إقـــــرار

أنا الموقع أدناه مقدم الرسالة التي تحمل العنوان:

Healthcare Resource Management by using Data mining – Predicting Length of Stay

أقر بأن ما اشتملت عليه هذه الرسالة إنما هو نتاج جهدي الخاص، باستثناء ما تمت الإشارة إليه حيثما ورد، وإن هذه الرسالة ككل أو أي جزء منها لم يقدم من قبل لنيل درجة أو لقب علمي أو بحثي لدى أي مؤسسة تعليمية أو بحثية أخرى.

DECLARATION

The work provided in this thesis, unless otherwise referenced, is the researcher's own work, and has not been submitted elsewhere for any other degree or qualification

Student's name:	اسم الطالب/ة: ناريمان سامي التتر
Signature:	التوقيع: ناريمان
Date:	التاريخ: 2 / 02 / 2016

بسم الله الرحمن الرحيم

Islamic University Gaza
Deanery of Post Graduate Studies
Faculty of Commerce
Business Administration Department

الجامعة الإسلامية – غزة
عمادة الدراسات العليا
كلية التجارة
قسم إدارة أعمال

Healthcare Resource Management by using Data mining – Predicting Length of Stay

Case Study: (Birthing Centers and Maternity Hospitals- Gaza Strip)

By

Nariman Sami El-Tater

Supervised By:

Dr. Wael Al-Daya

A Thesis Submitted as Partial Fulfillment of the Requirements for the Degree of Master in Business Administration

1437 H –2015 AD

بسم الله الرحمن الرحيم

الجامعة الإسلامية – غزة
The Islamic University - Gaza

مكتب نائب الرئيس للبحث العلمي والدراسات العليا هاتف داخلي 1150

الرقم.....ج س. غ/35/..... Ref

التاريخ.....2015/12/23م..... Date

نتيجة الحكم على أطروحة ماجستير

بناءً على موافقة شئون البحث العلمي والدراسات العليا بالجامعة الإسلامية بغزة على تشكيل لجنة الحكم على أطروحة الباحثة/ **ناريمان سامي صقر التتر** لنيل درجة الماجستير في **كلية *التجارة*/ قسم إدارة الأعمال** وموضوعها:

تطبيق تنقيب البيانات في إدارة الموارد الصحية – التنبؤ بمدة الإقامة

دراسة حالة: مراكز ومستشفيات التوليد – قطاع غزة

Healthcare Resource Management by using Data mining-predicting length of Stay Case Study: Birthing Centers and Maternity hospitals – Gaza Strip

وبعد المناقشة التي تمت اليوم الثلاثاء 12 ربيع الأول 1437هـ، الموافق 2015/12/23م الساعة الواحدة ظهراً، اجتمعت لجنة الحكم على الأطروحة والمكونة من:

د. وائل حمدي الدايـة	مشرفاً و رئيساً	.....
د. خالد عبد دهليز	مناقشاً داخلياً	.....
د. تامر سعد فطاير	مناقشاً خارجياً	.....

وبعد المداولة أوصت اللجنة بمنح الباحثة درجة الماجستير في **كلية *التجارة*/قسم إدارة الأعمال**.

واللجنة إذ تمنحها هذه الدرجة فإنها توصيها بتقوى الله ولزوم طاعته وأن تسخر علمها في خدمة دينها ووطنها.

والله ولي التوفيق ،،،

مساعد نائب الرئيس للبحث العلمي والدراسات العليا

أ.د. عبد الرؤوف علي المناعمة

"وَمَا أُوتِيتُمْ مِنَ الْعِلْمِ إِلَّا قَلِيلًا"

صدق الله العظيم
(سورة الإسراء : آية 85)

Acknowledgement

First, *I thank* ***Allah*** *for guiding me and taking care of me all the time. Praise is to Allah for giving me the power and help to accomplish this research.*

I would like to thank ***my family*** *especially* ***my parents*** *for encouraging and supporting my all the time.*

Also, I would like to express my sincere gratitude to my advisor ***Dr. Wael Al-*** **Daya** *for the continuous support of my study and related research, for his patience, motivation, and immense knowledge. His guidance helped me in all the time of research and writing of this thesis. I could not have imagined having a better advisor and mentor for my study.*

Sincere gratitude is expressed to the discussion board members, ***Dr. Khalid Dahleez*** *and* ***Dr. Tamer Fatayer*** *for their comments and discussion of this thesis.*

I wish to express my considerable gratitude to my special ***friend and colleges,*** *for there supports.*

Thank you all for being always there when I need you most. Thank you for believing in me and supporting me.

Abstract

Government hospitals in the Gaza Strip, Palestine are forced to use their scarce resources as efficient as possible. One of these resources is the hospital bed capacity. An optimal admission planning uses hospital bed capacity as efficient as possible. In order to achieve such a planning, early predictions of the expected discharge moment of patients are needed. Expected discharge moments can be predicted if the expected duration of admissions is known. In other words, predictions of the expected length of stay (LOS) at admission are required. In this study, LOS is defined as the number of days a patient is admitted to the hospital during an admission.

Predicting the LOS of patients in a hospital is important in providing them with better services and higher satisfaction, furthermore accurate predictions of patient LOS in the hospital can effectively manage hospital resources and increase efficiency of patient care.

The aim of this study is to applying data mining techniques to support successful decisions that will improve success of healthcare management and build accurate model to predict the LOS of maternity hospital. The Data were collected from childbirth database. The patient records of 22,461 instances were included in the analysis.

This thesis applied on three government maternity hospitals in Gaza strip. The dataset used information of pregnant women who delivered in a public hospital in Gaza Strip between 1 January 2013 and 31 December 2013.

The techniques used are classification with decision tree. LOS is the target variable, and 12 input variables are used for prediction. The LOS is categorized into three classes, LOS 1 less than 24 hours, LOS2 from 24 hours to 72 hours and LOS 3 greater

than 72 hours. A confusion matrix was obtained to calculate sensitivity, specificity, and accuracy. Results: The overall accuracy of decision tree was 79.99 % in the training set. Most normal delivery (77.54%) had an LOS ≤24 hours, whereas 20.5% of cesarean section had an LOS >24 hours. Moreover, the study shows that LOS class recall is 97.56%. Therefore the tree algorithms are able to predict LOS with various degrees of accuracy.

Keywords: Management Resources, Data Mining, Length of Stay, Childbirth, Normal Delivery

ملخص الدراسة

عنوان الرسالة :" تطبيق تنقيب البيانات في إدارة الموارد الصحية – التنبؤ بمدة الإقامة

دراسة حالة: مراكز ومستشفات التوليد – قطاع غزة"

إن شح الموارد في المستشفيات الحكومية في قطاع غزة يتطلب إدارة الموارد الصحية بشكل فعال، وإحدى هذه الموارد هي السعة السريرية للمستشفى. ولتحقيق هذا الهدف تبرز الحاجة إلى التوقع المبكر لوقت خروج المريض وتحديد مدة إقامته بشكل دقيق.

وبمعنى اخر فإن المطلوب هو التنبؤ بمدة إقامة المريض في المستشفى لتحديد الخدمات المطلوبة ويساهم التنبؤ بمدة إقامة المريض بالمستشفيات في تقديم الرعاية الصحية بشكل أفضل وتحقيق رضى المرضى مما يساعد المستشفيات بإدارة مواردها بشكل فعال.

وتهدف هذه الدراسة لإستخدام تقنية تنقيب البيانات "Data mining" للتنبؤ بمدة إقامة المريض في المتشفيات.

تم تطبيق هذه الدراسة على ثلاث مستشفيات من مستشفيات الولادة الحكومية باستخدام البيانات للسيدات اللواتي ولدن في هذه المستشفيات من تاريخ 1-1-2013 إلى تاريخ 31-12-2013 بواقع 22461 سجل دخول لهذه المستشفيات.

ولقد قمنا بتطبيق تقنية التصنيف باستخدام شجرة صنع القرار للتنبؤ بمدة الإقامة، وتم تقسيم مدة الإقانة إلى ثلاث فئات الفئة الأولى أقل من 24 ساعة والفئة الثانية من 24-72 ساعة والفئة الثالثة أكبر من 72 ياعة وأشارت النتائج إلى قدرة النموذج على التنبؤ بدقة 79.99 % وبدرجة حساسية 97.56 % للفئة الأولى.

كما تشير التحليلات أن معظم الولادات الطبيعية كانت ضمن الفئة الأولى بنسبة 77.54 % وأن معظم الولادات القيصرية ضمن الفئة الثانية والثالثة بنسبة 20.5 %.

الكلمات المفتاحية: تنقيب البيانات – إدارة الموارد – مدة الإقامة – الولادة الطبيعية – الولادات.

Table of Contents

List of Figures

List of Tables

List of Abbreviations

CRISP-DM	Cross-Industry Standard Process for Data Mining
DM	Data mining
KDD	Knowledge Discovery from Data
KDP	knowledge discovery process
MoH	Ministry of Health
CART	Classification and regression Trees
PHC	Primary Health Care
SVM	Support vector machine
AUC	Area Under (ROC) curve
MLP	Multi Layer Perceptrones
Weka	Weka Waikato Environment for Knowledge Analysis
CHAID	CHAID: Chai-squared Automation Interaction Detection
ANN	ANN: Artificial Neural Network
DM	DM: Data Mining
LOS	Length of Stay
AI	Artificial Intelligence

Chapter 1

Introduction

Chapter 1 Introduction

1.1 Overview

The advances in health informatics as Electronic Health Record (HER) make healthcare organizations overwhelmed with data. The wealth of data available within the healthcare industry is a key resource to be processed and analyzed for knowledge extraction. The knowledge discovery is the process of making low-level data into high-level knowledge (Khajehei & Etemady, 2010).

Data mining is a core component of the Knowledge Discovery in Databases (KDD) process. Data mining techniques are used in healthcare management which improve the quality and decrease the cost of healthcare services. Data mining algorithms are needed in almost every step in KDD process ranging from domain understanding to knowledge evaluation (Mahindrakar & Hanumanthappa, 2013).

Therefore, data mining is one of the most vital and motivating area of research with the objective of finding meaningful information from huge data sets (Tomar & Agarwal, 2013).

Data Mining is becoming popular in healthcare field because there is a need of efficient analytical methodology for detecting patterns of diagnosis, decisions, treatments, quality improvement strategies and enable improved cost efficiency while maintaining a high quality of care (Koh & Tan, 2011).

Healthcare organizations today are capable of generating and collecting a large amount of data. These data are very complex and voluminous and difficult to be analyzed by traditional tools and methods. With the use of data mining techniques it is possible to extract interesting and useful knowledge and regularities. Knowledge acquired in this

manner, can be used in appropriate area to improve work efficiency and enhance quality of decision making process (Milovic, 2012). The objective of healthcare resource management is to manage resource allocation effectively by carefully looking into high risk areas. Further, it should also predict the requirement and usage of various resources accordingly (Sharma & Mansotra, 2014).

The data mining helps in planning and implementation of healthcare resource management activities like identifying and tracking chronic diseases states and high risk patients and reducing the number of inpatients in hospital (Sharma & Mansotra, 2014).

However, a major challenge facing healthcare organizations (hospitals, medical centers) is the provision of quality services at affordable costs.

To this end, the main aim of this research is help the decision makers to plan and manage hospitals resources.

1.2 Statement of the Problem

According to Palestinian Central Bureau of Statistics (PCBS), Gaza strip is often referred to as one of the most densely populated places on earth, the population density in Palestine at the end of 2012 was 724 individuals per square kilometer (km2): 475 individuals/km2 in the West Bank and 4,583 individuals/km2 in Gaza Strip (Palestinian Central Bureau of Statistics (PCBS), 2013).

Recent years have a steady increase in the number of births, in 2008 was 50,390 and in 2013 was 55,897, and most of these births were in governmental hospitals by 71.3 % in 2013 according to the Palestinian Health Information Center (Health Information Center - Palestinian Ministry of Health, 2014).

In context the Occupancy rate in obstetric services was 146% in 2012 (Health Information Center - Palestinian Ministry of Health, 2014).

The problem is in increasing pressure on governmental maternity hospitals and birthing centers, so it requires more attention to this cause from the government and decision makers.

Predicting the length of stay (LOS) of patients in a hospital is important in providing them with better services and higher satisfaction, as well as helping the hospital management plan and managing hospital resources as meticulously as possible.

This study uses the methods of data mining to analyze the data in the central database of the childbirth registration hospitals and predict the Length of stay then estimate cost of childbirth in maternity hospitals.

The basic aim of this research is to answer the following question: How to implement data mining tools to improve childbirth services management and optimize medical resources?

1.3 Objectives

The major objective of this study was to:

- To understand the potential role of data mining tools in facilitating the managerial decision making process in healthcare resource management.
- To understand the data adequately enough to do efficient data mining, and identify the critical success factors in healthcare resource management.
- Explore the possibility of use the critical success factors in current data at healthcare fields.

- To check the quality of data in birth registration system within the Ministry of Health.
- To predict length of stay and demand of resources by uses data mining.
- Introduce the necessary recommendations.

