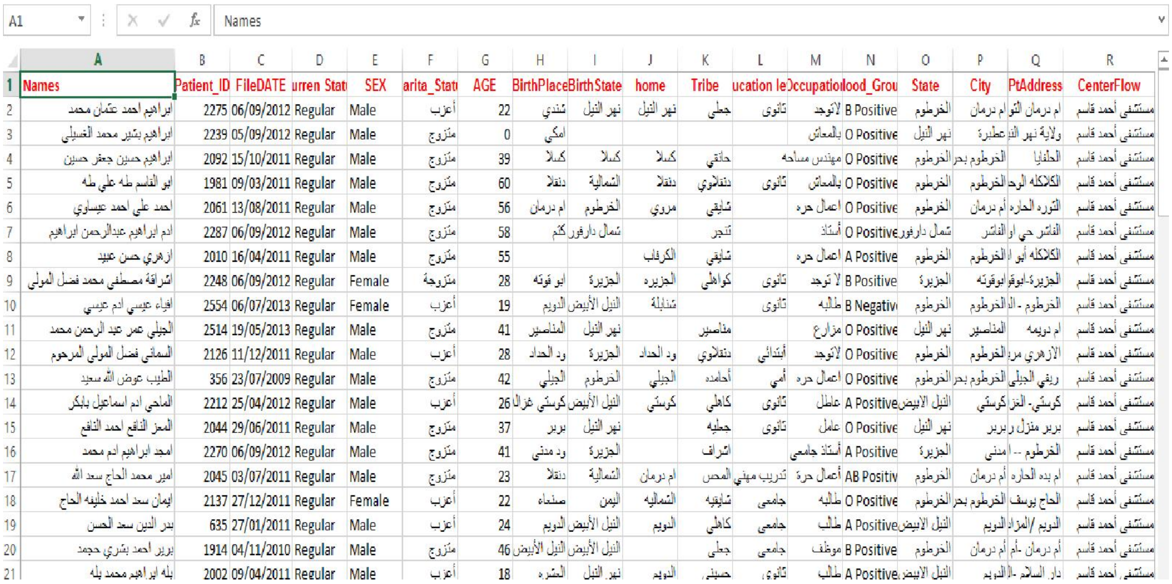
Refer for chest physician

— analysis for pleural fluid was clear exudat

Figure 5-5: Patient's File, page 4

We have read the files of patients who attend regularly visits and their number was 326 files, we were able to extract data from 100 files of them and we have integrated these data with Sudanese Kidney Transplanted Association excel sheet. And we've created a excel file with 100 records contains basic patient information's and other data we have obtained from the Sudanese Kidney Transplanted Association with follow-up data from the Ahmed Qassim for that patient.

The following figure illustrates the new Excel file



A1 | Names

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R
1	Names	Patient_ID	FileDATE	urren Stat	SEX	arita_Stat	AGE	BirthPlace	BirthState	home	Tribe	ucation le	Occupatio	lood_Grou	State	City	PtAddress	CenterFlow
2	ابراهيم احمد عثمان محمد	2275	06/09/2012	Regular	Male	أعزب	22	شندي	نهر النيل	نهر النيل	جعلي	ثانوى	لاتوجد	B Positive	الخرطوم	ام درمان	ام درمان الثو	مستشفى أحمد قاسم
3	ابراهيم بشير محمد الغسيلي	2239	05/09/2012	Regular	Male	متزوج	0	امكي					بالمعاش	O Positive	نهر النيل	عطبرة	ولاية نهر النب	مستشفى أحمد قاسم
4	ابراهيم حسين جعفر حسين	2092	15/10/2011	Regular	Male	متزوج	39	كسلا	كسلا	كسلا	حلنقي		مهندس مساحه	O Positive	الخرطوم	الخرطوم بحر	الحلفايا	مستشفى أحمد قاسم
5	ابو القاسم طه علي طه	1981	09/03/2011	Regular	Male	متزوج	60	دنقلا	الشمالية	دنقلا	دنقلاوي	ثانوى	بالمعاش	O Positive	الخرطوم	الخرطوم	الكلاكله الوحد	مستشفى أحمد قاسم
6	احمد علي احمد عيساوي	2061	13/08/2011	Regular	Male	متزوج	56	ام درمان	الخرطوم	مروي	شايقي		اعمال حره	O Positive	الخرطوم	أم درمان	الثوره الحاره	مستشفى أحمد قاسم
7	ادم ابراهيم عبدالرحمن ابراهيم	2287	06/09/2012	Regular	Male	متزوج	58	كتم	شمال دارفور		تنجر		أستاذ	O Positive	شمال دارفور	الفاشر	الفاشر حي او	مستشفى أحمد قاسم
8	ازهري حسن عبيد	2010	16/04/2011	Regular	Male	متزوج	55			الكرفاب	شايقي		اعمال حره	A Positive	الخرطوم	الخرطوم	الكلاكله أبو ا	مستشفى أحمد قاسم
9	اشراقة مصطفى محمد فضل المولى	2248	06/09/2012	Regular	Female	متزوجة	28	ابو قوته	الجزيرة	الجزيره	كواهلي	ثانوى	لا توجد	B Positive	الجزيرة	ابوقوته	الجزيرة-ابوقو	مستشفى أحمد قاسم
10	افياء عيسي ادم عيسي	2554	06/07/2013	Regular	Female	أعزب	19	الدويم	النيل الأبيض	شنابلة		ثانوى	طالبه	B Negativ	الخرطوم	الخرطوم	الخرطوم - ال	مستشفى أحمد قاسم
11	الجيلي عمر عبد الرحمن محمد	2514	19/05/2013	Regular	Male	متزوج	41	المناصير	نهر النيل		مناصير		مزارع	O Positive	نهر النيل	المناصير	ام دويمه	مستشفى أحمد قاسم
12	السماني فضل المولى المرحوم	2126	11/12/2011	Regular	Male	أعزب	28	ود الحداد	الجزيرة	ود الحداد	دنقلاوي	أبتدائي	لاتوجد	O Positive	الخرطوم	الخرطوم	الازهري مرب	مستشفى أحمد قاسم
13	الطيب عوض الله سعيد	356	23/07/2009	Regular	Male	متزوج	42	الجيلي	الخرطوم	الجيلي	أحامده	أمي	اعمال حره	O Positive	الخرطوم	الخرطوم بحر	ريفي الجيلي	مستشفى أحمد قاسم
14	الماحي ادم اسماعيل بابكر	2212	25/04/2012	Regular	Male	أعزب	26	كوستي غزال	النيل الأبيض	كوستي	كاهلي	ثانوى	عاطل	A Positive	النيل الابيض	كوستي	كوستي- الغزا	مستشفى أحمد قاسم
15	المعز النافع احمد النافع	2044	29/06/2011	Regular	Male	متزوج	37	بربر	نهر النيل		جعليه	ثانوى	عامل	O Positive	نهر النيل	بربر	بربر منزل ر	مستشفى أحمد قاسم
16	امجد ابراهيم ادم محمد	2270	06/09/2012	Regular	Male	متزوج	41	ود مدني	الجزيرة		اشراف		أستاذ جامعي	A Positive	الجزيرة	مدني	الخرطوم -- ا	مستشفى أحمد قاسم
17	امير محمد الحاج سعد الله	2045	03/07/2011	Regular	Male	متزوج	23	دنقلا	الشمالية	ام درمان	المحس	تدريب مهني	أعمال حرة	AB Positiv	الخرطوم	أم درمان	ام بده الحاره	مستشفى أحمد قاسم
18	ايمان سعد احمد خليفه الحاج	2137	27/12/2011	Regular	Female	أعزب	22	صنعاء	اليمن	الشماليه	شايقيه	جامعي	طالبه	O Positive	الخرطوم	الخرطوم بحر	الحاج يوسف	مستشفى أحمد قاسم
19	بدر الدين سعد الحسن	635	27/01/2011	Regular	Male	أعزب	24	الدويم	النيل الأبيض	الدويم	كاهلي	جامعي	طالب	A Positive	النيل الابيض	الدويم	الدويم /المزاد	مستشفى أحمد قاسم
20	برير احمد بشري حجمد	1914	04/11/2010	Regular	Male	متزوج	46	النيل الأبيض	النيل الأبيض		جعلي	جامعي	موظف	B Positive	الخرطوم	أم درمان	أم درمان -أم	مستشفى أحمد قاسم
21	بله ابراهيم محمد بله	2002	09/04/2011	Regular	Male	أعزب	18	الحشره	نهر النيل	الدويم	حسيني	ثانوى	طالب	A Positive	النيل الابيض	الدويم	دار السلام -ا	مستشفى أحمد قاسم

Figure 5-6: Basic and Follow-up Data

5.2 Data Set Description

Title:

Renal Failure Transplanted Data.

Sources:

Sudanese Kidney Transplanted Association
Cardiac surgery & renal transplantation center – Ahmed Qassim Hospital

Data Set Information

This data set includes the basic information fields about the members of the Sudanese Kidney Transplanted Association, and follow up data from Ahmed Qassim Hospital. It is includes 9 attributes for basic information and 8 attributes for follow up data.

Attribute Information:

Basic Information Data

1. Age :

Table 5-1: Clementine Age Values

Range	**0-14**	**15-21**	**22-40**	**41-60**	**61-120**	
Code	Child	Boy	Youth	Adult	Old	

2. Sex :

   Male Female

3. Current Statues :

   Alive Death Unknown

4. Blood group

Blood Groups	A+	B+	AB+	O+	A-	B-	AB-	O-

5. State/Residence
6. Duration after transplantation

   1 2 3

7. Relevance of patient

   Count – Type of relevance

8. The infect by Renal Failure Result form

   Blood Pressure, Diabetes, Malaria, Kidney infections, Atrophy, Other, unknown.

9. Other Disease

   Hypertensions, Diabetes, Others, Both, None

Follow up Data:-

1. Blood pressure

Missed values: 11
Values: Low-Normal-High

Represent: T F F/ F T F/ F F T

Missed Values Represent: F T F as Normal.

Table 5-2: Clementine Blood Pressure Category

Blood Pressure Category	**Systolic** mm Hg (upper #)		**Diastolic** mm Hg (lower #)
Normal	less than **120**	And	less than **80**
Prehypertension	**120 – 139**	Or	**80 – 89**
High Blood Pressure (Hypertension) **Stage 1**	**140 – 159**	Or	**90 – 99**
High Blood Pressure (Hypertension) **Stage 2**	**160** or higher	Or	**100** or higher
Hypertensive Crisis **(Emergency care needed)**	Higher than **180**	Or	Higher than **110**

2. Hemoglobin

Normal: Men 14-18, Women 12-16

Missed values: 4

Values: Low-Normal-High

Represent: T F F/ F T F/ F F T

Missed Values Represent: F T F as Normal.

3. White Blood Cells

Normal: Adults 4-10, Childs to 12

Missed values: 19

Values: Low-Normal-High

Represent: T F F/ F T F/ F F T

Missed Values Represent: F T F as Normal.

4. Platelets

Normal: 150-290

Missed values: 20

Values: Low-Normal-High

Represent: T F F/ F T F/ F F T

Missed Values Represent: F T F as Normal.

5. Creatinine

Normal: 0.1-1.4

Missed values: 3

Values: Low-Normal-High

Represent: T F F/ F T F/ F F T

Missed Values Represent: F T F as Normal.

6. Uric acid

Normal: 2-7

Missed values: 74

Values: Low-Normal-High

Represent: T F F/ F T F/ F F T

Missed Values Represent: F T F as Normal.