1.4 Significance of the thesis

The significant of this research comes from the following main points:

- Using Data mining in health care considered new and emergent science for healthcare fields.
- The importance of this study to other researchers who interested in the field of health care management resources.
- Encourage the health care organization using data mining technology to benefit from the large databases, and turning the massive data into meaningful information, to develop their work.
- Increase awareness about the importance of Data mining to understand the patient behaviors and delivering the offers and services to match their needs.
- Improving, strengthening and adjusting of quality indexes for data, standards, plans and treatments
- To foster scientific discussion and disseminate new knowledge discover data mining techniques applied to health care.

1.5 Research Methodology

The current research follows a quantitative approach in its research design to achieve the above stated objectives build a predictive model using data mining tools by

used the Cross-Industry Standard Process for Data Mining (CRISP-DM) model, which contains six phases as shown in Figure 1.1.

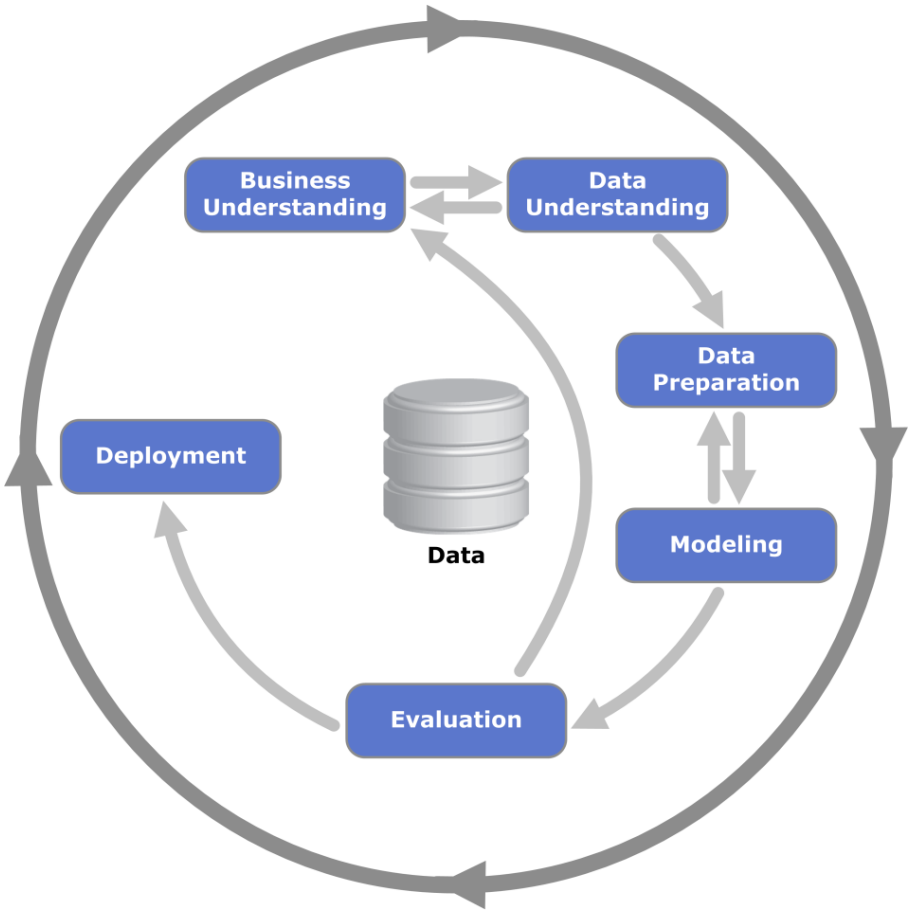


Figure 1.1 Phases of the CRISP-DM reference model (Chapman et al., 2000)

The life cycle of CRISP-DM model consists of six phases:

A. Business Understanding

This initial phase includes determining and understanding business objectives, and developing a project plan. The main objective of this work was to predict the length of stay to improve beds management.

B. Data Understanding

The data understanding phase starts with an initial data collection and proceeds with activities in order to get familiar with the data, to identify data quality problems, to

discover first insights into the data or to detect interesting subsets to form hypotheses for hidden information.

C. Data Preparation

This step covers all activities needed to construct the final dataset, which constitutes the data that will be fed into DM tool(s) in the next step. It includes Table, record, and attribute selection; data cleaning; construction of new attributes; and transformation of data.

D. Modeling

To carry out initial analysis various data mining software tools like visualization, generalized rule induction, cluster analysis are used to generate basic idea regarding usefulness of particular techniques on the available data.

E. Evaluation

The Models obtained are being evaluated in the context of established business objectives. This help in complete scan of other requirements, which results into moving towards. At the end of this phase, a decision about the use of the data mining results should be reached. The key sub steps in this step include evaluation of the results, process review, and determination of the next step.

1.6 Data Collection

The source of data about admission into maternity hospitals was obtained from internal computerized system. The database has attributes designed to store administrative information, obstetric and medical history, and preoperative care.

1.7 Research Population

This research applied on three government maternity hospitals in Gaza strip. The dataset used information of pregnant women who delivered in a public hospital in Gaza Strip between 1 January 2013 and 31 December 2013.

1.8 Research Limitations

First, this study can be limited by the accessibility of data because of heterogeneity of source of data (primary care data, laboratory, ..). Hence, data must be collected and integrated before used. One of solution proposed in the literature is the use of data warehouse. But it is a costly and time consuming project.

Secondly, data can be missing, or non-standardized such as pieces of information recorded in different format. For example, the variable positions of delivery researcher can find missing value when a delivery type is cesarean section, so it is important to specified the position in all delivery type.

1.9 Research Structure

This research structured as follow, Chapter 1 represents an introduction for the research where research importance, basis and problem are discussed, in addition to the research outline. Chapter 2 presents a theoretical framework about healthcare strategy and data mining Methodology. The literature review in Data Mining for healthcare management explained in chapter 3.Chapter 4 provides the analysis of the data collected and estimates the cost of project. Chapter 5 discusses the conclusions and the recommendations that investigated from the research.

Chapter 2

Theoretical Framework

Chapter 2 Theoretical Framework

Introduction

This chapter provides a review of data mining in healthcare management and Data Mining definitions, benefits of data mining and usage, importance of data mining in healthcare, in addition explores the data mining (knowledge Discovery) Process. In this chapter, the difference between data mining and statistical analysis will explain. And introduce to data mining techniques and their models.

2.1 Data, Information & Knowledge

The large amount of data in healthcare industry is a key resource to be processed and analyzed for knowledge extraction. The knowledge discovery is the process of making low-level data into high-level knowledge (Shelly, Dharminder, & Anand, 2011).

With the enormous amount of data stored in files, databases, and other repositories, it is increasingly important, if not necessary, to develop powerful means for analysis and perhaps interpretation of such data and for the extraction of interesting knowledge that could help in decision-making (Deshpande & Thakare, 2010).

The simple model (Figure 2.1) of the relationships between data, information and knowledge is important to be able to promote a shared understanding of how the components of a knowledge management system connect and contribute to achieving the desired business outcome (Rowley, 2007).

Figure 0.1: The Data-Information-Knowledge-W Hierarchy (Rowley, 2007)

That its importance to understanding the difference between these words and how to use to make healthcare systems work more efficiently and effectively.

Data:

Definitions of data are offered clearly and succinctly stated, in summary the definitions that:

- Data has no meaning or value because it is without context and interpretation (Jessup & Valacich, 2002).
- Data... data is raw. It simply exists and has no significance beyond its existence (in and of itself). It can exist in any form, usable or not. It does not have meaning of itself. In computer parlance, a spreadsheet generally starts out by holding data (Bellinger, Castro, & Mills, 2004).

In other words, data are set of symbols that have very little independent meaning, relevance, or purpose. Most data are gathered and stored in some form of technology, for

example, the vital signs and laboratory information in an electronic health record (Thompson & Warren, 2009).

Information:

- Information is data which adds value to the understanding of a subject (Chaffey & White, 2010).
- Information is data that has been given meaning by way of relational connection. This "meaning" can be useful, but does not have to be (Bellinger et al., 2004).

Depends on above definitions Information occurs when the data become relevant or develop a purpose. Data become information as they are processed and understood within the context of the healthcare situation, answering the "who," "what," "where," and "when" questions (Thompson & Warren, 2009).

Knowledge

- Knowledge is the combination of data and information, to which is added expert opinion, skills, and experience, to result in a valuable asset which can be used to aid decision making (Chaffey & White, 2010).
- Knowledge is data and/or information that have been organized and processed to convey understanding, experience, accumulated learning, and expertise as they apply to a current problem or activity (Turban, Rainer, & Potter, 2005).

Knowledge occurs when we apply the data and information to obtain relevant and purposeful meaning. It answers our "how" questions and lends depth and breadth to our understanding of what is happening to our patient (Thompson & Warren, 2009).

Wisdom

- Wisdom is accumulated knowledge, which allows you to understand how to apply concepts from one domain to new situations or problems (Jessup & Valacich, 2002)

- Wisdom is the highest level of abstraction, with vision foresight and the ability to see beyond the horizon (Awad & Ghaziri, 2004).
- Wisdom is the ability to act critically or practically in any given situation. It is based on ethical judgement related to an individual's belief system (Jashapara, 2004).
- Wisdom is thought of as an inherently human trait in which you must possess the ability to judge or distinguish what is right or wrong. It is based in the very human understanding of morals, values, ethics, and cultural contexts (Bellinger et al., 2004).

That is Wisdom is unlike the previous levels (see Figure 2.1), it asks questions to which there is no (easily-achievable) answer, and in some cases, to which there can be no humanly known answer period. Wisdom is therefore, the process by which we also discern, or judge, between right and wrong, good and bad. Wisdom is a uniquely human state (Bellinger et al., 2004).

These words help Stakeholders in the field efficiently and effectively systems that are used for clustering, processing, and transforming the data into useful information; and then used to take information and apply it within the contextual situation to produce working knowledge; and then experts use them to evaluate the knowledge to understand or gain wisdom about their work (Thompson & Warren, 2009).

2.2 Data mining

Data mining has attracted a great deal of attention in the information industry and in society as a whole in recent years, due to the wide availability of huge amounts of data and the imminent need for turning such data into useful information and knowledge (Han, Kamber, & Pei, 2011).

Data mining is a science to discover knowledge from databases. The database contains a collection of instances (records or case). Each instance used by machine learning and data mining algorithms is formatted using same set of fields (features, attributes, inputs, or variables) (Maimon & Rokach, 2008).

Data mining involves discovering novel, interesting, and potentially useful patterns from large data sets and applying algorithms to the extraction of hidden information. Many other terms are used for data mining, for example, knowledge discovery (mining) in databases (KDD), knowledge extraction, data/pattern analysis, data archeology, data dredging, and information harvesting (Han et al., 2011).

Data Mining is an interdisciplinary field that combines artificial intelligence, computer science, machine learning, database management, data visualization, mathematic algorithms, and statistics. With a combination of these techniques, it is possible to find different kinds of structures and relations in the data, as well as to derive rules and models that enable prediction and decision making in new situations. It is possible to perform classification, estimation, forecasts, affinity grouping, clustering and description and visualization (Diwani, Mishol, Kayange, Machuve, & Sam, 2013).

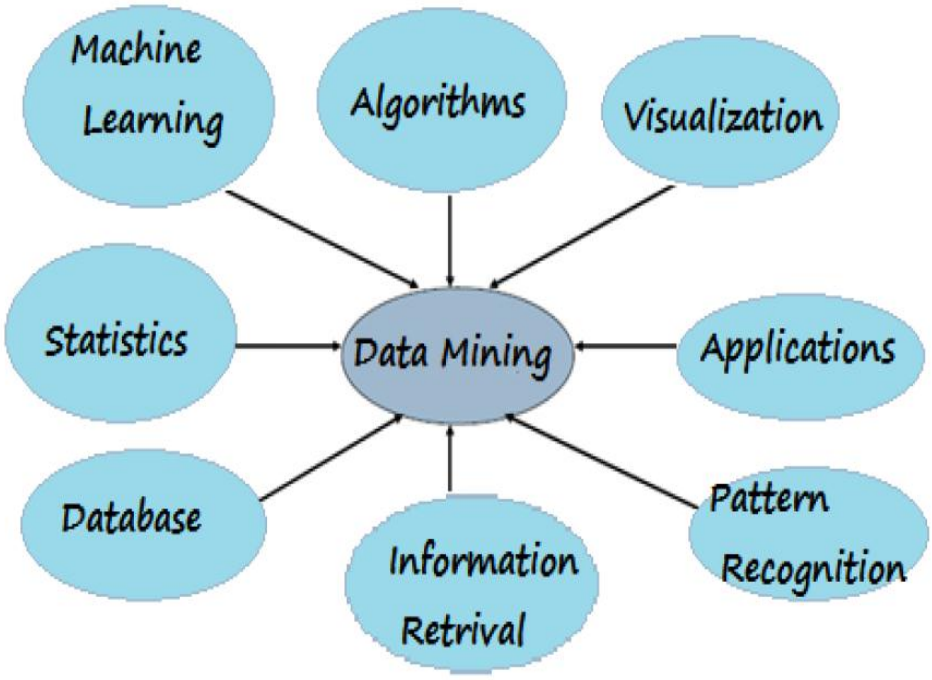


Figure 0.2: The Data Mining Architecture (Diwani et al., 2013)

Data mining approaches solving problem through analysis of massive data (Lan, Frank, & Hall, 2011) but data mining has been around much longer than analytics, at least in the context of analytics today. As analytics became an overarching term for all decision support and problem-solving techniques and technologies, data mining found itself a rather large space within that arc, ranging from descriptive exploration of identifying relationships and affinities among variables (e.g., market basket analysis) to developing models to estimate future values of interesting variable (Delen, 2014).

Obtain results from data mining can influence cost, revenue and operating efficiency while maintaining a high level of care. Healthcare organizations that perform data mining are better positioned to meet their long term needs. Data mining applications also can benefit healthcare providers such as hospitals, clinics, physicians, and patients

by predicting healthcare services and plan accordingly for expansion programs (Diwani et al., 2013).

2.2.1 The Data Mining (knowledge Discovery) Process

The objective of any data mining process is to build an efficient predictive or descriptive model of a large amount of data that not only best fits or explains it, but is also able to generalize to new data (Han et al., 2011). Based on a broad view of data mining functionality, data mining is the process of discovering interesting knowledge from large amounts of data stored in either databases, data warehouses, or other information repositories (Chen et al., 2015).

Although data mining's roots can be traced back to the late 1980s, for most of the 1990s the field was still in its infancy. Data mining was still being defined, and refined. In 1999, several sizeable companies including auto maker Daimler-Benz, insurance provider OHRA, hardware and software manufacturer NCR Corp. and statistical software maker SPSS, Inc. began working together to formalize and standardize an approach to data mining. The result of their work was CRISP-DM, the CRoss-Industry Standard Process for Data Mining. The process was designed independent of any specific tool. It was written in such a way as to be conceptual in nature—something that could be applied independent of any certain tool or kind of data. The process consists of six steps or phases, as illustrated in Figure 2.3.(North, 2012)

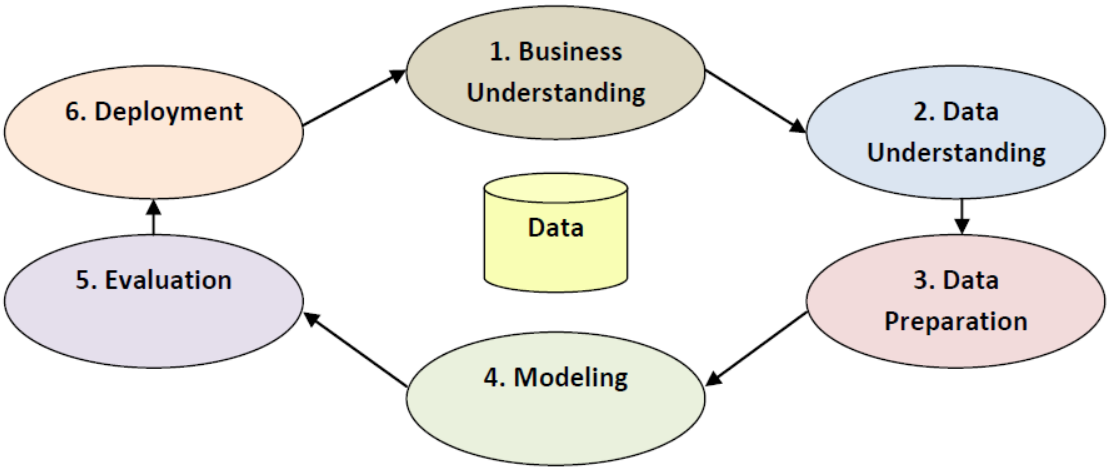


Figure 0.3: CRISP-DM reference model (North, 2012)

A. Business Understanding

The first phase of the CRISP-DM is focused on understanding the project objectives and requirements from the business point of view. This phase includes determining business objectives, assessing the current scenario, data mining goals, and developing a project plan, such as “increasing the response rate of the direct marketing".

B. Data Understanding

In this phase; after establishment of business objectives and project plan, data requirements are taken into consideration. This step includes data collection, description an exploration, and the data quality verification.

C. Data Preparation

After identifying the available data resources, they are further selected, cleaned, made into the desired form, and also being formatted. Data cleaning and data transformation are two processes which are being carried out in this phase.

D. Modeling

To carry out initial analysis various data mining software tools like visualization, generalized rule induction, cluster analysis are used to generate basic idea regarding usefulness of particular techniques on the available data. After being thorough with the understanding of data, appropriate models can be applied on given data type.

E. Evaluation

The Models obtained are being evaluated in the context of established business objectives. This help in complete scan of other requirements, which results into moving towards previous phases of CRISP-DM. The iterative procedure of CRISP-DM model helps in better understanding of business requirements. The results obtained by model evaluation provide a deeper insight into business operations of an organization.

F. Deployment

The effective models can be obtained by use of the knowledge discovered from the previous phases of the CRISP-DM, which can be further applied to particular situation or predictions. These models are further monitored depending upon the changes in operating conditions of particular business domain. If there is any major change in the operating conditions of a particular business domain, then the model has to design afresh so as to have an effective model.

This set of previous processes is summarized in the following diagram (Figure 0.4).

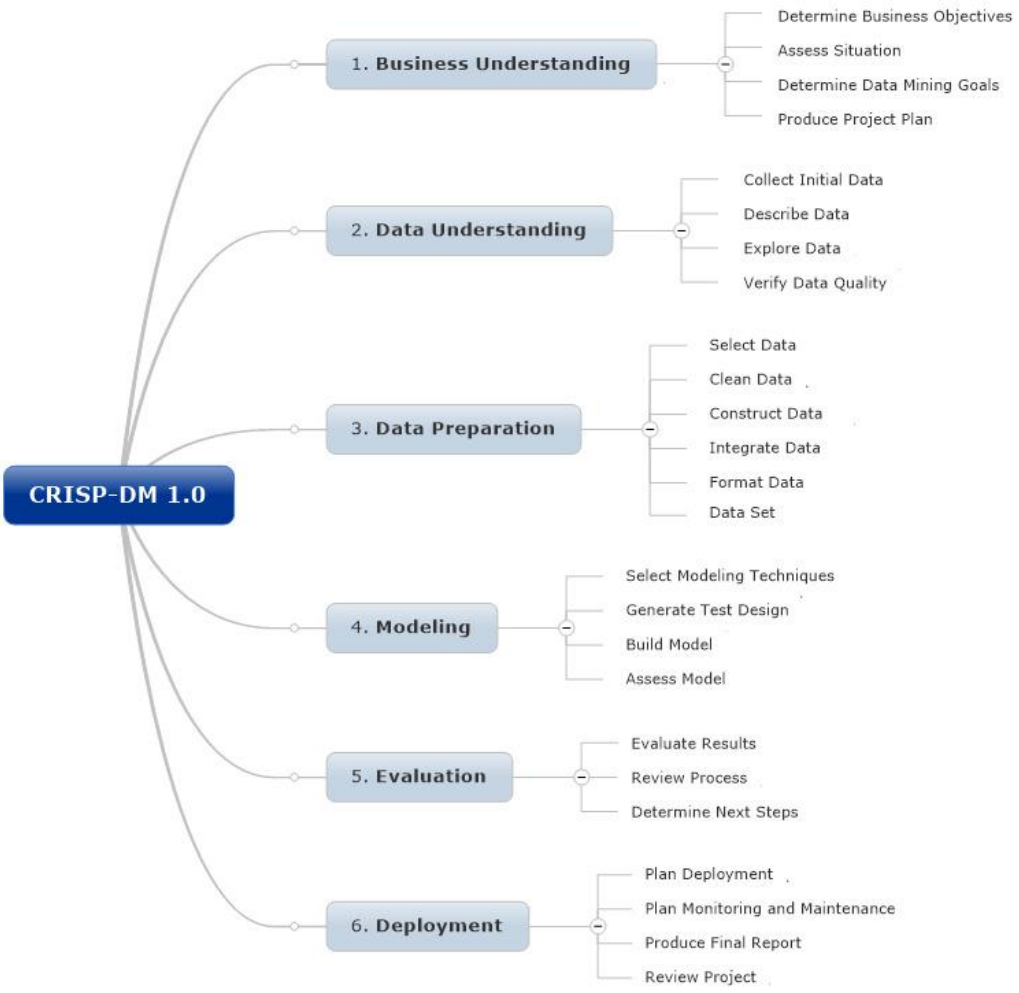


Figure 0.4: CRISP-DM summarize model (Chapman et al., 2000)

2.3 Data mining techniques

Data mining techniques are used in healthcare management for, diagnosis and treatment, healthcare resource management, customer relationship management and fraud and anomaly detection.

Data mining uses two strategies; supervised and unsupervised learning. In supervised learning, a training set is used to learn model parameters whereas in unsupervised learning no training set is used. Unsupervised learning refers to modeling

with an unknown target variable. The models are solely descriptive. The goal of the process is to build a model that describes interesting regularities in the data.

Tasks of Data mining can be separated into descriptive and predictive. Descriptive tasks have a goal on finding human interpreted forms and associations, after reviewing the data and the entire construction of the model, prediction tasks tend to predict an outcome of interest.

The prediction is one of the data mining techniques that discover relationship between independent variables and relationship between dependent and independent variables. The main predictive and descriptive data mining tasks can be classified as following:

2.3.1 Classification Models

Classification is important for management of decision making. This technique employs a set of pre-classified examples to develop a model that can classify the population of records at large (S. El-Sappagh, El-Masri, Riad, & Elmogy, 2013).

The goal of classification is to accurately predict the target class for each case in the data. The accuracy of the classification rules are estimated using test data (Kesavaraj & Sukumaran, 2013).

There are many methods to classify the data, including decision tree induction, frame-based or rule-based expert systems, hierarchical classification, neural networks, Bayesian network, and support vector machines (see Figure2.5) (Chen et al., 2015).

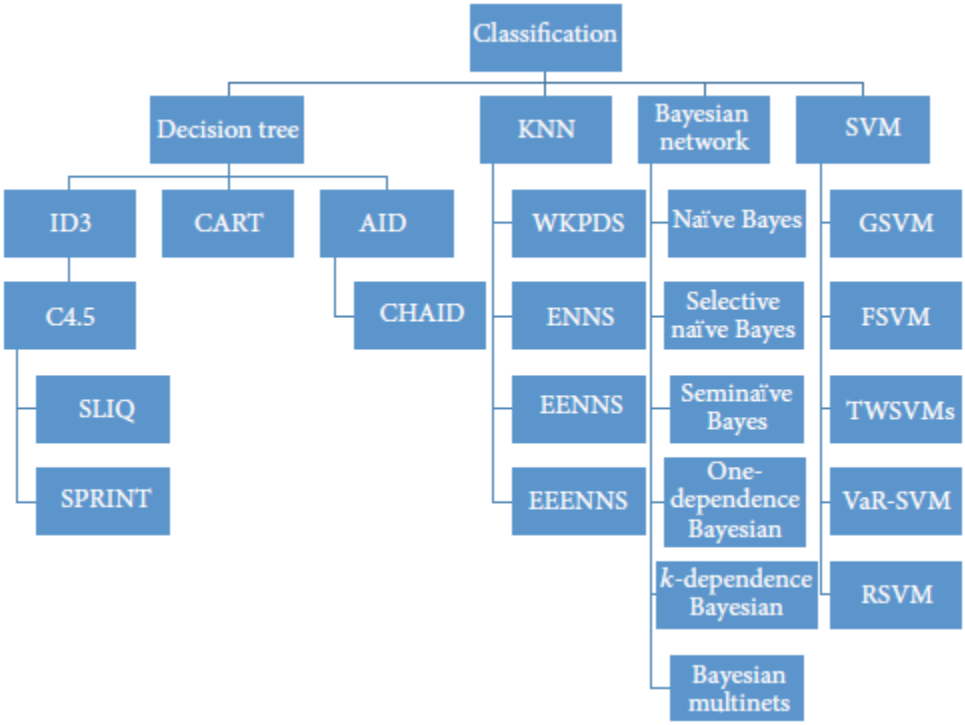


Figure 0.5: The research structure of classification (Chen et al., 2015)

Decision Tree:

DT is similar to the flowchart in which every non-leaf nodes denotes a test on a particular attribute and every branch denotes an outcome of that test and every leaf node have a class label. The node at the top most labels in the tree is called root node.(Tomar & Agarwal, 2013). Using Decision Tree, decision makers can choose best alternative and traversal from root to leaf indicates unique class separation based on maximum information gain (Apté & Weiss, 1997). A typical decision tree is shown in Figure 2.6 below.