7. Urea

Normal: 15-45

Missed values: 28

Values: Low-Normal-High

Represent: T F F/ F T F/ F F T

Missed Values Represent: F T F as Normal.

8. Prograf

Normal: 5-9

Missed values: 45

Values: Low-Normal-High

Represent: T F F/ F T F/ F F T

Missed Values Represent: F T F as Normal.

5.3 Apriori Algorithm

In data mining, Apriori is a classic algorithm for learning association rules. Apriori is designed to operate on databases containing transactions (for example, collections of items bought by customers, or details of a website frequentation). [10]

Other algorithms are designed for finding association rules in data having no transactions (Winepi and Minepi), or having no timestamps (DNA sequencing).

The whole point of the algorithm (and data mining, in general) is to extract useful information from large amounts of data. For example, the information that a customer who purchases a keyboard also tends to buy a mouse at the same time is acquired from the association rule below:

Support: The percentage of task-relevant data transactions for which the pattern is true.

Support (Keyboard -> Mouse) =

$$\frac{\text{Number of transactions containing both keyboard and mouse}}{\text{Number of total transactions}}$$

Confidence: The measure of certainty or trustworthiness associated with each discovered pattern.

Confidence (Keyboard -> Mouse) =

$$\frac{\text{Number of transactions containing both keyboard and mouse}}{\text{Number of transactions containing (keyboard)}}$$

The algorithm aims to find the rules which satisfy both a minimum support threshold and a minimum confidence threshold (Strong Rules). [10]

Item: article in the basket, Item set: a group of items purchased together in a single transaction.

5.3.1 Algorithm

1. Find all frequent item sets:

- ○ Get frequent items: Items whose occurrence in database is greater than or equal to the minimum support threshold.
- ○ Get frequent item sets: Generate candidates from frequent items. Prune the results to find the frequent item sets.

2. Generate strong association rules from frequent item sets

- ○ Rules which satisfy the minimum support and minimum confidence threshold.

5.3.2 High level Design

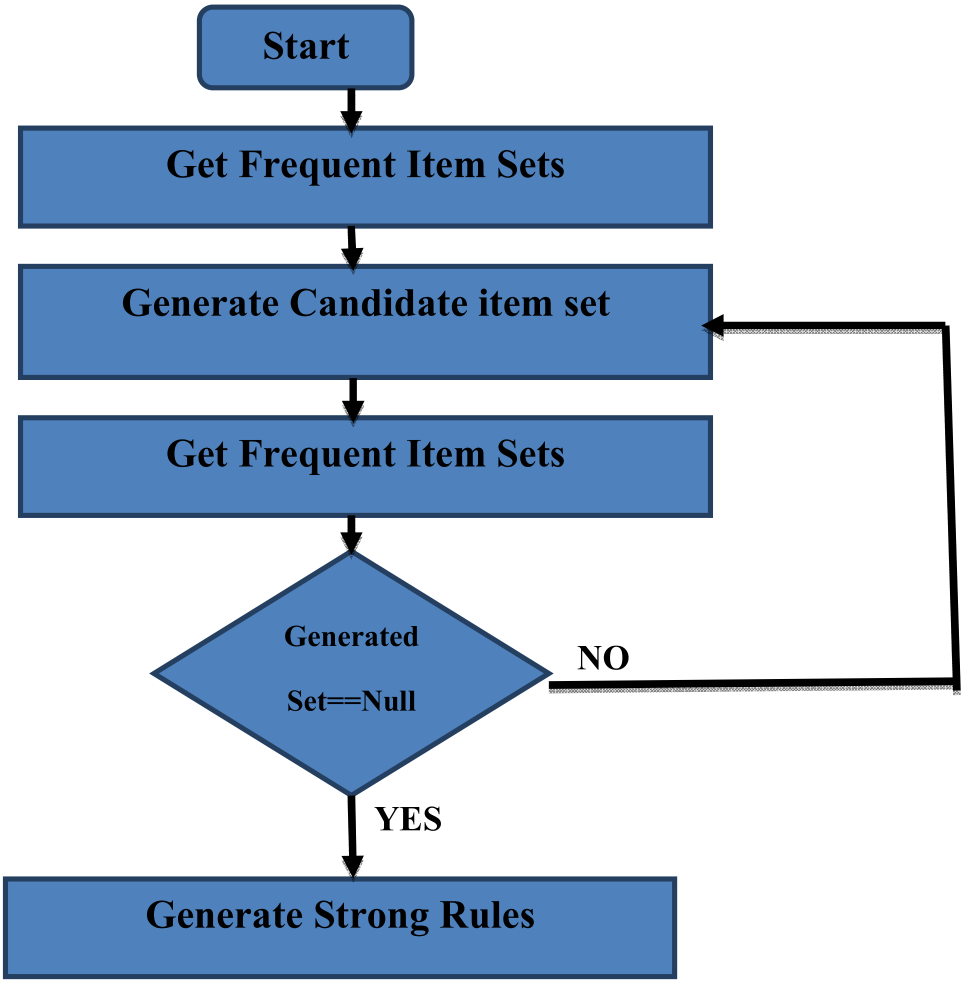
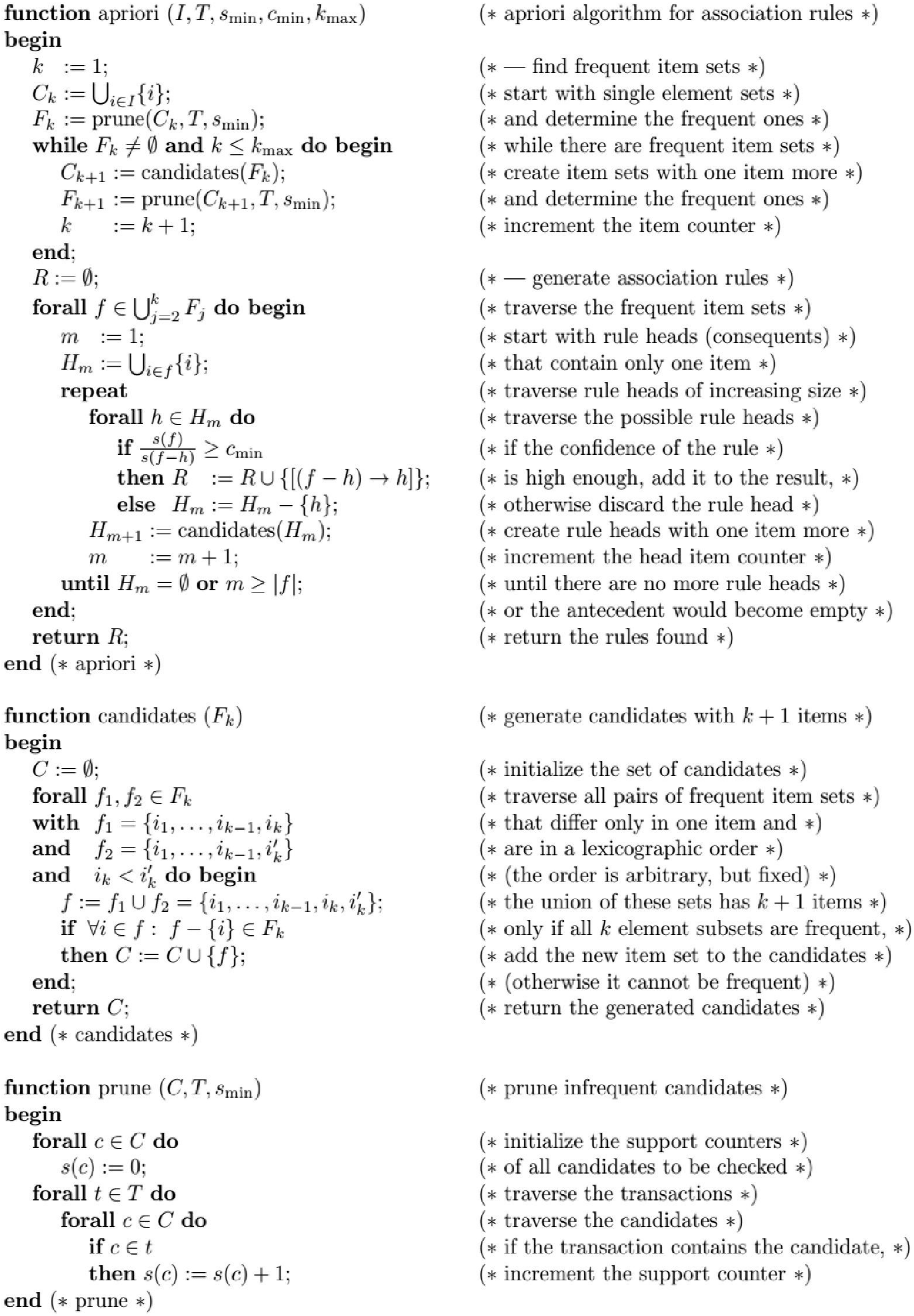


Figure 5.8: High level design

5.3.3 Algorithm Pseudo Code



```
function apriori ($I, T, s_{\min}, c_{\min}, k_{\max}$)      (* apriori algorithm for association rules *)
begin
  $k$ := 1;                                                  (* — find frequent item sets *)
  $C_k := \bigcup_{i \in I}\{i\}$;                           (* start with single element sets *)
  $F_k :=$ prune($C_k, T, s_{\min}$);                        (* and determine the frequent ones *)
  while $F_k \neq \emptyset$ and $k \leq k_{\max}$ do begin  (* while there are frequent item sets *)
    $C_{k+1} :=$ candidates($F_k$);                          (* create item sets with one item more *)
    $F_{k+1} :=$ prune($C_{k+1}, T, s_{\min}$);              (* and determine the frequent ones *)
    $k := k + 1$;                                            (* increment the item counter *)
  end;
  $R := \emptyset$;                                          (* — generate association rules *)
  forall $f \in \bigcup_{j=2}^{k} F_j$ do begin              (* traverse the frequent item sets *)
    $m$ := 1;                                                (* start with rule heads (consequents) *)
    $H_m := \bigcup_{i \in f}\{i\}$;                         (* that contain only one item *)
    repeat                                                   (* traverse rule heads of increasing size *)
      forall $h \in H_m$ do                                  (* traverse the possible rule heads *)
        if $\frac{s(f)}{s(f-h)} \geq c_{\min}$               (* if the confidence of the rule *)
        then $R := R \cup \{[(f-h) \to h]\}$;                (* is high enough, add it to the result, *)
        else $H_m := H_m - \{h\}$;                           (* otherwise discard the rule head *)
      $H_{m+1} :=$ candidates($H_m$);                        (* create rule heads with one item more *)
      $m := m + 1$;                                          (* increment the head item counter *)
    until $H_m = \emptyset$ or $m \geq |f|$;                 (* until there are no more rule heads *)
  end;                                                       (* or the antecedent would become empty *)
  return $R$;                                                (* return the rules found *)
end (* apriori *)

function candidates ($F_k$)                                  (* generate candidates with $k+1$ items *)
begin
  $C := \emptyset$;                                          (* initialize the set of candidates *)
  forall $f_1, f_2 \in F_k$                                  (* traverse all pairs of frequent item sets *)
  with $f_1 = \{i_1, \ldots, i_{k-1}, i_k\}$                 (* that differ only in one item and *)
  and  $f_2 = \{i_1, \ldots, i_{k-1}, i'_k\}$                (* are in a lexicographic order *)
  and  $i_k < i'_k$ do begin                                 (* (the order is arbitrary, but fixed) *)
    $f := f_1 \cup f_2 = \{i_1, \ldots, i_{k-1}, i_k, i'_k\}$;   (* the union of these sets has $k+1$ items *)
    if $\forall i \in f: \ f - \{i\} \in F_k$                (* only if all $k$ element subsets are frequent, *)
    then $C := C \cup \{f\}$;                                (* add the new item set to the candidates *)
  end;                                                       (* (otherwise it cannot be frequent) *)
  return $C$;                                                (* return the generated candidates *)
end (* candidates *)

function prune ($C, T, s_{\min}$)                            (* prune infrequent candidates *)
begin
  forall $c \in C$ do                                        (* initialize the support counters *)
    $s(c) := 0$;                                             (* of all candidates to be checked *)
  forall $t \in T$ do                                        (* traverse the transactions *)
    forall $c \in C$ do                                      (* traverse the candidates *)
      if $c \in t$                                           (* if the transaction contains the candidate, *)
      then $s(c) := s(c) + 1$;                               (* increment the support counter *)
end (* prune *)
```

5.4 Tools used

We used three tools to test the Apriori algorithm clementine, matrix laboratory, ARMDA association rule mining tool.