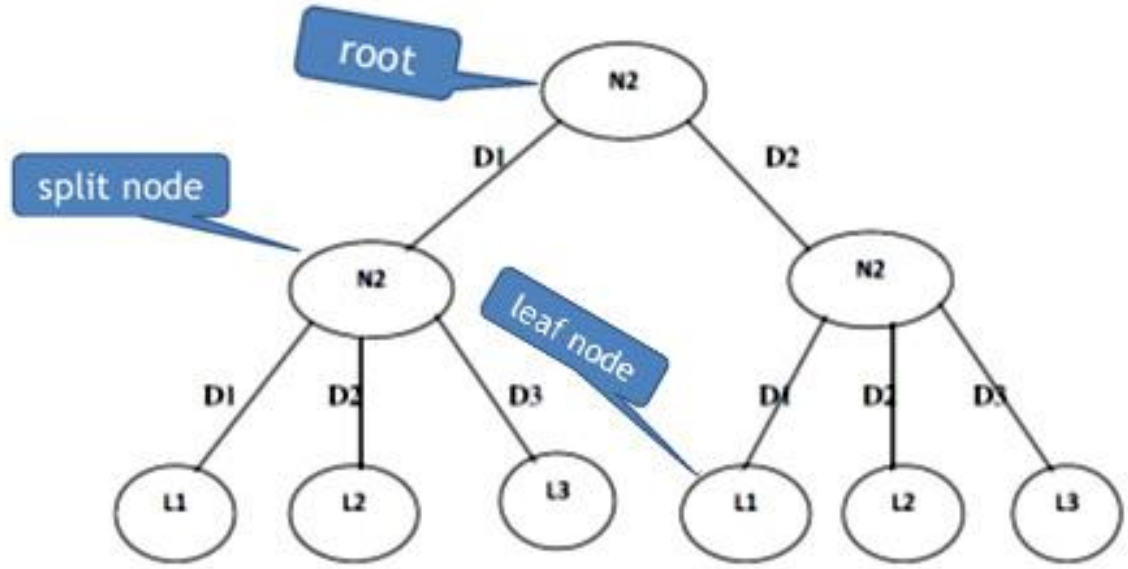


Figure 0.6: A Decision Tree with Decision (Ni) and Leaf (Li) nodes, and decisions (Di) (Tomar & Agarwal, 2013)

Decision tree algorithms include Iterative Dichotomiser 3 (ID3), assistant algorithm; C4.5, C5, and CART .The decision tree is performed with separate recursive observation in branches to construct a tree for prediction. The splitting algorithms – i.e. Information gain (used in ID3, C4.5, C5), Gini index (used in CART), and Chi-squared test (used in CHAID) – are used to identify a variable and the corresponding threshold, and then split the input observation into two or more subgroups (Xing, Wang, Zhao, & Gao, 2007)

K-Nearest Neighbors:

The k Nearest Neighbor algorithm (k-NN) is an instance based machine learning algorithm. k-NN is very simple to understand but works amazingly well (Thirumuruganathan, 2010). The idea behind k-NN method for classifying objects is based on the closest training cases in the feature space. The k-NN finds the k closest instances to a predefined instance and decides its class label by identifying the most

frequent class label among the training data that have the minimum distance between the query instance and training instances. Figure 2.7 shows an example of k-NN.

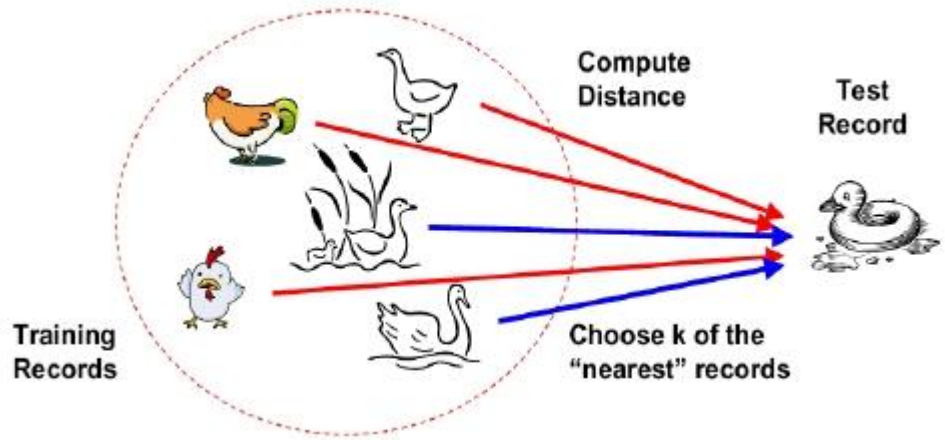


Figure 0.7: Example of k-NN (Pang-Ning, Steinbach, & Kumar, 2006)

There are a lot of different improvements for the traditional KNN algorithm, such as the Wavelet Based K-Nearest Neighbor Partial Distance Search (WKPDS) algorithm (Hwang & Wen, 1998), Equal-Average Nearest Neighbor Search (ENNS) algorithm (Jeng-Shyang, Yu-Long, & Sheng-He, 2004).

Bayesian networks:

Bayesian networks are directed acyclic graphs whose nodes represent random variables in the Bayesian sense. Edges represent conditional dependencies; nodes which are not connected represent variables which are conditionally independent of each other. The structural model (Figure 2.8). Based on Bayesian networks, these classifiers have many strengths, like model interpretability and accommodation to complex data and classification problem settings (Bielza & Larranaga, 2014). The research includes Naïve Bayes, selective naıve Bayes, seminaïve Bayes, one-dependence Bayesian classifiers, K-dependence Bayesian classifiers, Bayesian network-augmented Naïve Bayes, unrestricted Bayesian classifiers and Bayesian multinets (Chen et al., 2015).

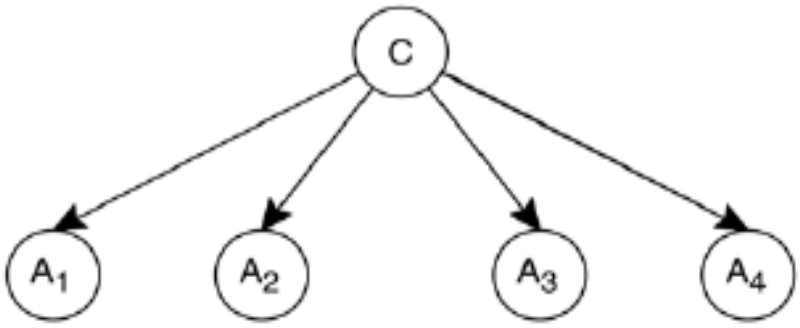


Figure 0.8: A Representation of a Bayesian Classifier Structure (Jiang, Zhang, & Cai, 2009)

Support Vector Machines:

The Support Vector Machines (SVM) is a method used for the classification of linear and non-linear data. This method uses the mapping of non-linear data to transform the training data in higher dimensions (Han et al., 2011). The basic intuition behind SVM is to find a hyperplane that discriminates between two classes (implementations for more than two classes are available), but with an extra constraint that the margin between classes should be maximized. The instances that are the closest to the hyperplane are called support vectors (Morton et al., 2014).

2.3.2 Clustering

Clustering can be considered as identification of similar classes of objects and divide data into meaningful groups (see Figure 2.9). Clustering techniques can find out overall distribution pattern and correlations among data attributes. Clustering various methods are applicable as partitioning methods (e.g. k-means, k-medoids), hierarchical methods (e.g. chameleon, CURE), density-based methods (e.g. DBSCAN, OPTIC), grid-based methods (e.g. STING, CLIQUE) and model-based methods (e.g. statistical and neural network approaches) (S. El-Sappagh et al., 2013).

Searching for clusters involves unsupervised learning (Ansari et al., 2013). In information retrieval, for example, the search engine clusters billions of web pages into different groups, such as news, reviews, videos, and audios. One straightforward example of clustering problem is to divide points into different groups (Sun, 2011).

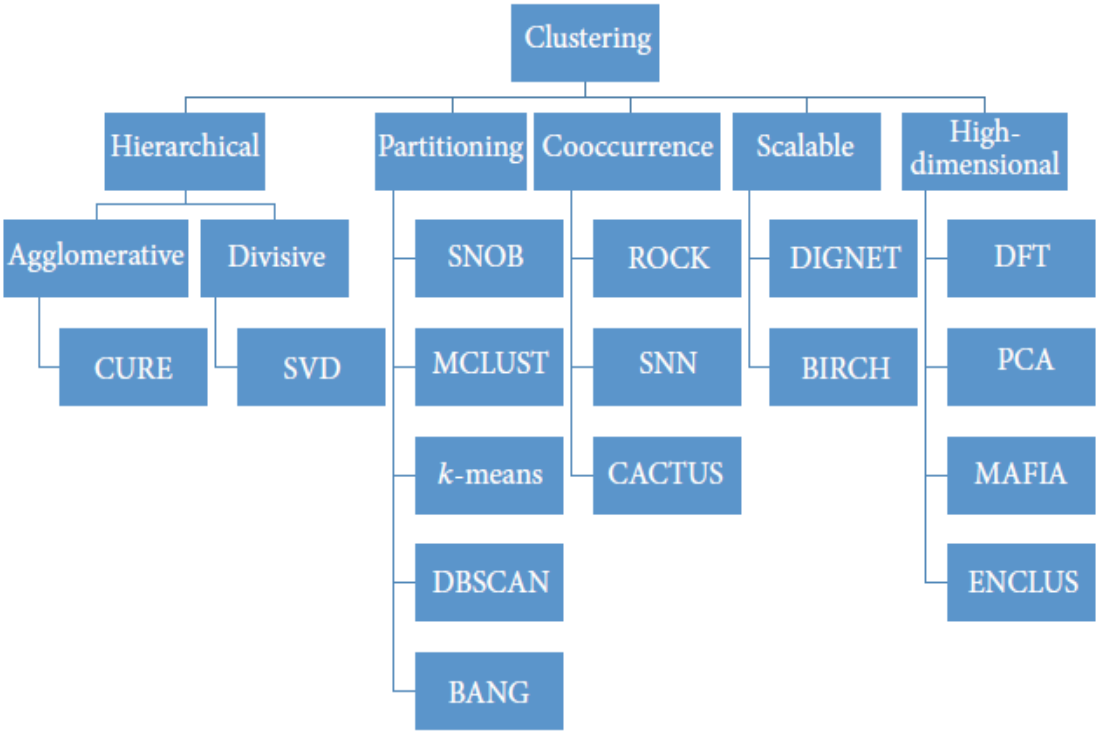


Figure 0.9: The research structure of clustering (Chen et al., 2015)

2.3.3 Association Analysis

Association rule analysis is descriptive data mining task which includes determining patterns, or associations, between elements in data sets. Associations are represented in the form of rules, or implications (Milovic, 2012). It focuses on the market basket analysis or transaction data analysis, and it targets discovery of rules showing attribute value associations that occur frequently and also help in the generation of more general and qualitative knowledge which in turn helps in decision making(Gosain &

Bhugra, 2013). The research structure of association analysis is shown in Figure 2.10. There are several classifications of these algorithms such as according to the types of values handled (e.g. Boolean and quantitative rules), according to the number of dimensions (e.g. single and multi-dimensional) and according to the level of abstraction (e.g. single and multi-level rules) (S. H. El-Sappagh, El-Masri, Elmogy, Riad, & Saddik, 2014).

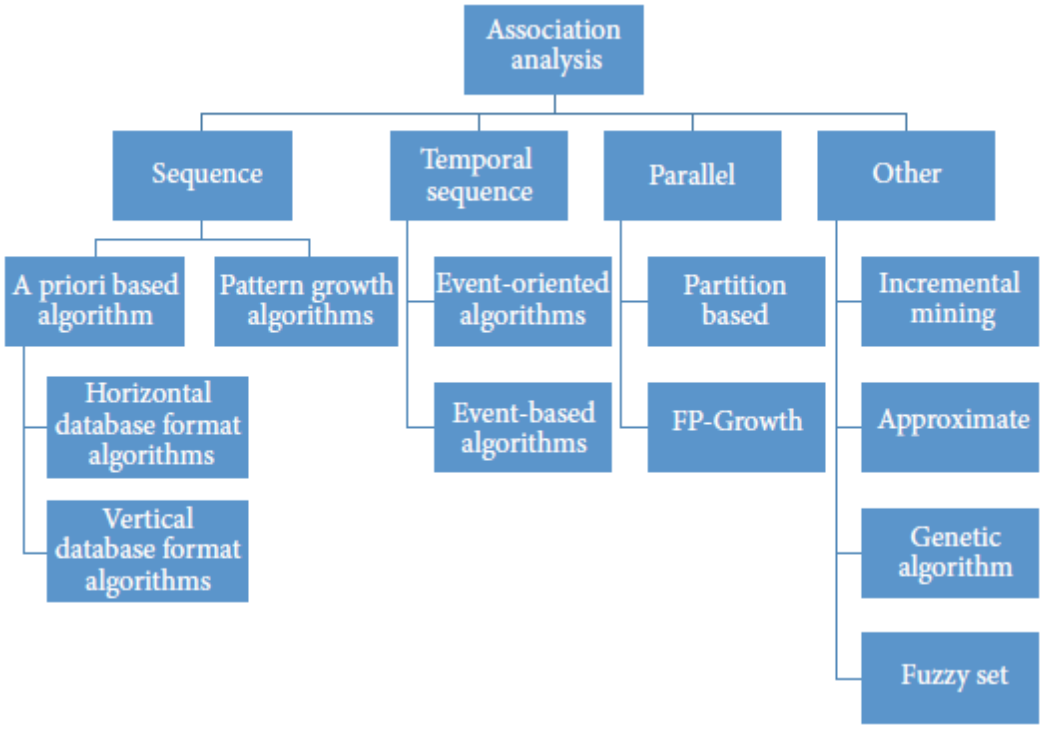


Figure 0.10: The research structure of association analysis (Chen et al., 2015)

2.4 Data Mining vs. Statistical Analysis

The difference between traditional 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 (Hand, 2007).

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 (Bath, 2004).

Data mining has various purposes and different in the traditional statistical techniques, the following table to illustrate this difference (Pareek, 2006).

Table 0.1: Data Mining vs. Statistical Analysis (Pareek, 2006)

Data Mining	Statistical Analysis
Data mining does not require a hypothesis.	Statisticians usually start with a hypothesis (a question or assumption).
Data-mining algorithms in the tool can automatically develop the equations.	Statisticians have to develop their own equations to match their hypotheses.
Data-mining tools can use different types of data, not just numerical data.	Statistical analysis uses only numerical data.
Data mining depends on clean, well documented data.	Statisticians can find and filter dirty data during their analysis.
Data-mining results are not easy to interpret, and a statistician must still be involved in analyzing the data-mining results and conveying the findings to the business managers and executives.	Statisticians interpret their own results and convey these results to business managers and executives.

In statistical analyses, decision makers formulate a hypothesis that then has to be confirmed on the basis of sample evidence by using different statistical validation techniques. But learning models, which represent the core of data mining projects, are capable of playing an active role by generating predictions and interpretations which actually represent new knowledge available to the users. The following table illustrate this (Carlo, 2009).

Table 0.2: Data Mining vs. Statistics (Carlo, 2009)

Data Mining	Statistics
Identification of patterns and recurrences in data to obtain knowledge. Ex. characterization of home loan applicant and prediction of future applicants.	Verification of hypotheses formulated by analysts to validate it. Ex. analysis of variance of incomes of home loan applicants.

2.5 Software Tools for Data Mining

There are many data mining software and application packages available on the market nowadays. According to a survey by KD-nuggets, the most popular data mining tool used by real projects is RapidMiner (King & Satyanarayana, 2013).

RapidMiner is probably the best open-source data mining tool these days. It can directly access/export data from/to different storages (databases, .csv, .xml, .arff files etc.) and it can handle large datasets. RapidMiner (formerly YALE (Yet Another Learning Environment)) for text preprocessing and classification (Mierswa, Wurst, Klinkenberg, Scholz, & Euler, 2006). Previous versions (v. 5 or lower) were open source. The latest one (v. 6) is proprietary for now, with several license options (Starter, Personal, Professional, Enterprise). The Starter version is free with limitations only in respect to maximum allocated memory size (1 GB) and input files (.csv, Excel). The tool has become very popular in several recent years and has a large community support. RapidMiner offers an integrating environment with visually appealing and user-friendly GUI. Everything in RapidMiner is focused on processes that may contain subprocesses. Processes contain operators in the form of visual components. Operators are implementations of DM algorithms, data sources, and data sinks. The dataflow is constructed by drag-and-drop of operators and by connecting the inputs and outputs of corresponding operators. RapidMiner also offers the option of application wizards that

construct the process automatically based on the required project goals (e.g. direct marketing, churn analysis, sentiment analysis). There are tutorials available for many specific tasks so the tool has a stable learning curve (Jovic, Brkic, & Bogunovic, 2014).

RapidMiner provides more than 1,000 operators for all main machine learning procedures, including input and output, and data preprocessing and visualization. Process Documents from files is a RapidMiner operator that Generates word vectors from a text collection stored in multiple files. It also provides different term weighting schemes, and term pruning options (Mierswa et al., 2006).

The graphical user interface of this software consists of windows, from which windows “Operators“, “Process“, “Parameters“ are the most important. Useful are also windows “Repositories“, “Problems“ or “Help“. The concept of RapidMiner is based on a visual programming and process creation. The work with this software is intuitive while creating “light” processes. To evaluate data or create a model, additional knowledge of this tool is a big advantage. “Operators” represent functions and they form a process under “Process” window. In “Result Overview” window, if operators are connected properly. If not, “Quick Fixes” option under “Problems” window can help to handle the problem. “Parameters” are used to set properties of operators. “Help” is the description of operators. “Repositories” is the list of folders and RapidMiner files. Example of a process and RapidMiner environment is shown in Appendix A.

Thus, by the results mentioned above, RapidMiner is the free available software that could be the best place for learning data mining and performing the decision tree method in this research. Therefore, RapidMiner was selected for the performing decision tree method.

2.6 Measuring Data Mining Performance

To obtain reliable results, the extracted knowledge should be evaluated by a comparison of results obtained with various supervised classification methods and using several measures (Dunham, 2006). Performance evaluation of classifiers can be measured by hold-out, random sub-sampling, cross-validation and bootstrap (Esfandiari, Babavalian, Moghadam, & Tabar, 2014). Furthermore, performance measures can be used to analyze predictive models. They based on four values of the confusion matrix (Figure 2.11): true positives (TP), false positives (FP), true negatives (TN), false negatives (FN).

Confusion matrix: - A confusion matrix is a very useful tool for understanding results; Confusion matrix shows the counts of the actual versus predicted class values. It shows not only how well the model predicts, but also presents the details needed to see exactly where things may have gone wrong.

Given two-class, a confusion matrix may be used to summarize the predictive performance of a classifier on test data. It is commonly encountered in a two-class format, but can be generated for any number of classes. Suppose we have a two-class problem with classes referred to as positive and negative. (Han et al., 2011).

		Observed	
		True	False
Predicted	True	True Positive (TP)	False Postive (FP)
	False	False Negative (FN)	True Negative (TN)

Figure 0.11 confusion matrix (Benbelkacem, Kadri, Chaabane, & Atmani, 2014)

True positives (TP): These refer to the positive tuples that were correctly labeled by the classifier. Let TP be the number of true positives.

True negatives (TN): These are the negative tuples that were correctly labeled by the classifier. Let TN be the number of true negatives.

False positives (FP): These are the negative tuples that were incorrectly labeled as positive (e.g., tuples of class buys computer = no for which the classifier predicted buys computer= yes). Let FP be the number of false positives.

False negatives (FN): These are the positive tuples that were mislabeled as negative (e.g., tuples of class buys computer = yes for which the classifier predicted buys computer= no). Let FN be the number of false negatives.

Four performance measures are used:

Accuracy: is a rate of correct classification defined by: $\frac{TP+TN}{TP+TN+FP+FN}$

Precision: is a rate of true positive classification defined by: $\frac{TP}{TP+FP}$

Recall: is the evaluation and ranking of each sample based on positive class defined by: $\frac{TP}{TP+FN}$

Kappa statistic: measures the agreement of prediction with the true class labels. A score value of 1.0 signifies complete agreement, and a value greater than 0 means that the classifier is doing better than pure random behavior (Azari, Janeja, & Mohseni, 2012).

2.7 Data Mining in Healthcare Management

Data mining techniques has been used intensively and extensively by many organizations. In healthcare, data mining is gradually increasing popularity, if not by any case, becoming increasingly essential. Data mining applications can greatly benefit all parties involved in the healthcare industry (Hájek, Holeňa, & Rauch, 2010).

For example, data mining can help healthcare insurers detect fraud and abuse; healthcare organizations can make customer relationship management decisions; physicians can identify effective treatments and best practices; and patients receive better and more affordable healthcare services. Data mining can be defined as "the process of finding previously unknown patterns and trends in databases and using that information to build predictive models" (Ngai, Hu, Wong, Chen, & Sun, 2011).

2.7.1 Application in Healthcare Management

The key directions in applying data mining for Healthcare management have been broadly classified into the following categories (Koh & Tan, 2011).

Diagnosis and Treatment: Research studies have suggested that unaided human analysis of data for decision making is unintentionally flawed. Applying data mining to even small data sets can provide protection against error-prone unaided human inference and could consequently support improved treatment decisions. Data mining could be particularly useful in medicine when there is no dispositive evidence favoring a particular

treatment option. Other key areas where data mining has been proved as an effective tool are disease diagnosis, detection and prediction.

Healthcare Resource Management: The goal here is to effectively manage resource allocation by identifying high risk areas and predicting the need for and usage of various resources. For example, a key problem in healthcare is measuring the flow of patients through hospitals and other healthcare facilities. If the inpatient length of stay (LOS) can be predicted efficiently, the planning and management of hospital resources can be greatly enhanced.

Customer Relationship Management: The principles of applying of data mining for customer relationship management in other industries are also applicable to the healthcare industry. The identification of usage and purchase patterns and the eventual satisfaction they result in can be used to improve overall customer satisfaction. The customers could be patients, pharmacists, physicians or clinics. In many cases prediction of purchasing and usage behavior can aid in designing proactive initiatives to reduce overall cost and increase customer satisfaction.

Fraud and Anomaly Detection: Data mining has been used very successfully in aiding the prevention and early detection of medical insurance fraud. The ability to detect anomalous behavior based on purchase, usage and other transactional behavior information has made data mining a key tool in variety of organizations to detect fraudulent claims, inappropriate prescriptions and other abnormal behavioral patterns. Another key area where data mining based fraud detection is useful is the detection and prediction of faults in medical devices.

2.7.2 Prediction of inpatient length of stay

A key problem in the healthcare area is the measurement of flow of patients through hospitals and other health care facilities (Dao et al., 2008).

If the inpatient length of stay can be predicted efficiently, the planning and management of hospital resources can be greatly enhanced. Hence, it can effectively manage the resource allocation by identifying high risk areas and predicting the need and usage of various resources.

2.8 Chapter Summary

This chapter presented a background study of main machine learning and data mining technologies used in the present research. It also presented data mining in the field of healthcare.

In recent years, data mining has emerged as a new dimension of machine learning in Artificial Intelligence (AI). Association rule mining and classification rule mining are well-recognized areas of data mining.

The next chapter will present some related prior work on different data mining techniques, missing feature values and feature selection techniques, and the technique used in this thesis described.

Chapter 3

Related Work

3.1 Previous Studies

3.2 Commentary

3.3 The Study Contribution

Chapter 3 Related Work

Introduction

The Healthcare industry is among the most information intensive industries. Medical information, knowledge and data keep growing on a daily basis. The ability to use these data to extract useful and new information for quality healthcare is crucial. In the recent years researchers of machine learning techniques have shown interest in the medical field. Many healthcare applications are developed for enhancement of the healthcare management as well as patients' care.

Research in biomedicine, genetics and medical diagnostics involves a variety of machine learning techniques. These techniques have been used for comparison, prediction and predictive analysis. In this chapter, researcher will discuss a few studies that are relevant to work in healthcare management field. Everyday, a large collection of medical data are stored in electronic format in healthcare organizations. The analysis of such an immense amount of data is a complex and very difficult task, since each patient has his/her medical history and disease complexities and conditions. Therefore, the investigation of medical data is an open and challenging issue. To cope with this problem, data mining techniques have taken great attention to investigate large collection of medical data.

In this chapter, researcher will discuss a few studies that are relevant to work in healthcare management field.

3.1 Previous Studies

1. **(Banjari, Kenjerić, Šolić, & Mandić, 2015),** *"Cluster Analysis as a Prediction Tool for Pregnancy Outcomes".*

The aim of this study is to predict certain pregnancy outcomes by using the classification of pregnant women at early pregnancy. This study based on an approach from intelligent data mining, cluster analysis. Cluster analysis method is a statistical method which makes possible to group individuals based on sets of identifying variables. 222 pregnant women from two general obstetric offices were recruited. The main orient was set on characteristics of these pregnant women: their age, pre-pregnancy body mass index (BMI) and hemoglobin value. Cluster analysis gained a 94.1% classification accuracy rate with three groups of pregnant women.

The results are showing that Pregnant women both of older age and higher pre-pregnancy BMI have a significantly higher incidence of delivering baby of higher birth weight but they gain significantly less weight during pregnancy. Their babies are also longer, and these women have significantly higher probability for complications during pregnancy (gestosis) and higher probability of induced or caesarean delivery. Cluster analysis is a useful method in predicting and grouping pregnant women at early pregnancy, and can be used for the prevention of certain risk factors during pregnancy.

2. **(He, 2014),** *"Data Mining for Improving Health-Care Resource Deployment"*

The aim of this study is to address the problem of predicting future hospitalization periods (in days) for patients from a given set of historical patient data. The data mining techniques used linear regression, random forest and gradient boosting. For each technique used different historical data sets. The combination of data mining techniques

and historical datasets enabled to compare access and choose the combination which provides the best prediction of hospitalization period of a set of patients. In the dataset, the basic information about members is sex, age and memberId. The claims data, drugcoutn data and labcount data were available for Year1 and Year2. We also have DaysInHospitals (DIH) data for Year2.

The goal is to predict Year3 DaysInHospitals data. Findings showed that the each algorithm provided different accuracy with each model and reflected the inherent properties of the algorithm.