5.4.1 Matrix Laboratory (MATLAB)

MATLAB is a multi-paradigm numerical computing environment and fourth generation programming language It is allows matrix manipulations, plotting of functions and data, implementation of algorithms, creation of user interfaces, and interfacing with programs written in other languages, including C, C++, Java, and Fortran, It is intended primarily for numerical computing, an optional toolbox uses the MuPAD symbolic engine, allowing access to symbolic computing capabilities. An additional package, Simulink, adds graphical multi-domain simulation and Model-Based Design for dynamic and embedded systems.

5.4.1.1 Syntax

The MATLAB application is built around the MATLAB language, and most use of MATLAB involves typing MATLAB code into the Command Window (as an interactive mathematical shell), or executing text files containing MATLAB codes, including scripts and/or functions.

5.4.1.2 Structures

MATLAB has structure data types. Since all variables in MATLAB are arrays, a more adequate name is "structure array", where each element of the array has the same field names. In addition, MATLAB supports dynamic field names (field look-ups by name, field manipulations, etc.). Unfortunately, MATLAB JIT does not support MATLAB structures, therefore just a simple bundling of various variables into a structure will come at a cost.

5.4.1.3 Interfacing with other languages

MATLAB can call functions and subroutines written in the C programming language or FORTRAN. A wrapper function is created allowing MATLAB data types to be passed and returned. The dynamically loadable object files created by compiling such functions are termed "MEX-files" (for MATLAB executable).

Libraries written in Perl, Java, ActiveX or .NET can be directly called from MATLAB, and many MATLAB libraries (for example XML or SQL support) are implemented as wrappers around Java or ActiveX libraries. Calling MATLAB from Java is more complicated, but can be done with a MATLAB toolbox which is sold separately by MathWorks, or using an undocumented mechanism called JMI (Java-to-MATLAB Interface), (which should not be confused with the unrelated Java Metadata Interface that is also called JMI).

As alternatives to the MuPAD based Symbolic Math Toolbox available from MathWorks, MATLAB can be connected to Maple or Mathematica.

Libraries also exist to import and export MathML.

5.4.2 ARMADA

ARMADA is a Data Mining tool designed by James Malone that extracts Association Rules from numerical data files using a variety of selectable techniques and criteria. The program integrates several mining methods which allow the efficient extraction of rules, while allowing the thoroughness of the mine to be specified at the user discretion.[13]

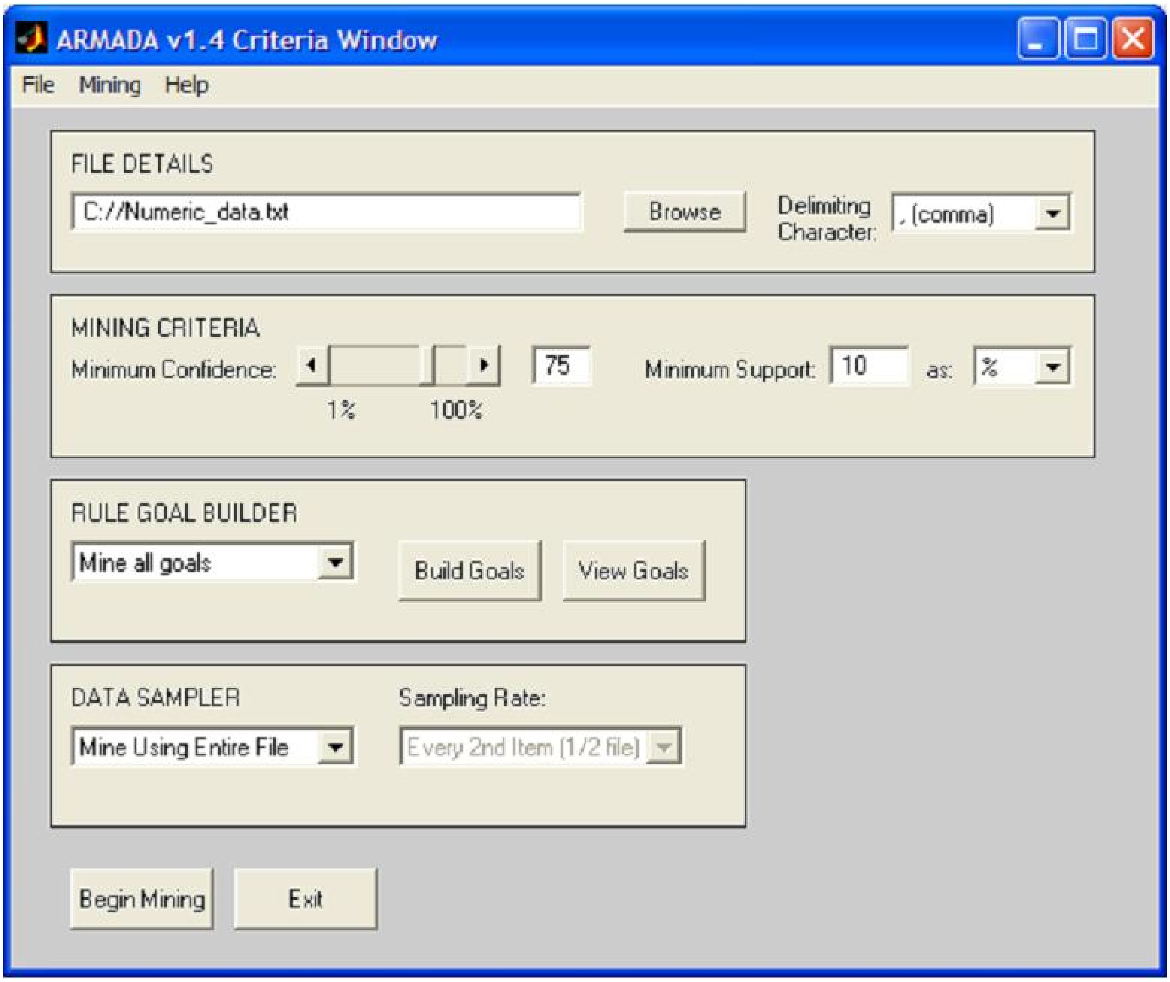
The name ARMADA stands for Association Rule Miner and Deduction Analysis. The program was designed as a tool to assist in the analysis of both the knowledge extracted and the deduction processes by which such a task is undertaken. However, the program can also be used as a straightforward Data Mining tool for the efficient extraction of Association Rules. [6]

The actual knowledge extracted is presented in the form of easy-to-understand rules, while the details of the process are conveniently summarized in the 'Mining Report' section. These mining results can also be saved and opened for analysis. The program also allows the results to be displayed through various graphical representations, such as bar charts and line graphs.[6]

5.4.2.1 The Parts of ARMADA System

Familiarization with the Program:

The ARMADA Criteria Window. This is the initial pre-mining part of the program which deals with specifying the criteria by which the mining process is going to be undertaken.



C://Numeric_data.txt

Delimiting Character: , (comma)

Minimum Confidence: 75

Minimum Support: 10 as: %

Mine all goals

Mine Using Entire File

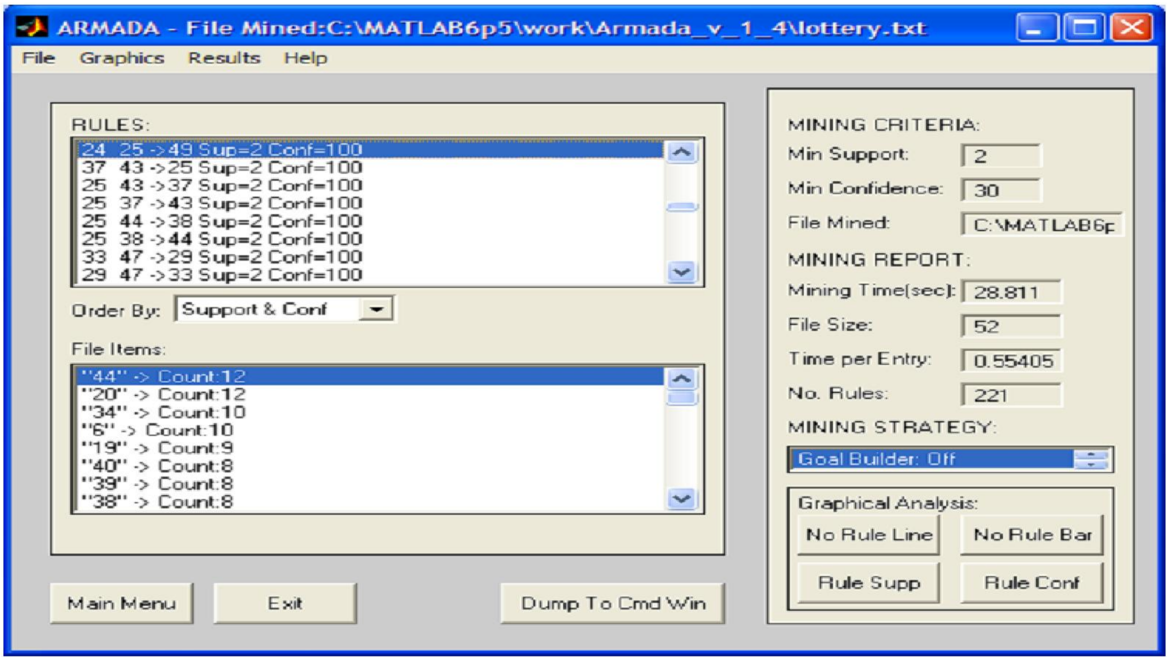
Sampling Rate: Every 2nd Item (1/2 file)

Figure 5-7: ARMADA Criteria Window

This window can be broken down into four further parts

- The File Details section. This deals with the selection of the file related criteria, such as the file and path name and the delimiting character which indicates the character that separates one numeric item from the next within the file.
- The Mining Criteria section. This deals with the specifying of two important attributes used to evaluate Association Rules – that of Minimum Confidence and Minimum Support.
- The Rule Goal Builder section. This allows the creation and viewing of goals by which rules are mined.
- The Data Sampler section. This section specifies the thoroughness by which the mining is undertaken, allowing the data set to be analyzed in full, as a specified sample or as both for analysis purposes. [6]

The Mining Results Window in figure 5.8 shows the post-mining part of the program which displays the Association Rules that have been extracted and a report down the right hand side.



RULES:

```
24  25 ->49 Sup=2 Conf=100
37  43 ->25 Sup=2 Conf=100
25  43 ->37 Sup=2 Conf=100
25  37 ->43 Sup=2 Conf=100
25  44 ->38 Sup=2 Conf=100
25  38 ->44 Sup=2 Conf=100
33  47 ->29 Sup=2 Conf=100
29  47 ->33 Sup=2 Conf=100
```

Order By: Support & Conf

File Items:

```
"44" -> Count:12
"20" -> Count:12
"34" -> Count:10
"6" -> Count:10
"19" -> Count:9
"40" -> Count:8
"39" -> Count:8
"38" -> Count:8
```

MINING CRITERIA:

Min Support: 2

Min Confidence: 30

File Mined: C:\MATLAB6p

MINING REPORT:

Mining Time(sec): 28.811

File Size: 52

Time per Entry: 0.55405

No. Rules: 221

MINING STRATEGY:

Goal Builder: Off

Figure 5-8: The Mining Results Window

5.4.2.2 Data Preparation

This data set includes the basic information fields about patients records gathered, especially prepared for ARMADA tool and Mat Lab implementation.

Basic Information is collected from Sudanese Kidney Transplanted Association, and the follow up data is from Ahmed Qassim Hospital.

It is includes 16 attributes for all information.

1. Age :

Table 5-3: ARMADA Age Values

Range	0-14	15-21	22-40	41-60	61-80	80-130
Code	21	22	23	24	25	26
Name	Child	Boy	Youth	Adult	Old	Immortal

2. Sex :

Table 5-4: ARMADA Sex Code

Sex	Male	Female
Code	11	12

3. Current Statues :

Table 5-5: Current Status

Status	Death	Alive
Code	91	92

4. Blood group

Table 5-6: Blood Group Code

Blood Groups	A+	B+	AB+	O+	A-	B-	AB-	O-
Code	31	32	33	34	35	36	37	38

5. Place of Live/State

Table 5-7: States Code

State	Khartoum	River Nile	Jazeera	White Nile	North Darfur	West Darfur	Central Darfur	West Kordofan

Code	41	42	43	44	45	46	47	48
State	North Kordofan	South Kordofan	Northern	Gadaref	Kassala	Sennar	Read Sea	
Code	49	410	411	412	413	414	415	

6. The Relevant of patient that have infections

Table 5-8: Relevant Code

Element	Yes	No
Code	1	2

7. The cause of Renal Failure

Table 5-9: Renal Failure Cause Code

Cause	Unknown	Atrophy	Blood Pressure	defects	Diabetes	Kidney Infections	Malaria	other
Code	61	62	63	64	65	66	67	68

8. The Other Disease with the patients

Table 5-10: Other Disease Code

Disease	Diabetes	Blood Pressure	asthma
Code	81	82	83

9. Blood pressure

Blood pressure categorize:

Table 5-11: ARMADA Blood Pressure Code

Blood Pressure Category	Systolic mm Hg (upper #)		Diastolic mm Hg (lower #)	Code
Low	Less than 90	And	Less than 60	181
Normal	less than 120	And	less than 80	182
Prehypertension	120 – 139	Or	80 – 89	183
High Blood Pressure (Hypertension) Stage 1	140 – 159	Or	90 – 99	184
High Blood Pressure (Hypertension) Stage 2	160 or higher	Or	100 or higher	185
Hypertensive Crisis (Emergency care needed)	Higher than 180	Or	Higher than 110	186

10. Hemoglobin

Table 5-12: Hemoglobin Code

Sex	Category	Range	Code
Man	High	More than 18	121
	Normal	14-18	122
	Low	Less than 14	123
Women	High	More than 16	124
	Normal	12-16	125
	Low	Less than 12	126

11. White Blood Cells

Table 5-13: White Blood Cells Code

Age	Category	Range	Code
New born	High	More than 15	111
	Normal	15	112
	Low	Less than 15	113
Child	High	More than 12	114
	Normal	12	115
	Low	Less than 12	116
Adult	High	More than 10	117
	Normal	4-10	118
	Low	Less than 4	119

12. Platelets

Table 5-14: Platelets Code

Range	Less than 150	150-400	More than 500
Code	131	132	133

13. Creatinine

Table 5-15: Creatinine Code

Sex	Category	Range	Code
Man	High	More than 1.5	141
	Normal	0.5-1.5	142
	Low	Less than 0.5	143
Women	High	More than 1.2	144
	Normal	0.6-1.2	145
	Low	Less than 0.6	146

14. Uric acid

Table 5-16: Uric Acid Code

Range	Less than 2	2-7	More than 7
Code	151	152	153

15. Urea

Table 5-17: Urea Code

Range	Less than 150	150-400	More than 500
Code	131	132	133

16. Prograf

Table 5-18: Prograf Code

Range	Less than 5	5-9	More than 9
Code	171	172	173

5.4.3 Clementine

Clementine is a data mining workbench that enables user to quickly develop predictive models using business expertise and deploy them into business operations to improve decision making. Designed around the industry-standard CRISP-DM model, Clementine supports the entire data mining process, from data to better business results. [11]

The Clementine Client/Server: This release distributes client requests for resource-intensive operations to powerful server software, resulting in faster performance on larger data sets. Additional products or updates beyond those listed here may also be available.

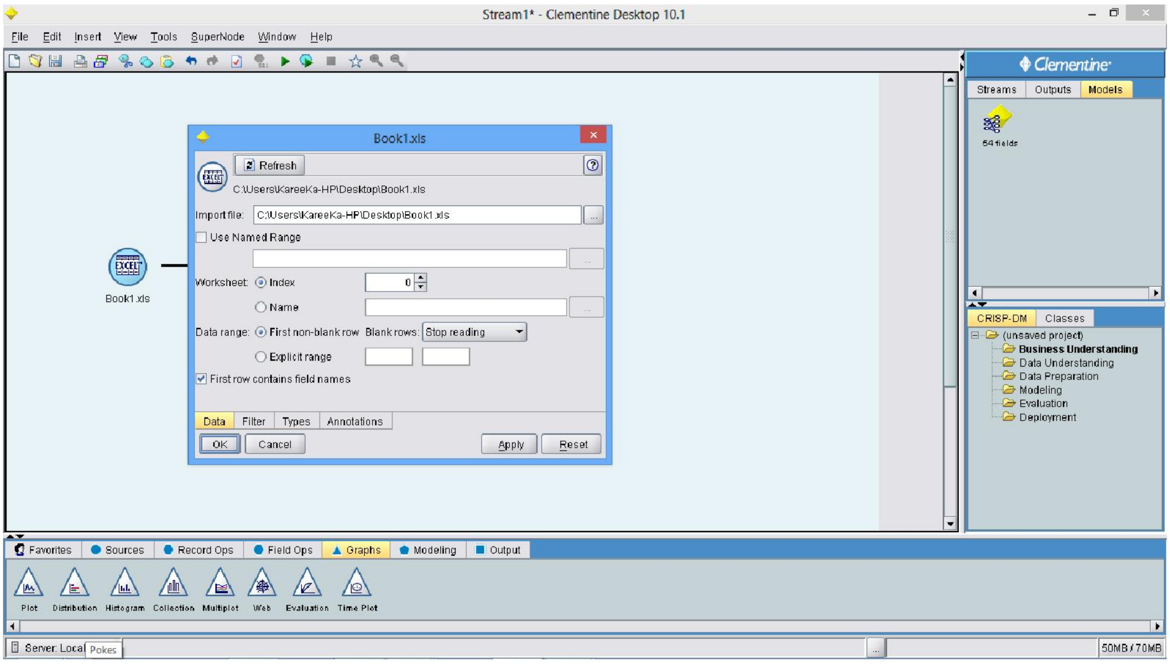
- Clementine Client: Clementine Client is a functionally complete version of the product that is installed and run on the user’s desktop computer. It can be run in local mode as a stand-alone product or in distributed mode along with Clementine Server for improved performance on large data sets.
- Clementine Server: The Clementine Server runs continually in distributed analysis mode together with one or more Clementine Client installations. This provides superior performance on large data sets, because memory-intensive operations can be done on the server without downloading data to the client computer.

Clementine provides different approaches to solve the user's data mining problem such as:

- Feature selection: This approach uses fewer predictors and therefore is less expensive (fast model-building).
- Screening approaches: Anomaly Detection and modeling approach based on a Neural Net. These two approach can be used together.
- Discover affinities or links in a database: This is done by modeling visualization (using a web display).

We used this tool to test the data we gathered and to compare the results with the program that we implemented to see if there is a lot of differences or not. [11]

5.4.3.1 Upload data to Clementine


Stream1* - Clementine Desktop 10.1
File Edit Insert View Tools SuperNode Window Help
Book1.xls
Clementine
Streams Outputs Models
54 fields
CRISP-DM Classes
(unsaved project)
Business Understanding
Data Understanding
Data Preparation
Modeling
Evaluation
Deployment
Favorites Sources Record Ops Field Ops Graphs Modeling Output
Plot Distribution Histogram Collection Multiplot Web Evaluation Time Plot
Server: Local
50MB / 70MB

Book1.xls

Refresh

C:\Users\KareeKa-HP\Desktop\Book1.xls

Import file: C:\Users\KareeKa-HP\Desktop\Book1.xls

□ Use Named Range

Worksheet: ● Index 0

○ Name

Data range: ● First non-blank row Blank rows: Stop reading

○ Explicit range

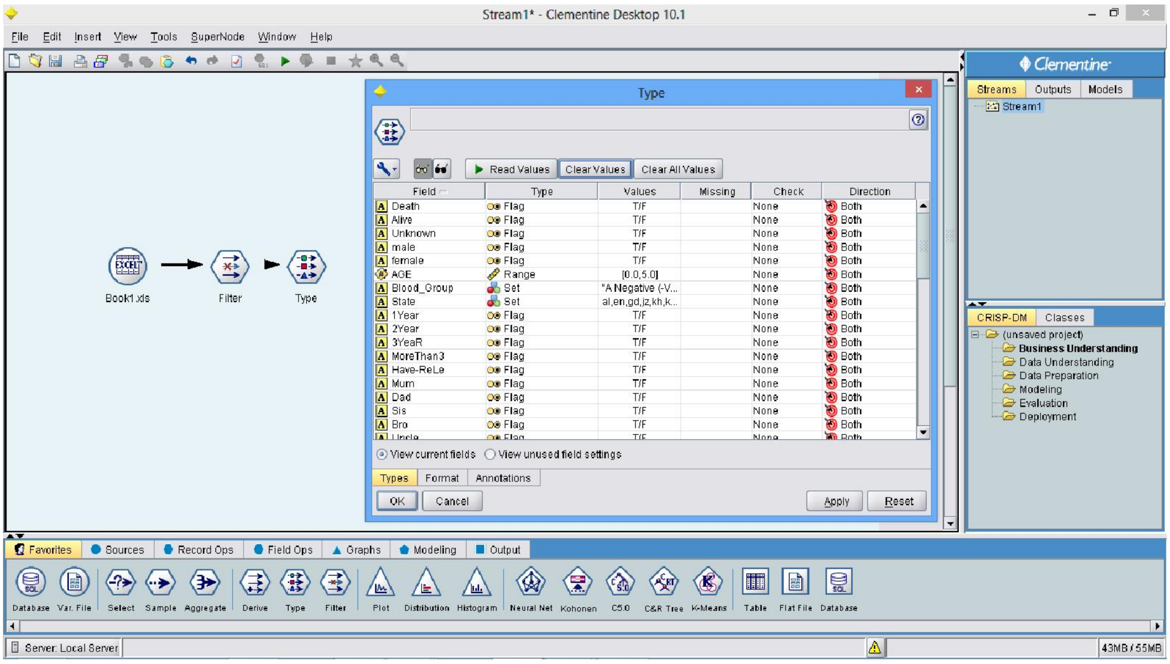
☑ First row contains field names

Data | Filter | Types | Annotations

OK Cancel Apply Reset

Figure 5-9: Upload Data to Clementine

5.4.3.2 Data Filtering


Stream1* - Clementine Desktop 10.1
File Edit Insert View Tools SuperNode Window Help
Book1.xls
Filter
Type
Clementine
Streams Outputs Models
Stream1
CRISP-DM Classes
(unsaved project)
Business Understanding
Data Understanding
Data Preparation
Modeling
Evaluation
Deployment
Favorites Sources Record Ops Field Ops Graphs Modeling Output
Database Var. File Select Sample Aggregate Derive Type Filter Plot Distribution Histogram Neural Net Kohonen C5.0 C&R Tree K-Means Table Flat File Database
Server: Local Server
43MB / 55MB