The conclusion from this research is that the random forest techniques are the best techniques of prediction patient hospitalization periods with this dataset. The historical dataset was used had 112 attributes (e.g. Memberid, num_ProviderID).Some of those are may be relatively unimportant to the prediction of hospitalization period (e.g. Memberid).

3. **(Mohammed, Mohammed, Fiaidhi, Fong, & Kim, 2014),** *"Clinical Narratives Context Categorization: The Clinician Approach using RapidMiner"*

The aim of this study is to describe how RapidMiner as a visual programming environment can be used for tokenization and categorization of clinical narratives. It also describes how to select the best classifier for categorization.

For this purpose the researcher collected equal number of clinical narratives from the MTSamples.com7 which is a web repository designed to give you access to a big collection of transcribed medical reports. The samples were contained from eight different clinical categories (Autopsy, Diet, Discharge Summaries, Chiropractic, Cosmetic, Dental, ENT and Radiology). By divided samples into training and testing

(80% training and 20% testing) where the textual documents for each category have been assigned to a specific directory. The K-NN has been used as the classifier.

Findings showed that the K-NN proves to provide the highest accuracy (95.5%) compared to the other classifiers. After using more sensitive tokenization such as identifying tokens based on the linguistic features, the accuracy have been raised to 97.5%. Moreover, one can enhance the accuracy further by choosing more careful document vector pruning such as the absolute pruning instead of the traditional perceptual pruning. Both the precision and recall identified K-NN to be the best compared to the other classifiers. This is an encouraging result for categorizing clinical. By focused on two categories (Current Smoker and Non-Smoker) and extracted 48 narratives for each of the two categories as well as 13 narratives test cases.

The results also show that the performance measures were as follows :(Accuracy: 80.36% -Classification Error: 19.69% -Kappa: 0.616 -Average Class Precision: 85.39% -Average Class recall: 81.25% -Absolute Error: 0.196 -Relative Error: 19.69 -Correlation: 0.661).

4. **(Ramezankhani et al., 2014)** *"Applying decision tree for identification of a low risk population for type 2 diabetes. Tehran Lipid and Glucose Study"*

The aim of this study is to create a prediction model using data mining approach to identify low risk individuals for incidence of type 2-diabetes. Using the Tehran Lipid and Glucose Study (TLGS) database for a 6647 population without diabetes, aged _20 years, followed for 12 years, a prediction model was developed using classification by the decision tree technique. Seven hundred and twenty-nine (11%) diabetes cases

occurred during the follow-up. Predictor variables were selected from demographic characteristics, smoking status, medical and drug history and laboratory measures.

The findings of the study show that overall classification accuracy was 90.5%, with 31.1% sensitivity, 97.9% specificity; and for the subjects without diabetes, precision and f-measure were 92% and 0.95, respectively.

The decision tree analysis, using routine demographic, clinical, anthropometric and laboratory measurements, created a simple tool to predict individuals at low risk for type 2-diabetes. In particular, of the 1776 cases without diabetes in the test dataset, the decision tree correctly classified 1738 cases as without diabetes, an accuracy rate of 97.9%. Further, for the 1889 individuals classified by the decision tree as cases without diabetes, 1738 (92%) cases are actually without diabetes. For the 219 person with diabetes in the test dataset, the decision tree correctly classified 68 individuals, an accuracy rate of 31.1% and for the 106 individuals classified.

5. **(Shenas, Raahemi, Tekieh, & Kuziemsky, 2014),** *"Identifying high-cost patients using data mining techniques and a small set of non-trivial attributes"*

The aim of this paper is to build predictive models and determined a small set of non-trivial attributes to identify very high-cost patients in the top 5 percentile among the general population in order to enable better proactive estimation of healthcare costs. This paper used data mining techniques, namely neural networks and decision trees. A large empirical dataset from the Medical Expenditure Panel Survey with 98,175 records was used. After pre-processing, partitioning and balancing the data, the refined dataset of 31,704 records was modeled by Decision Trees (including C5.0 and CHAID), and Neural

Networks. The performances of the models are analyzed using various measures including accuracy, G-mean, and Area under ROC curve.

The findings demonstrated that the CHAID classifier returns the best G-mean and AUC measures for top performing predictive models ranging from 76% to 85%, and 0.812 to 0.942 units, respectively. The 5 non-trivial attributes among a primary set of 66 attributes are age, history of blood cholesterol check, history of physical/sensory/mental limitations, and history of colonic prevention measures.

On other hand when the number of input attributes is high (as in our case with 39 attributes), or if the attributes are continuous values (instead of binary or discrete values) it takes longer for neural networks to converge, which impacts its accuracy. In general, the decision tree models perform better than the neural networks for analysis of the dataset and results of this study can be used by healthcare data analysts, policy makers, insurer, and healthcare planners to improve the delivery of health services.

6. **(Benbelkacem, Kadri, Chaabane, & Atmani, 2014),** *"A Data Mining-Based Approach To Predict Strain Situations In Hospital Emergency Department Systems"*

Aim of this paper is to illustrate with a real world case study, how data mining can be benefit methods to make pediatric emergency department (PED) management system more proactive face to strain situations and predict length of stay (LOS) in emergency departments (ED).

This paper presented an approach based on supervised data mining classification methods in order to predict the patient length of stay (LOS) at the pediatric emergency department in Lille regional center, France. The used data were collected from the

pediatric emergency department (PED) in Lille regional hospital center, France. The valuation of models is carried out using 10-fold cross validation on the PED data set.

The findings of the study show that the maximum accuracy score equal to 79.942 is obtained by SVM and the lowest one equal to 72.291 is obtained by Id3. On other hand the highest precision score belongs to BayesNet (0.763), but the recall was slightly better with SVM. So according all performance measures BayesNet and NB give the best result. It also shows that the decision tree models are for the most, more than 80% closer to the maximum and decision trees give comparable results compared to NB and BayesNet.

However decision tree methods obtain comparable results and can successfully be suitable for the prediction of patient's length of stay, and the information can be used to make the decision more proactive and useful for PED manager to predict LOS and detect the beginning of strain situation.

7. **(Ahmed & Hannan, 2013),** *"Data Mining Techniques to Find Out Heart Diseases"*

This research paper proposed to find out the heart diseases through data mining, Support Vector Machine (SVM), Genetic Algorithm, rough set theory, association rules and Neural Networks.

This study examined that out of the above techniques Decision tree and SVM is most effective for the heart disease. So it is observed that, the data mining could help in the identification or the prediction of high or low risk heart diseases.

The results of this paper showed that there are relatively differences in different techniques. Decision tree and SVM perform classification more accurately than the other methods. Researchers suggest that the age, sex, chest pain, blood pressure, personnel

history, previous history, cholesterol, fasting blood sugar, resting ECG, Maximum heart rate, slope, etc. that may be used as reliable indicators to predict presence of heart disease. They also suggest that data should be explored and must be verified from the team of heart disease specialist doctors.

In future, they will try to increase the accuracy for the heart disease patient by increasing the various parameters suggested from the doctors by using different data mining techniques.

8. **(Hachesu, Ahmadi, Alizadeh, & Sadoughi, 2013),** *"Use of Data Mining Techniques to Determine and Predict Length of Stay of Cardiac Patients"*

The aim of this study is to extract useful knowledge and draw an accurate model to predict the length of stay (LOS) of heart patients. The techniques used are data min¬ing classification with three algorithms, decision tree, support vector machines (SVM), and artificial neural network (ANN).

Data were collected from patients with coronary artery disease (CAD). The patient records of 4,948 patients who had suffered CAD were included in the analysis LOS is the target variable, and 36 input variables are used for prediction. Created models were evaluated using training data and then applied test data to compare the results. A confusion matrix was obtained to calculate sensitivity, specificity, and accuracy.

Findings showed that all three algorithms are able to predict LOS with various degrees of accuracy. The ensemble algorithm showed a stronger performance than other algorithms with a sensitivity of 98.2%, and the overall accuracy of SVM was 96.4% in the training set. Most single patients (64.3%) had an LOS ≤5 days, whereas 41.2% of

married patients had an LOS >10 days. The most significant variables were drug categories, such as nitrates and anticoagulants as well as CAD diagnosis.

Moreover, the study showed that Comorbidity is also a strong predictor of prolonged LOS. The presence of comorbidities, an ejection fraction <2, being a current smoker, and having social security type insurance in coronary artery patients led to longer LOS than other subjects. There was a significant tendency for LOS to be longer in patients with lung or respiratory disorders and high blood pressure. Sex and Age played a notable role t in predicting LOS since men had longer LOS than women and patients aged <50 and ≥80 statistically had increased mean LOS. The insurance type had a predictive power. Patients with social security and rural medical insurance were in LOS>10.

9. **(Ndour et al., 2013),** *"Predicting In-Hospital Maternal Mortality in Senegal and Mali"*

The aim of this study is to identify the predictors of in-hospital maternal mortality among women attending referral hospitals in Mali and Senegal. The study used a cross-sectional epidemiological survey in 46 referral hospitals in Mali and Senegal from October 1st 2007 to October 1st 2008. Data included 89,518 women who delivered in the 46 hospitals during this period. Researchers developed a tree-like classification rule (classification rule) to identify patient subgroups at high risk of maternal in-hospital mortality and fourteen features variables are used to predict in-hospital maternal mortality (age, parity, previous cesarean section, prenatal care attendance, vaginal bleeding during pregnancy, Malaria, premature rupture of the membranes and multiple pregnancy, referral from another health facility, labor induction, mode of delivery

"emergency ante-partum cesarean delivery and intra-partum cesarean delivery", hemorrhage, Prolonged/ obstructed labor, and uterine rupture).

The findings of the study association rule method will help health care professionals to identify mothers at high risk of in-hospital death. Also patients with uterine rupture, hemorrhage or prolonged/obstructed labor, and those who have an emergency ante-partum cesarean delivery have an increased risk of in-hospital mortality, especially if they are referred from another health facility. On other hand twenty relevant patterns, based on fourteen predictors variables, are used to predict in-hospital maternal mortality with 81.41% sensitivity (95% CI = [77.12%–87.70%]) and 81.6% specificity (95% CI = [81.16%–82.02%]). Furthermore, the classification approach developed in this study shows that patients who were referred from other health care facilities or had emergency ante-partum cesarean delivery (ante- and intra-partum) also have an increased risk of hospital-based mortality.

10. (Freitas et al., 2012), *"Factors influencing hospital high length of stay Outliers"*

The aim of this paper is to study variables associated with high length of stay (LOS) outliers and their evolution over time. Due to (LOS) outliers is important for the management and financing of hospitals. Data were collected from inpatient episodes in public acute care hospitals in the Portuguese National Health Service (NHS), between years 2000 and 2009. The dependent variable, LOS outliers, was calculated for each diagnosis related group (DRG) using a trim point defined for each year by the geometric mean plus two standard deviations. Logistic regression models, including a multivariable logistic regression, were used in the analysis. All the logistic regressions were fitted using generalized estimating equations (GEE).

Findings showed that the near nine million inpatient 3.9% are high LOS outliers, accounting for 19.2% of total inpatient days. The number of hospital patient discharges increased between years 2000 and 2005 and slightly decreased after that. The proportion of outliers ranged between the lowest value of 3.6% (in years 2001 and 2002) and the highest value of 4.3% in 2009. In the last years, both average LOS and high LOS outliers are increasing in Portuguese NHS hospitals. The age, type of admission, and hospital type were significantly associated with high LOS outliers.

In addition, the increasing complexity of both hospitals and patients may be the single most important determinant of high LOS outliers and must therefore be taken into account by health managers when considering hospital costs. On other hand, the outliers increased with age, from near 2.5% between 0 and 45 years, to about 5.5% for patients with more than 66 years.

11. **(Gharehchopogh, Mohammadi, & Hakimi, 2012),** *"Application of Decision Tree Algorithm for Data Mining in Healthcare Operations: A Case Study"*

The aim of this study is to explain utilization of medical data mining in determination of medical operation methods. The Data were collected from pregnant women's information that referred to delivery in Tabriz health center (80 pregnant), and then applied decision tree C4.5 algorithm to exploit operational and drastic information by genuine data. The The aim of this study is to explain utilization of medical data mining in determination of medical operation methods. The Data were collected from pregnant women's information that referred to delivery in Tabriz health center (80 pregnant), and then applied decision tree C4.5 algorithm to exploit operational and drastic

information by genuine data. The medical field is (age, number of pregnant, delivery time, blood of pressure and heart status).

The researchers classified delivery time to Premature, Timely and Latecomer, blood of pressure in three statuses of Low, Normal and High moods, and Existence or absence of heart problem.

The results indictor was refer to direct relevance between heart failure rate and deliveries those ends to caesarian section. The Heart status play key role in designed system bye Decision trees. Other results indicate that more than 75% women with inept heart status didn't have a natural delivery and over 65% of those have inept heart case savors an abnormal pressure of blood.

The conclusion of this paper investigates decision tree decision tree for 86.25% cases predicted correct results. So it can be considered as useful information obtained in important decision in determination of medicine operations.