Type

Read Values | Clear Values | Clear All Values

Field	Type	Values	Missing	Check	Direction
Death	Flag	T/F		None	Both
Alive	Flag	T/F		None	Both
Unknown	Flag	T/F		None	Both
male	Flag	T/F		None	Both
female	Flag	T/F		None	Both
AGE	Range	[0.0,5.0]		None	Both
Blood_Group	Set	"A Negative (-V...		None	Both
State	Set	al,en,gd,jz,kh,k...		None	Both
1Year	Flag	T/F		None	Both
2Year	Flag	T/F		None	Both
3YeaR	Flag	T/F		None	Both
MoreThan3	Flag	T/F		None	Both
Have-ReLe	Flag	T/F		None	Both
Mum	Flag	T/F		None	Both
Dad	Flag	T/F		None	Both
Sis	Flag	T/F		None	Both
Bro	Flag	T/F		None	Both
Uncle	Flag	T/F		None	Both

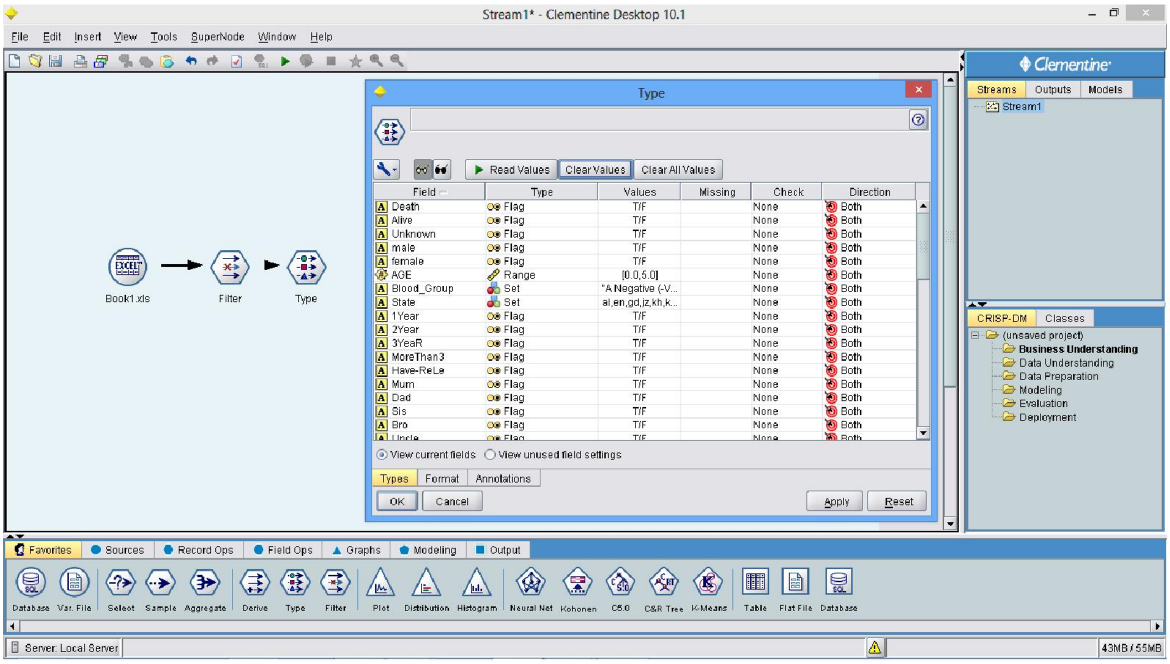
● View current fields ○ View unused field settings

Types | Format | Annotations

OK Cancel Apply Reset

Figure 5-10: Filter Data

5.4.3.3 Choose Types of data


Stream1* - Clementine Desktop 10.1
File Edit Insert View Tools SuperNode Window Help
Type
Read Values
Clear Values
Clear All Values
Book1.xls
Filter
Type
Clementine
Streams Outputs Models
Stream1
CRISP-DM Classes
(unsaved project)
Business Understanding
Data Understanding
Data Preparation
Modeling
Evaluation
Deployment
Favorites Sources Record Ops Field Ops Graphs Modeling Output
Database Var. File Select Sample Aggregate Derive Type Filter Plot Distribution Histogram Neural Net Kohonen C5.0 C&R Tree K-Means Table Flat File Database
Server: Local Server
43MB / 55MB

Field	Type	Values	Missing	Check	Direction
Death	Flag	T/F		None	Both
Alive	Flag	T/F		None	Both
Unknown	Flag	T/F		None	Both
male	Flag	T/F		None	Both
female	Flag	T/F		None	Both
AGE	Range	[0.0,5.0]		None	Both
Blood_Group	Set	"A Negative (-V...		None	Both
State	Set	al,en,gd,jz,kh,k...		None	Both
1Year	Flag	T/F		None	Both
2Year	Flag	T/F		None	Both
3YeaR	Flag	T/F		None	Both
MoreThan3	Flag	T/F		None	Both
Have-ReLe	Flag	T/F		None	Both
Mum	Flag	T/F		None	Both
Dad	Flag	T/F		None	Both
Sis	Flag	T/F		None	Both
Bro	Flag	T/F		None	Both
Uncle	Flag	T/F		None	Both

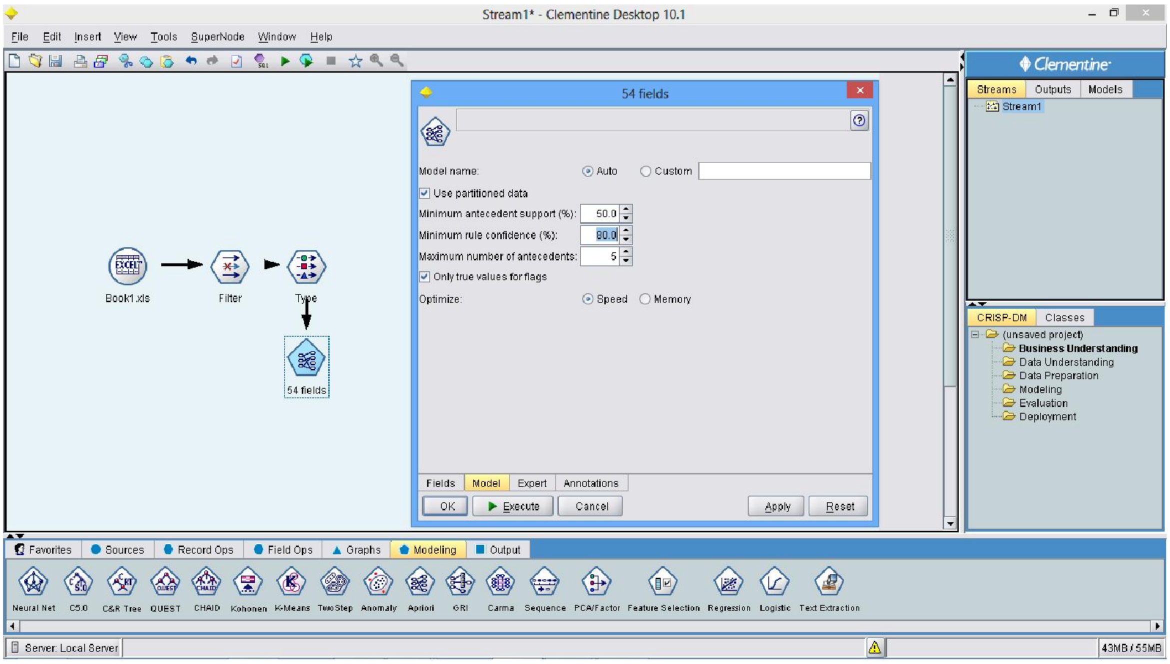
● View current fields ○ View unused field settings

Types | Format | Annotations

OK | Cancel | Apply | Reset

Figure 5-11: Read Data Types

5.4.3.4 Algorithm option


Stream1* - Clementine Desktop 10.1
File Edit Insert View Tools SuperNode Window Help
54 fields
Book1.xls
Filter
Type
54 fields
Clementine
Streams Outputs Models
Stream1
CRISP-DM Classes
(unsaved project)
Business Understanding
Data Understanding
Data Preparation
Modeling
Evaluation
Deployment
Favorites Sources Record Ops Field Ops Graphs Modeling Output
Neural Net C5.0 C&R Tree QUEST CHAID Kohonen K-Means TwoStep Anomaly Apriori GRI Carma Sequence PCA/Factor Feature Selection Regression Logistic Text Extraction
Server: Local Server
43MB / 55MB

Model name: ● Auto ○ Custom

☑ Use partitioned data

Minimum antecedent support (%): 50.0

Minimum rule confidence (%): 80.0

Maximum number of antecedents: 5

☑ Only true values for flags

Optimize: ● Speed ○ Memory

Fields | Model | Expert | Annotations

OK | Execute | Cancel | Apply | Reset

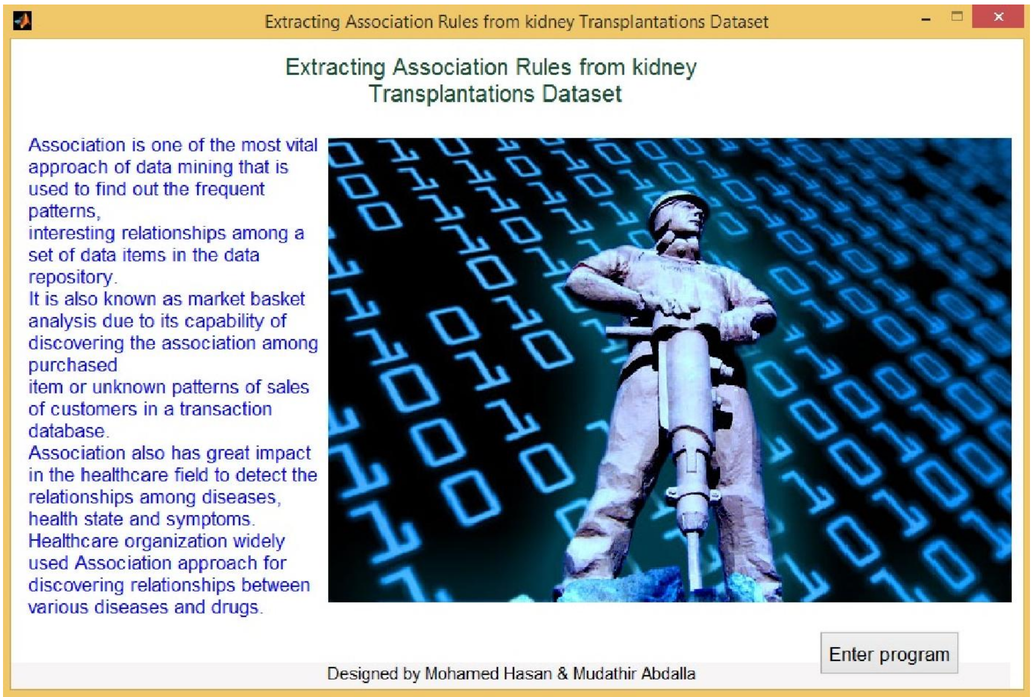
Figure 5-12: Algorithm Options

5.4.4 Extracting Association using MatLab

We developed a Matlab implementation where doctors can extract association rules from kidney transplantation dataset through its graphical user interface, to help in detecting relations between different tests and patient's information, and acquire a lot of knowledge from it. Christian Borgelt developed version of Apriori algorithm have been adopted. The simple GUI minimizes the options to be easily understood. It consists of three windows: Introduction window, Algorithm parameters and Option window and Result window.

5.4.4.1 Introduction Window

This Window is Identification of program, it is divide in three parts. Part one in lift side of window contain information of association rule data mining technique that provides briefly explain, in right side of window contain relative slide show image to provide idea of program to any person, part three it is button in right corner of bottom window that will move users to next window as shown in figure 5-14.



Extracting Association Rules from kidney Transplantations Dataset

Extracting Association Rules from kidney
Transplantations Dataset

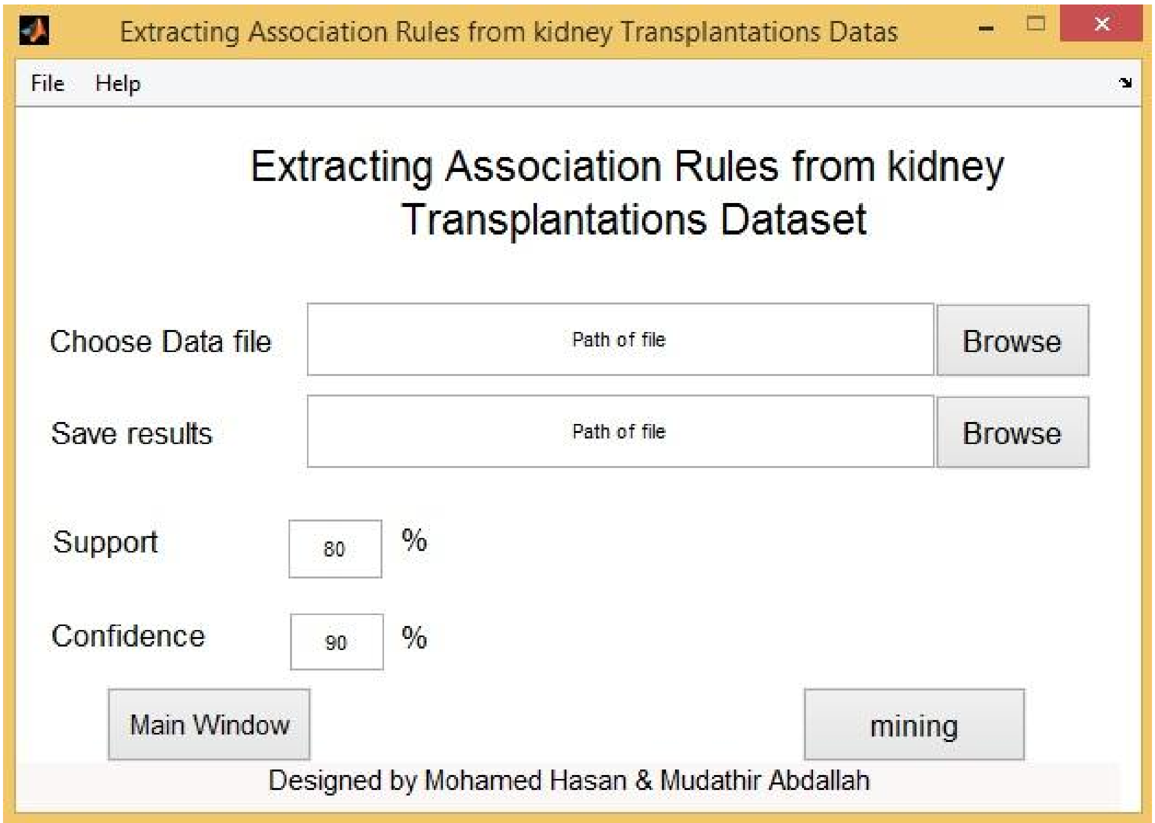
Association is one of the most vital approach of data mining that is used to find out the frequent patterns,
interesting relationships among a set of data items in the data repository.
It is also known as market basket analysis due to its capability of discovering the association among purchased
item or unknown patterns of sales of customers in a transaction database.
Association also has great impact in the healthcare field to detect the relationships among diseases, health state and symptoms.
Healthcare organization widely used Association approach for discovering relationships between various diseases and drugs.

Designed by Mohamed Hasan & Mudathir Abdalla

Figure 5-13: Introduction Window

5.4.4.2 Options Window

Options window provide users control of algorithm, it is consist basic option of algorithm that specific files and set support and confidence. After select data file and set support and confidence mining button will execute algorithm and move users to results window or press main window button for back to introduction window as shown in figure 5-15.



Extracting Association Rules from kidney Transplantations Datas

Extracting Association Rules from kidney Transplantations Dataset

Choose Data file: Path of file

Save results: Path of file

Support: 80 %

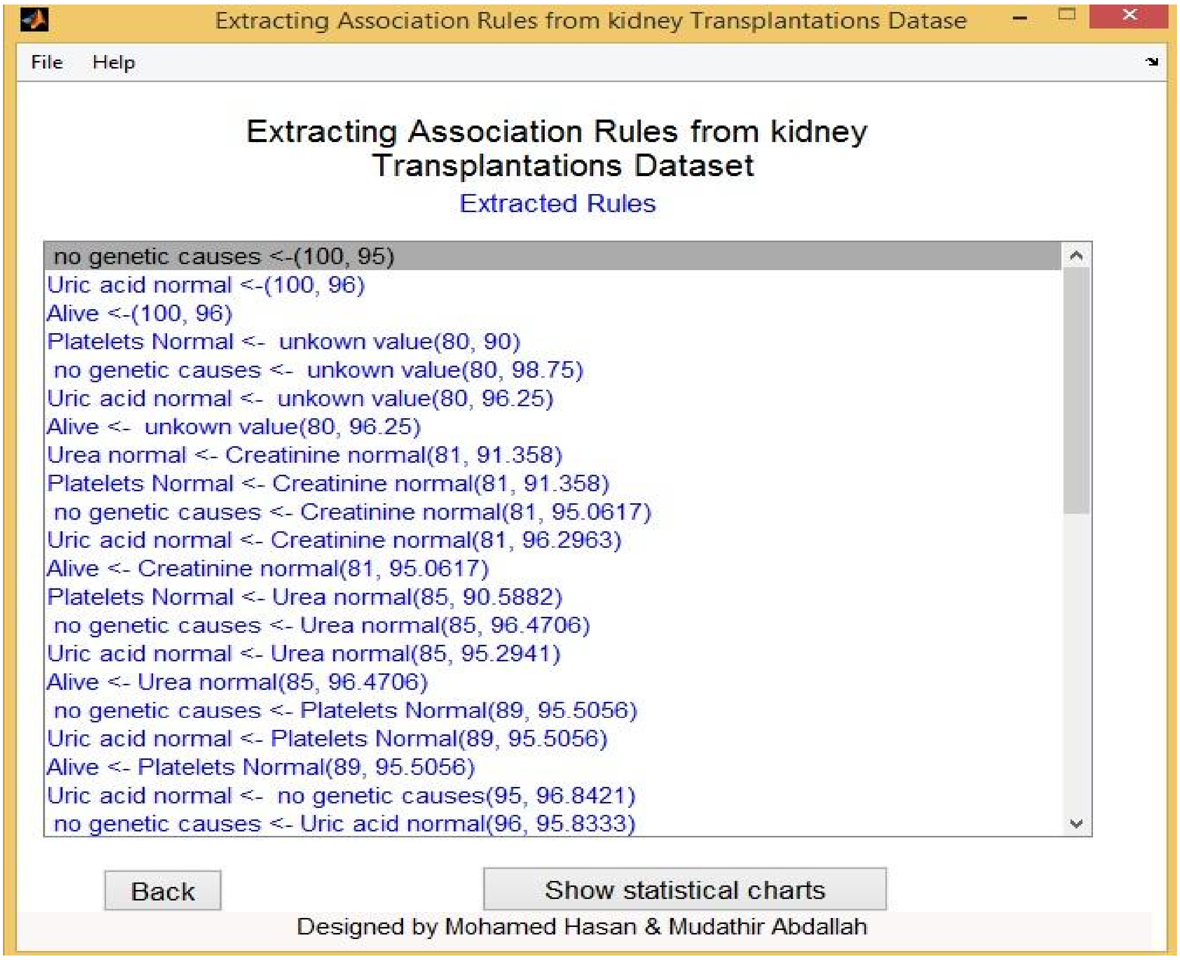
Confidence: 90 %

Designed by Mohamed Hasan & Mudathir Abdallah

Figure 5-14: Options Window

5.4.4.3 Results Window

This window shown extracted rules. The rules shown in a list in middle of window. The show statistical charts button is show many charts for kidney transplants of Sudan and back button will returned user to previous window as shown in figure 5-16.


Extracting Association Rules from kidney Transplantations Datase
File
Help
Back
Show statistical charts

Extracting Association Rules from kidney Transplantations Dataset

Extracted Rules

```
 no genetic causes <-(100, 95)
Uric acid normal <-(100, 96)
Alive <-(100, 96)
Platelets Normal <-  unkown value(80, 90)
 no genetic causes <-  unkown value(80, 98.75)
Uric acid normal <-  unkown value(80, 96.25)
Alive <-  unkown value(80, 96.25)
Urea normal <- Creatinine normal(81, 91.358)
Platelets Normal <- Creatinine normal(81, 91.358)
 no genetic causes <- Creatinine normal(81, 95.0617)
Uric acid normal <- Creatinine normal(81, 96.2963)
Alive <- Creatinine normal(81, 95.0617)
Platelets Normal <- Urea normal(85, 90.5882)
 no genetic causes <- Urea normal(85, 96.4706)
Uric acid normal <- Urea normal(85, 95.2941)
Alive <- Urea normal(85, 96.4706)
 no genetic causes <- Platelets Normal(89, 95.5056)
Uric acid normal <- Platelets Normal(89, 95.5056)
Alive <- Platelets Normal(89, 95.5056)
Uric acid normal <-  no genetic causes(95, 96.8421)
 no genetic causes <- Uric acid normal(96, 95.8333)
```

Designed by Mohamed Hasan & Mudathir Abdallah

Figure 5-15: Results Window

CHAPTER SIX

RESULT & DISCUSSION

6 Result and Discussion

This part discusses the extracted rules that represent the data mining output and the new knowledge that was derived. Each application has a unique result that is shows up below:

6.1 Clementine result and rules

The rules of clementine has always one consequent and one or more antecedent as shown in the table bellow