12. **(Sana, Razzaq, & Ferzund, 2012),** *"Automated Diagnosis and Cause Analysis of Cesarean Section Using Machine Learning Techniques"*

The aim of this study is to evaluate different machine learning techniques for birth classification (cesarean or normal). The data were collected from 15 hospitals of Sargodha, Pakistan between 2nd February 2010 to 28th February 2010, by interviewing the medical experts in this area and the patients.

The researchers identified about 50 factors that can influence the type of birth. These factors include Pre pregnancy factors like maternal age, body weight, education, drinking routine, diabetes, and hypertension. Some factors are identified during pregnancy like HIV, uterine rupture, blood sugar, abnormal presentation. By

applied Chi Square test on pairs, with each pair have one attribute of birth and the other is selected from the list of attributes.

A birth classification model is built using decision tree and artificial neural networks.

The findings of the study show that the model can classify the births into normal and cesarean with an average accuracy, precision and recall of 80%, 85% and 84% respectively. The birth type is influenced by age, income, condition, blood pressure, pulse surgery, exercise, pre-pregnancy, food and use of multivitamins. Therefore use of multivitamins is associated with normal births. High blood pressure and pulse rate are associated with cesarean births. Lack of education is also associated with cesarean births.

13. **(BISET, 2011),** *"Predicting Low Birth Weight Using Data Mining Techniques On Ethiopia Demographic And Health Survey Data Set"*

The aim of this study is to Predict LBW in various communities in the country, and addressing the factors associated with low birth weight. Data mining techniques were used to achieve the goal of this study; data was collected from EDHS 2005 (Ethiopia Demographic Health Survey) data set. A total of 9861 records were used for the experiments. Weka tool was used to discretize some attributes of numeric values and process missing values. The selected data mining techniques for predicting low birth weight was classification. J48 decision tree classifier and PART rule induction algorithms were selected for experiments.

These experiments has been done using pruning with all and reduced attributes, by giving J48 classifiers parameters in different values. Findings showed that the decision tree with prunning can improve decision tree's classification performance compared to

without tree pruning. The factors that determinant and predicted low birth weight (antenatal visits during pregnancy, mother's educational level, and marital status, Iodine contents in salt, region, and age of mother, numbers of birth , wealth index , place of residence. The overall best performance was achieved by J48 decision tree classifiers using pruned technique , confidence factor of 0.5 , minimum numbers of instance at 2 with all attributes data set, with a recall (true positive rate) of 95%, a false positive rate of 1%, a precision (positive predictive value) of 95%, and an accuracy of 94.7%. The second algorithm conducted in this research is PART rule induction algorithms.

The result show accuracy of 94.35% and correctly and Incorrectly Classified Instances are 9304 and 557 respectively and with recall (true positive rate) of 94%. In general, the results from this study were encouraging; it can be used as decision support aid for health practitioner. The extracted rules in both the algorithms (J48 decision tree, PART rule induction) are very effective for the prediction of low birth weight.

14. **(H.-Y. Chen, Chuang, Yang, & Wu, 2011),** *"Exploring the risk factors of preterm birth using data mining"*

The aim of this study is to explore the risk factors of preterm. Hence the Preterm birth is the leading cause of perinatal morbidity and mortality. Data mining with neural network and decision tree C5.0 were used. The original medical data were collected from a prospective pregnancy cohort by a professional research group in National Taiwan University.

Using the nest case-control study design, a total of 910 mother–child dyads were recruited from 14,551 in the original data. Thousands of variables are examined in this data including basic characteristics, medical history, environment, and occupation factors

of parents, and variables related to infants. A neural network was used to mine the 15 most important factors related to preterm with coefficients larger than 0.0300.

These factors were as follows: number of birth, paternal smoking, hemorrhage during pregnancy, parity, maternal age, paternal occupation, maternal hypertension, medicines taken during pregnancy, maternal gynecological diseases, maternal body height, maternal body weight before pregnancy, paternal age, paternal drinking, previous preterm birth, and vitamins taken during pregnancy. A decision tree C5.0 was used to classify the risk factors, so high risk groups for preterm birth would be detected.

Findings showed that multiple birth, hemorrhage during pregnancy, age, disease, previous preterm history, body weight before pregnancy and height of pregnant women, and paternal life style risk factors related to drinking and smoking are the important risk factors of preterm birth. Through the construction of a decision tree, 17 rules were explored to predict preterm birth. Ten of these rules, with an accuracy of 80% or more the multiple birth was the highest risk factor, with hemorrhage during pregnancy the second most important the specific risk factors related to pregnant women were previous preterm birth, diseases, body height, and weight before pregnancy, while the, paternal risk factors were smoking, drinking, age, and occupation.

15. **(GOKSEN, EMINAGAOGLU, & DOGAN, 2011),** *"Data Mining In Medical Records For The Enhancement Of Strategic Decisions: A Case Study"*

The aim of this study is to explore the role of Data Mining Techniques for the enhancement of strategic decisions in medical records. In the first phase of the study, the data are processed and transformed into a suitable format for data mining. In the second phase, some of the association rule algorithms are applied to the data set in.

The data were collected from the hospital's outpatient clinic Oracle 9i v.9.2.0.1 database system. 256816 records and 9 different fields were used for data mining analysis in the study (Patient Gender Male, Female - Day of Week - Date - Time -2-Hour Period - Department Code -Patient Type Index-Diagnosis Code-Case Explanation). Weka version 3.6.0 was used as the software for the data mining analysis phase in the study.

The findings demonstrated that Predictive algorithm derived 100 different association rules. The rule with the highest accuracy had a value of 0.99498 and the one with lowest accuracy was observed as 0.9733. Some few but strategic and meaningful conclusions were achieved which were confirmed by the senior management in the hospital. First, on any day of the week, if the patient arrival time period is 22:00-00:00 at night and the patient is recorded as a non-standard emergency outpatient clinic and if the patient is also retired, then it could be a female patient. Second, if the patient's arrival time is 00:00-02:00 at night and the day is Saturday or Sunday and the department is emergency service and if it needs an immediate operation for surgery, it is probably a male patient. Finally, some necessary precautions and some necessary changes in the daily operations might be achieved by the hospital management that could improve the efficiency and quality of services given to outpatients.

This study it is shown that only a few relations and knowledge based conclusions could be made to support and enhance the decisions and managerial strategies of the hospital management.

16. **(Tanuja, Acharya, & Shailesh, 2011),** *"Comparison of different data mining techniques to predict hospital length of stay"*

The aim of this paper is to present the performance analysis of different data mining techniques to predict the inpatient hospital length of stay in a super specialty hospital.

Data set used for the analysis is real time data taken from super specialty hospital. This data set consists of 401 records with 16 parameters.

This paper was investigated four data mining techniques: Multilayer back propagation NN (MLP), Naive Bayes Classifier, K-NN method, J48 class of C4.5 decision tree. This study was analyzed around 40,000 electronic discharge summaries of super specialty hospital available in HTML file format. The analysis of these records requires two major steps: processing large volumes of textual data and developing useful predictions based on the data.

The results of this study show that the analysis of MLP has achieved better performance compared to the other three techniques. In this study, the accuracy of four data mining techniques MLP, Naives Bayes, K-NN, J48 are (87.8 %), (85.8%), (62.6%), (75.1%) respectively.

17. **(Bertsimas et al., 2008),** *"Algorithmic Prediction of Health-Care Costs"*

The aim of this paper is to utilize modern data-mining methods, specifically classification trees and clustering algorithms. Methods along with claims data from over 800,000 insured individuals over three years, to provide rigorously validated predictions of health-care costs in the third year. The data set includes both medical and pharmaceutical claims, as well as information on the period an individual (and his or her

family) was covered by the insurance policy. The data also contain basic demographic information such as age and gender based on medical and cost data from the first two years. Data mining methods were used (The Baseline Method, Classification Trees, and Clustering).

The results of this study showed that clustering method results in better predictions for current high-cost bucket members and consistently better absolute prediction error. The members with overall costs between $12,300 and $16,000 in the last 12 months of the observation period and who have acute cost profiles. The members take no more than 14 different therapeutic drug classes during that period, and have not had a heart. In addition Members in cost buckets 2 and 3 with nonacute cost profiles, less than $2,400 in pharmacy costs and on fewer than 13 therapeutic drug classes, but who have received Zyban (prescription medication designed to help smokers quit) after a seizure.

3.2 Commentary

The following can be concluded from the previous mentioned studies and the others discussed studies through this thesis:

1. This topic of Data mining is poor in the Arab countries especially in healthcare fields and most of these studies took place in foreign countries
2. There is no published paper or academic research dedicated in Palestine, which deals with the topic of data mining in healthcare domain.
3. Data mining can be applied in healthcare filed using different deployment and technology models.

4. Knowledge gained with the use of techniques of data mining can be used to make successful decisions that will improve success of healthcare organization and health of the patients.
5. Healthcare institutions that use data mining applications have the possibility to predict future requests, needs, desires, and conditions of the patients and to make adequate and optimal decisions about their treatments.
6. The need is for algorithms with very high accuracy as medical diagnosis is a significant task that needs to be carried out precisely and efficiently.
7. Many tools are available that helps stockholders to applying Data mining methods and generate models like "RapidMiner – WEKA - Orange – R".
8. The accuracy and performance is different in models according the used methods.
9. It is necessary to identify and evaluate the most common data mining algorithms implemented in modern healthcare services as data mining algorithms may give in better results for one type of problems while others may be suitable for different ones. It may be impracticable to find the best data mining algorithm suitable for all of the medical domains. Table 3.1 shows the summary of some previous studies.

Table 3.1 the summary of some previous studies.

Study	Used Techniques	Main variables	Main Finding
Banjari, Kenjerić, Šolić, & Mandić (2015)	Clustering	Age Pre-pregnancy body mass index (BMI) Hemoglobin value.	Cluster analysis gained a 94.1% classification accuracy rate with three groups of pregnant women. Pregnant women both of older age and higher pre-pregnancy BMI have a significantly higher incidence of

Study	Used Techniques	Main variables	Main Finding
			delivering baby of higher birth weight, and these women have significantly higher probability for complications during pregnancy and higher probability of induced or caesarean delivery.
He (2014)	Linear regression - Random Forest - Gradient Boosting	Sex - Age claims data – drug count data – lab count data – Days In Hospitals	That the each algorithm provided different accuracy with each model and reflected the inherent properties of the algorithm. The conclusion from this research is that the random forest techniques are the best techniques of prediction patient hospitalization periods with this dataset. The historical dataset was used had 112 attributes (e.g. Memberid, num_ProviderID).Some of those are may be relatively unimportant to the prediction of hospitalization period (e.g. Memberid).
Mohammed, Mohammed, Fiaidhi, Fong, & Kim (2014)	Classifier - K-NN	Autopsy-Diet-Discharge-Summaries-Chiropractic, Cosmetic- Dental ENT-Radiology	The K-NN proves to provide the highest accuracy (95.5%) compared to the other classifiers. Both the precision and recall identified K-NN to be the best compared to the other classifiers. That the performance measures were as follows :(Accuracy: 80.36% - Class Precision: 85.39% -Average Class recall: 81.25%).
Ramezankha	classification -	Demographic	Overall classification accuracy was

Study	Used Techniques	Main variables	Main Finding
ni et al. (2014)	Decision Tree	characteristics-smoking status-medical-drug history-laboratory measures.	90.5%, with 31.1% sensitivity, 97.9% specificity; and for the subjects without diabetes, precision and f-measure were 92% and 0.95, respectively. 1776 cases without diabetes in the test dataset, the decision tree correctly classified 1738 cases as without diabetes, an accuracy rate of 97.9%. For the 219 person with diabetes in the test dataset, the decision tree correctly classified 68 individuals, an accuracy rate of 31.1% and for the 106 individuals classified.
Shenas, Raahemi, Tekieh, & Kuziemsky (2014)	Neural Networks-Decision Trees (C5.0- CHAID)	age, history of blood cholesterol check, history of physical/sensory/mental limitations, and history of colonic prevention measures	The findings demonstrated that the CHAID classifier returns the best G-mean and AUC measures for top performing predictive models ranging from 76% to 85%, and 0.812 to 0.942 units, respectively. In general, the decision tree models perform better than the neural networks for analysis of the dataset and results of this study can be used by healthcare data analysts, policy makers, insurer, and healthcare planners to improve the delivery of health services.
Benbelkacem, Kadri, Chaabane, & Atmani	Data mining Classification		Maximum accuracy 79.942 is obtained by SVM and the lowest one equal to 72.291 is obtained by Id3. On other hand the highest precision score

Study	Used Techniques	Main variables	Main Finding
(2014)			belongs to BayesNet (0.763), but the recall was slightly better with SVM. Decision tree methods suitable for the prediction of patient's length of stay, and the information can be used to make the decision more proactive.
Ahmed & Hannan (2013)	Data mining SVM ,Genetic Algorithm, rough set theory, association rules, Neural Networks	age, sex, chest pain, blood pressure, personnel history, previous history, cholesterol, fasting blood sugar, resting ECG, Maximum heart rate, slope	Decision tree and SVM perform classification more accurately than the other methods. Main variables as age, sex, chest pain, blood pressure, resting ECG, Maximum heart rate, slope, etc. that may be used as reliable indicators to predict presence of heart disease. In future, they will try to increase the accuracy for the heart disease patient by increasing the various parameters suggested from the doctors by using different data mining techniques.
Hachesu, Ahmadi, Alizadeh, & Sadoughi (2013)	Data mining classification Decision tree Support vector machines (SVM) Artificial neural network (ANN).	36 input variables are used for prediction. Comorbidity Current Smoker Social security type insurance Sex, Age, high blood	The ensemble algorithm showed a stronger performance than other algorithms with a sensitivity of 98.2%, and the overall accuracy of SVM was 96.4% in the training set. Most single patients (64.3%) had an LOS ≤5 days, whereas 41.2% of married patients had an LOS >10 days. Sex and Age played a notable role t in predicting LOS. The insurance type had a predictive power. Patients with social security and rural medical insurance were in LOS>10.