Table 6-1: Clementine Results

Consequent	Antecedent	Support %	Confidence %
1. Urea-Normal = T	Creatinine-Normal = T	82.0	91.463
2. White Blood Cells-Normal = T	Creatinine-Normal = T	82.0	90.244
3. Alive = T	Creatinine-Normal = T	82.0	93.902
4. Uric Acid-Normal = T	Creatinine-Normal = T	82.0	96.341
5. White Blood Cells-Normal = T	Urea-Normal = T	85.0	91.765
6. Alive = T	Urea-Normal = T	85.0	95.294
7. Uric Acid-Normal = T	Urea-Normal = T	85.0	95.294
8. Alive = T	White Blood Cells-Normal = T	90.0	94.444
9. Uric Acid-Normal = T	White Blood Cells-Normal = T	90.0	95.556
10. Uric Acid-Normal = T	Alive = T	95.0	96.842
11. Alive = T	Uric Acid-Normal = T	96.0	95.833
12. White Blood Cells-Normal = T	Urea-Normal = T and Alive = T	81.0	91.358

13. White Blood Cells-Normal = T	Urea-Normal = T and Uric #Acid-Normal = T	81.0	91.358
14. Uric Acid-Normal = T	Urea-Normal = T and Alive = T	81.0	96.296
15. Alive = T	Urea-Normal = T and Uric Acid-Normal = T	81.0	96.296
16. Uric Acid-Normal = T	White Blood Cells-Normal = T and Alive = T	85.0	96.471
17. Alive = T	White Blood Cells-Normal = T and Uric Acid-Normal = T	86.0	95.349

6.2 ARMADA results

The rules of this tools are encrypted, they are shows as:

Table 6-2: ARMADA Results

Consequent	Antecedent	Support %	Confidence %
152	92	93	96.875
92	152	93	96.875
1	152	92	95.8333
1	92	92	95.8333
152	132	85	95.5056
92	132	85	95.5056
92	162	82	96.4706
1	162	82	96.4706
152	162	81	95.2941
152	1, 92	90	97.8261
92	1, 152	90	97.8261
1	92, 152	90	96.7742
152	92, 132	82	96.4706
92	132, 152	82	96.4706
152	1, 132	82	96.4706
1	132, 152	82	96.4706
92	1, 132	82	96.4706
1	92, 132	82	96.4706

6.2.1 Graphical Analysis

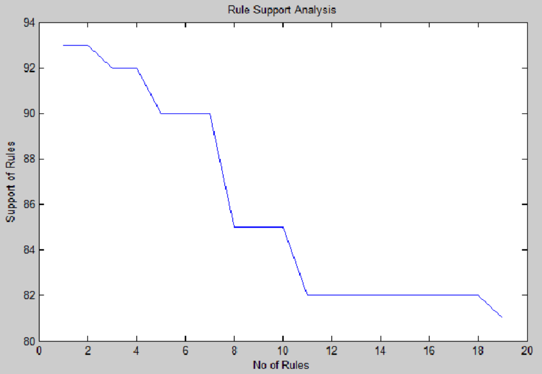


Figure 6-1: Rule Support Analysis

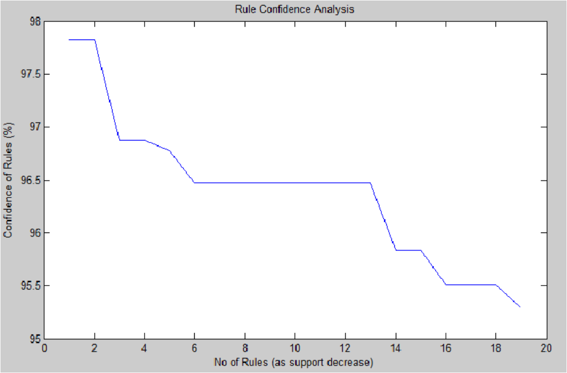


Figure 6-2: Rule Confidence Analysis

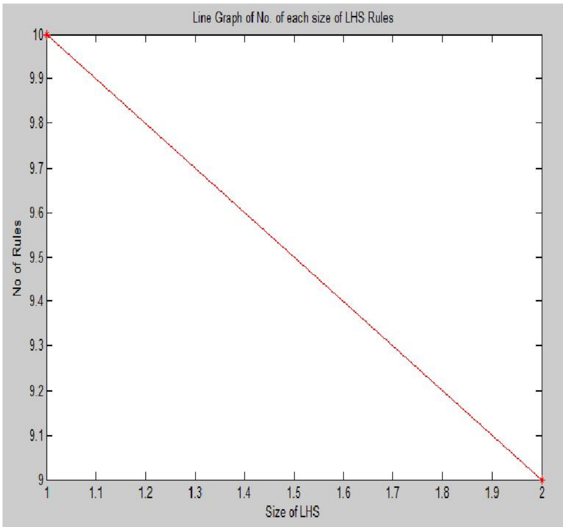


Figure 6-3: No of Antecedent Size

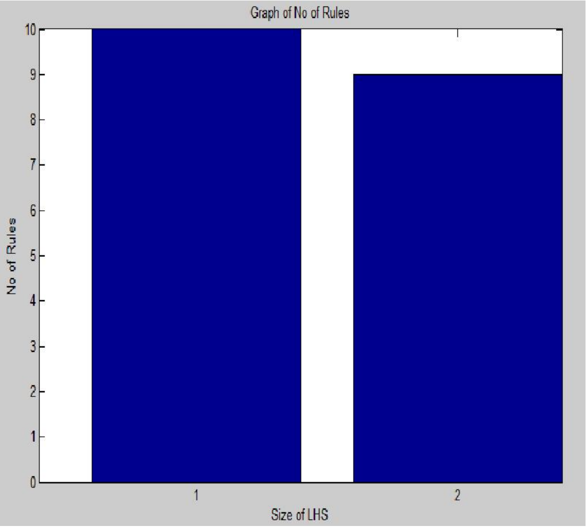


Figure 6-4: Graph of NO of Rules

6.3 Extracting Association Rules Results

This results when support 80% and confidence 90%, it was extracted 46 rules but we were filtered it to 19 rules.

Table 6-3: Extracting Association Rules Results

Consequance	Antecedent			Support %	Confidence %
1. no genetic causes	Creatinine normal			81	95.0617

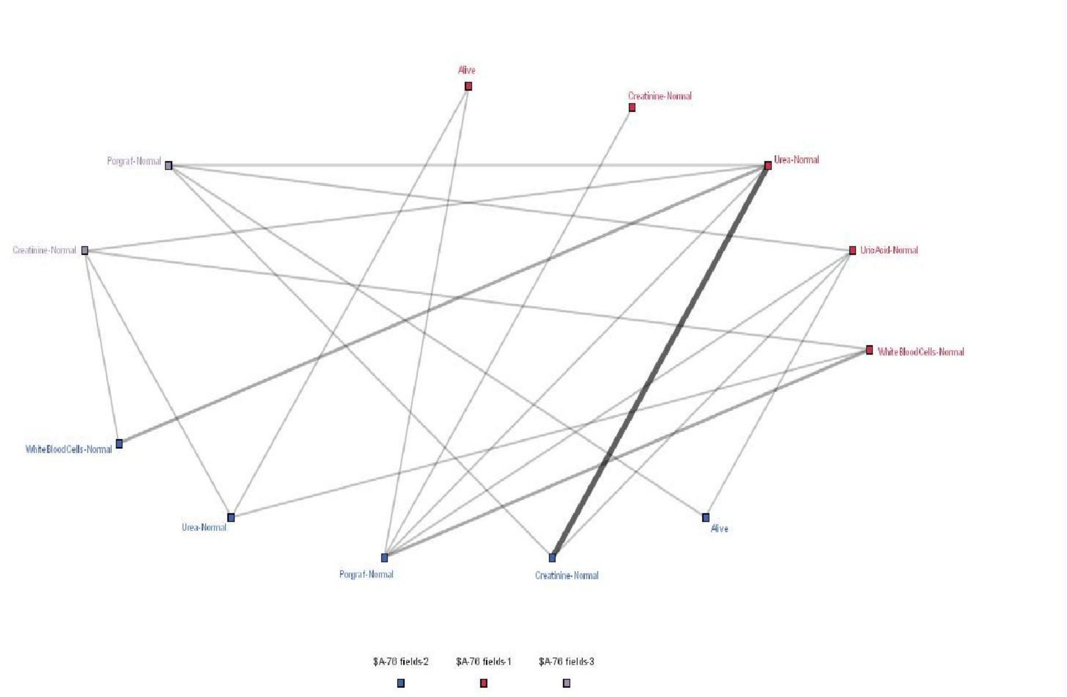
2. Uric acid normal	Creatinine normal			81	96.2963
3. no genetic causes	Urea normal	Uric acid normal		81	97.5309
4. Alive	Urea normal	no genetic causes		82	96.3415
5. no genetic causes	Urea normal	Alive		82	96.3415
6. Alive	Urea normal	Uric acid normal		81	97.5309
7. Uric acid normal	Urea normal	Alive		82	96.3415
8. Uric acid normal	Platelets Normal	no genetic causes		85	96.4706
9. no genetic causes	Platelets Normal	Uric acid normal		85	96.4706
10. Alive	Platelets Normal	no genetic causes		85	96.4706
11. no genetic causes	Platelets Normal	Alive		85	96.4706
12. Alive	Platelets Normal	Uric acid normal		85	96.4706
13. Uric acid normal	Platelets Normal	Alive		85	96.4706
14. Alive	no genetic causes	Uric acid normal		92	97.8261
15. Uric acid normal	no genetic causes	Alive		92	97.8261
16. no genetic causes	Uric acid normal	Alive		93	96.7742
17. Alive	Platelets Normal	no genetic causes	Uric acid normal	82	97.561
18. Uric acid normal	Platelets Normal	no genetic causes	Alive	82	97.561
19. no genetic causes	Platelets Normal	Uric acid normal	Alive	82	97.561

6.4 Discussion

The results of this tools is likely, the rules of clementine number 1, 4, 7 are good rules because their items are related to renal functions, rule number 8 is good, but in case of abnormality the patient is alive but threaten if the Platelets are low he has immune deficiency, if they are high it means that he has an infection, rule number 8 is true; the patient is alive but having complications such as renal stones.

The rest of rules are refused because the factors are not related and there is no biological explanation or scientific proof and the factors that affect white blood cells and the factors that affect uric acid level are not essentially related to each other.

The figure bellow illustrate the strong relations between items of the result rules in a heavy lines the degree of heaviness is increase for the items more various occur between them.



The rules of extracting association rules in general is good, the rule number 2 mean if level of creatinine normal that implis to level of uric acid normal, this rule is same rule number 4 of clementine.

Figure 6-5: Relation of Fields

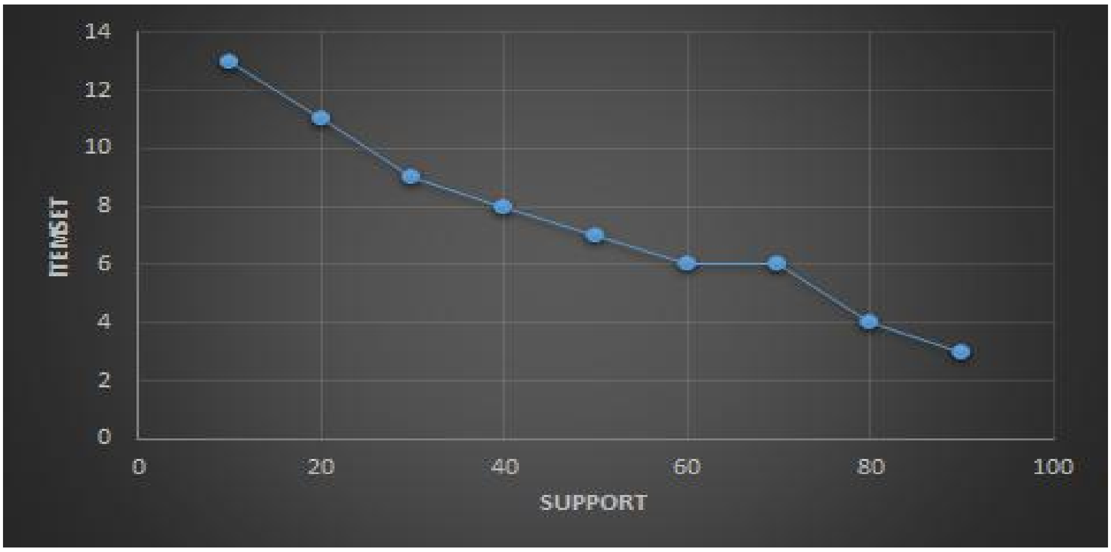


Figure 6-6: Relation of Item Set and Support

6.5 Statistical Charts

This results was derived from basics patient's data. It is statistical information about patients in Sudan. Consist of three charts: Percentage of patients in states, Percentage of patient's age by sex and Percentage of patients' blood group.

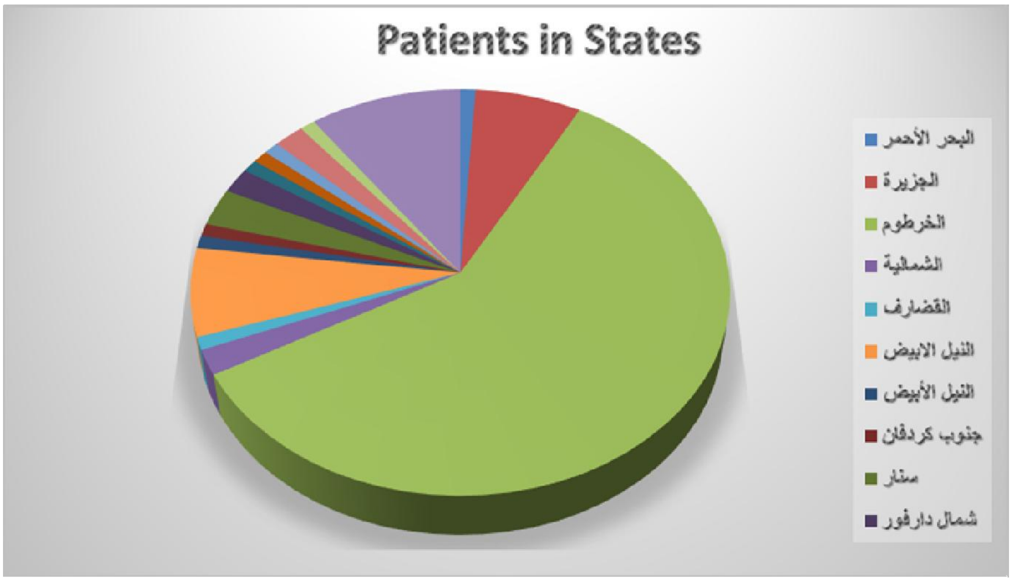


Figure 6-7: Patients in States

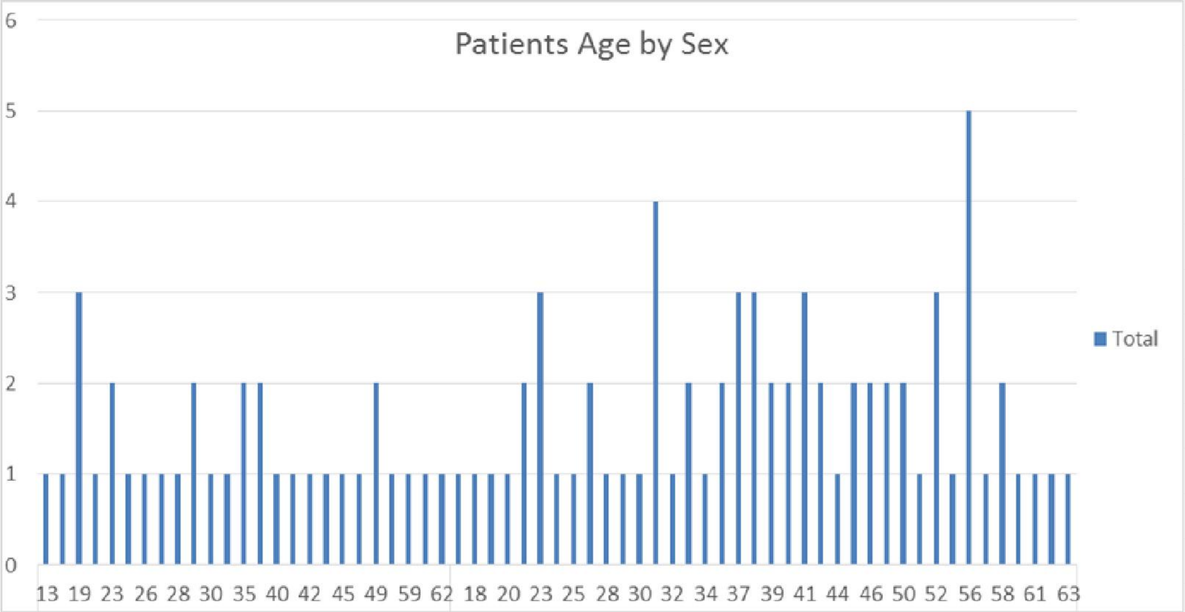


Figure 6-8: Patients Age by Sex

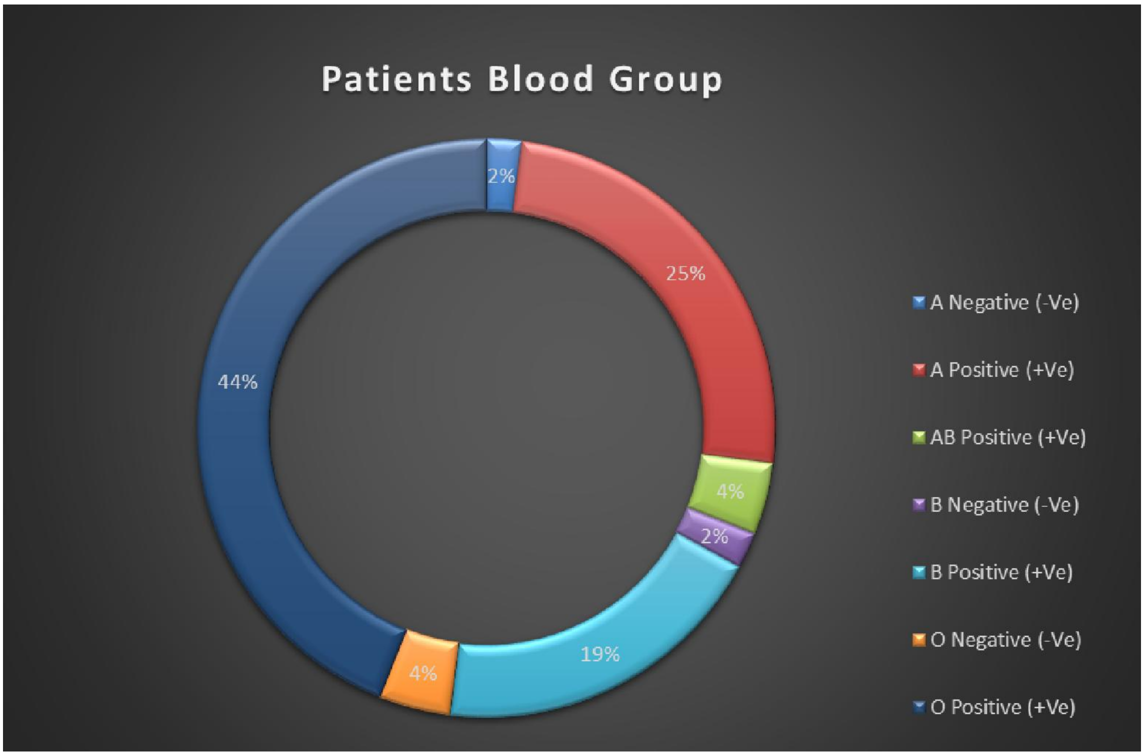


Figure 6-9: Patients Blood Group

CHAPTER SEVEN

CONCLUSION & RECOMMENDATION

7 Conclusion & Recommendation

This part illustrates the final aspects that were found from studying and mining the data set collected.

7.1 Conclusion

This study showed the success of the operations of a new kidney transplantation is highly affected by the Urea, Uric Acid and serum Creatinine reading and results of tests which have been included in this program. Using the algorithm led to extracting relationships between different values of data set and come out with multifactorial result.

Finally Association rule mining was able to extract information from kidney transplantation data set, the physicians were agreed the knowledge that founded which help them to understands the relation between items and their impact on patients

7.2 Recommendation

- Data set building by gathering the each available information in all Cardiac Surgery & Renal Transplantation centers to make any member of Sudanese kidney association has complete records contain his basic information and all follow-up data after the period of transplantation.
- More studies should be done using the data that Sudanese kidney association has store it and evaluate the results with other hospitals or centers of kidney failure disease.
- The Cardiac Surgery & Renal Transplantation Center in Ahmed Qassim Hospital should develop a computer system that should be capable of archiving the information of every patient’s visits and the tests which are done to facilitate the future research.

7.3

8 Sources and references

[1] H. C. Koh and G. Tan, “Data mining applications in healthcare.,” *J. Healthc. Inf. Manag.*, vol. 19, no. 2, pp. 64–72, Jan. 2005.

[2] D. Tomar and S. Agarwal, “A survey on Data Mining approaches for Healthcare,” *Int. J. Bio-Science Bio-Technology*, vol. 5, no. 5, pp. 241–266, 2013.

[3] J. Han and M. Kamber, “Data Mining : Concepts and Techniques (2nd edition) Bibliographic Notes for Chapter 6 Classification and Prediction,” 2006.

[4] “Original article : Incidence of renal failure in postoperative period of cardiovascular surgeries in India Abstract :,” no. 1, pp. 113–117, 2013.

[5] M. Rezapour, M. Khavanin Zadeh, and M. M. Sepehri, “Implementation of predictive data mining techniques for identifying risk factors of early AVF failure in hemodialysis patients.,” *Comput. Math. Methods Med.*, vol. 2013, p. 830745, Jan. 2013.

[6] B. J. Malone, “Rule And Deduction Analysis User Manual.”

[7] http://www.wifinotes.com/computer-networks/what-is-data-mining.html 17/03/2014 7:40pm

[8] http://www.contrib.andrew.cmu.edu/~tjabban/datamining.html

31/03/2014 22:21pm

[9] http://www.dataminingarticles.com/association-analysis/association-rules-algorithms/

30/07/2014 21:20 PM

[10] http://www.codeproject.com/Articles/70371/Apriori-Algorithm

02/04/2014 12:55am

[11] http://web.soccerlab.polymtl.ca/soccerlab/English/Soccer-lab-tool-19_E.html

27/07/2014 01:44AM

[12] http://www.borgelt.net/doc/apriori/apriori.html

28/07/2014 15:31PM

[13] http://www.mathworks.com/matlabcentral/fileexchange/3016-armada-data-mining-tool-version-1-4

29/7/2014 12:17

[14] Mohammed Abdallah, Mram Awad, Hibat Allah Al Fadel, “Using Data Mining Techniques for Heart Diseases”, 2010