Study	Used Techniques	Main variables	Main Finding
Ndour et al., (2013)	classification rule	age, parity, previous cesarean section, prenatal care attendance, vaginal bleeding during pregnancy, Malaria, premature rupture of the membranes and multiple pregnancy, referral from another health facility, labor induction, mode of delivery , hemorrhage, obstructed labor, uterine rupture	Patients with uterine rupture, hemorrhage or prolonged/obstructed labor, and those who have an emergency ante-partum cesarean delivery have an increased risk of in-hospital mortality, based on fourteen predictors variables, are used to predict in-hospital maternal mortality with 81.41% ,the classification approach developed ,shows that patients who were referred from other health care facilities or had emergency ante-partum cesarean delivery (ante- and intra-partum) also have an increased risk of hospital-based mortality.
Freitas et al., (2012)	Logistic regression	diagnosis related group (DRG)	Near nine million inpatient 3.9% are high LOS outliers, accounting for 19.2% of total inpatient days. The proportion of outliers ranged between the lowest value of 3.6% (in years 2001 and 2002) and the highest value of 4.3% in 2009. The age, type of admission, and hospital type were significantly associated with high LOS outliers. The outliers increased with age, from near 2.5% between 0 and 45 years, to about 5.5% for patients with more than 66 years.
Gharehchopogh,	decision tree C4.5	age, number of pregnant, delivery	The results indictor was refer to direct relevance between heart failure rate

Study	Used Techniques	Main variables	Main Finding
Mohammadi , & Hakimi (2012)		time, blood of pressure, heart status	and deliveries those ends to caesarian section. Result investigates decision tree decision tree for 86.25% cases predicted correct results. So it can be considered as useful information obtained in important decision in determination of medicine operations.
Sana, Razzaq, & Ferzund (2012)	classification (cesarean or normal) Decision tree Artificial neural networks.	Pre pregnancy factors like maternal age, body weight, education, drinking routine, diabetes, and hypertension. Some factors are identified during pregnancy like HIV, uterine rupture, blood sugar, abnormal presentation.	The model can classify the births into normal and cesarean with an average accuracy, precision and recall of 80%, 85% and 84% respectively. The birth type is influenced by age, income, condition, blood pressure, pulse surgery, exercise, pre-pregnancy, food and use of multivitamins. Therefore use of multivitamins is associated with normal births. High blood pressure and pulse rate are associated with cesarean births. Lack of education is also associated with cesarean births.
BISET (2011)	Data mining Classification J48 decision tree, PART	antenatal visits during pregnancy, mother's educational level, and marital status, Iodine contents in salt, region, and age of mother, numbers of birth , wealth index , place of residence	Findings showed that the decision tree with prunning can improve decision tree's classification performance compared to without tree pruning. The overall best performance was achieved by J48 decision tree with accuracy of 94.7%. The PART rule with accuracy of 94.35%. The extracted rules in both the algorithms (J48 decision tree, PART rule induction) are very effective for the

Study	Used Techniques	Main variables	Main Finding
			prediction of low birth weight.
H.-Y. Chen, Chuang, Yang, & Wu (2011)	neural network and decision tree C5.0	number of birth, paternal smoking, hemorrhage, parity, age, paternal occupation, hypertension, medicines taken, gynecological diseases, height, weight, paternal age, paternal drinking, previous preterm birth, and vitamins taken.	Findings showed that multiple birth, hemorrhage during pregnancy, age, disease, previous preterm history, body weight before pregnancy and height of pregnant women, and paternal life style risk factors related to drinking and smoking are the important risk factors of preterm birth. Through the construction of a decision tree, 17 rules were explored to predict preterm birth. Ten of these rules, with an accuracy of 80%.
GOKSEN, EMINAGAOGLU & DOGAN (2011)	Data mining association rule	Patient Gender Male, Female - Day of Week - Date - Time - 2-Hour Period - Department Code - Patient Type Index- Diagnosis Code-Case Explanation	The findings demonstrated that Predictive algorithm derived 100 different association rules. The rule with the highest accuracy had a value of 0.99498 and the one with lowest accuracy was observed as 0.9733. Some few but strategic and meaningful. As on any day of the week, if the patient arrival time period is 22:00-00:00 at night and the patient is recorded as a non-standard emergency outpatient clinic and if the patient is also retired, then it could be a female patient.
Tanuja, Acharya, & Shailesh	Multilayer back propagation NN (MLP),		The results of this study show that the analysis of MLP has achieved better performance compared to the other

Study	Used Techniques	Main variables	Main Finding
(2011)	Naive Bayes Classifier, K-NN method, J48 decision tree.		three techniques. In this study, the accuracy of four data mining techniques MLP, Naives Bayes, K-NN, J48 are (87.8 %), (85.8%), (62.6%), (75.1%) respectively.
Bertsimas et al. (2008)	Data mining The Baseline Method, Classification Trees, Clustering	Demographic information such as age and gender based on medical and cost data from the first two years.	Clustering method results in better predictions for current high-cost bucket members and consistently better absolute prediction error. The members with overall costs between $12,300 and $16,000 in the last 12 months of the observation period and who have acute cost profiles. The members take no more than 14 different therapeutic drug classes during that period, and have not had a heart.

3.3 The Study Contribution

This topic of Data mining is still new in the Arab countries, and most of these studies took place in foreign countries .This study fills the gap in Data mining researches In Arabic world especially in the field of Healthcare. This is revealed in exploring the factors and challenges that affect the healthcare management resources. On other hand, predictive analytics as length of stay (LOS) can be predicted efficiently, the planning and management of hospital resources can be greatly enhanced.

In addition, there is an increased level of activity in the biomedical and health informatics world (e-prescribing, electronic health records, personal health records) that,

in the near future, will yield a wealth of available data that researchers can exploit meaningfully to strengthen knowledge building and evidence creation, and ultimately improve clinical and preventive care by using data mining techniques. This study encourage the health care organization using data mining technology to benefit from the large databases, and turning the massive data into meaningful information, to develop their work. Increase awareness about the importance of Data mining to understand the patient behaviors and delivering the offers and services to match their needs. It is also foster scientific discussion and disseminate new knowledge discover data mining techniques applied to healthcare management.

Chapter 4

BUSINESS UNDERSTANDING, PREPROCESSING AND MODEL BUILDING

Chapter 4 BUSINESS UNDERSTANDING, PREPROCESSING AND MODEL BUILDING

4.1 Data Collection

In this study, the decision tree method was performed on a real childbirth dataset. However, this data were collected from three government maternity hospitals of Gaza Strip, from January, 2013 to December, 2013.

This dataset contains 22507 records with 48 attributes which include a single noisy, missing, and inconsistent attribute values are preprocessed to handle the this problem from the analysis. The results are thus based on 22,461 records and illustrate the information about the childbirth situation and characteristics in the following sections.

4.2 Business Understanding

Governmental hospitals in Gaza Strip are forced to use their scarce resources as efficient as possible One of these resources is the hospital bed capacity. An optimal admission planning uses hospital bed capacity as efficient as possible. In other words Predicting the length of stay (LOS) of patients in a hospital is important in providing them with better services and higher satisfaction, as well as helping the hospital management plan and managing hospital resources as meticulously as possible. The aim of this study is apply data mining techniques to extract useful knowledge that could help in decision-making and build an accurate model to predict the LOS of maternity hospitals.

The application of maternity hospitals is part of full computerized hospitals system, so admission and discharge of the hospital will follow with the addition of some medical data of this section.

The system consists of three phases:

- Open a new medical record to give the mother's medical file number to save the medical history and administrative information
- The data of admission, it includes information about the case of the mother when entering
- The data of discharge, it includes many variables and medical indicators of the state of mother and child.

Appendix B shows the screens of the system and medical form for all above phases.

4.3 Data Understanding

4.3.1 Description of the Process of Accessing the Dataset

The source of data about childbirth was obtained from internal application In Ministry of Health maternity hospitals (Al Shifaa -Al Aqsa Martyrs –Nasser Medical Complex – Al-Helal Al-Emarati). As part of administrative tasks, hospital has staff responsible for recording, administrating and maintaining the database. Each hospital has its own Oracle database, but in the same structure. The database has attributes designed to store information on obstetric and medical history, and information on enter delivery date and discharge delivery date. The datasets in the oracle database are exported to four excel files. Data found in electronic format more than 90,000 records from 2012 to 2015, therefore researcher depends on three hospitals data from the period 01/01/2013 to 01/12/2013 to build LOS predictive model and then estimate the cost for this model.

Removing those identifying variables such as name of mothers, identification number, and name of the doctors, was done by the database administrator for the purpose of protecting privacy. The number of records before preprocessing is 22507

Finally, access was obtained to analyze the dataset for the objectives specified in the thesis. The childbirth database contains the following attributes listed in table 4.1

Table 4.1: Attributes listed

No	Attribute name	Description	Values	Data Type
1	AGE	Age of mother at birth of the child	Mothers age in 5-year intervals(15-19, 20-24, 25-29, 30-34, 35-39, 40-(44, 45-49	Nominal
2	R_NAME_AR	Region Name	Region Name	nominal
3	C_NAME_AR	City Name	City Name	nominal
4	ADMISSION_ENTER_DATE	Date of admission	Date and time admission	Date/Time
5	ADMISSION_OUT_DATE	Date of discharge	Date and time discharge	Date/Time
6	ADMISSION_DELIVERY_DATE	Date of delivery	Date and time of delivery	Date/Time
7	OC_NAME_AR	Type of discharge	under request, Good, ran away, referred to another section, outpatient clinics, referred to hospital in Gaza Strip, dead, referred to hospital outside Gaza Strip	Nominal
8	DELIVERY_NAME_AR	Type of delivery	Vaginal breech, vacuum extractor, Spontanuos Vaginal, Emergency CS, Elective CS, Forceps, Normal /emergency	Nominal
9	PRE_RISK_FACTOR	Risk Status when admission	Unknown, low, high	Nominal
10	POST_RISK_FACTOR	Risk Status after admission	Unknown, low, high	Nominal
11	REASON_NAME_AR	Reason of admission	labor pain, Vaginal bleeding, water pocket explosion, Medical disease associated with pregnancy, gynecology, planned caesarean section, planned abortion, other	Nominal
12	ADMISSION_TW	If TWINS	0 or 1 (Yes or No)	Nominal

No	Attribute name	Description	Values	Data Type
	INS			
13	BI_WEIGHT_GM	WEIGHT of baby	Weight of baby (grams)	Numeric
14	BOC_NAME_AR	Status of baby	live, still birth antepar, still birth intrapar, dead after birth	Nominal
15	PRE_NAME_AR	Delivery position	cephalic, breech, brow, face, compound, shoulder	Nominal
16	BORN_EXAM	Exam of baby	Yes or No	Nominal
17	EXM_NAME_AR	Result of exam if it Yes	Normal, Abnormal	Nominal
18	CA_NAME_AR	Congenital Anomalies	Hydrops fetalis due to hemolytic disease, Ancephaly and similar malformation, Encephacele, Microcephaly, Congenital hydrocephalu Spina bifida, , Malformation of female genitalia, Congenital malformation of hip, Down's Syndrome, None, others	Nominal
19	BI_APAGAR_1	Evaluation of apgar sum in first minute	From 1- 10	Numeric
20	BI_APAGAR_5	Evaluation of apgar sum in 5th minute	From 1- 10	Numeric
21	PAIN_RELIEF_NAME_EN	The Name of RELIEF	Pethedine, Tramal, Panadol, Phenergan, Scobotel, Non pharmaceutical pain relief, None, Others	Nominal
22	ICD_CD	International Statistical Classification of Diseases	International Statistical Classification of Diseases	Nominal
23	ICD_NAME_EN	The description of ICD_CD	The description of ICD_CD	Nominal
24	ADMISSION_GESTATIONAL_WEEKS	Number of gestational weeks	Number of gestational weeks	Numeric
25	NICU	admission of neonatal intensive	Yes/No	Nominal

No	Attribute name	Description	Values	Data Type
		care unit		
26	PARTOGRAM	PARTOGRAM	Yes/NO	Numeric
27	GENERATOR_NAME_AR	Type of GENERATOR	doctor, midwife, resisdent doctor	Nominal
28	STATUS_NAME_AR	Status of Placenta	complete, Incomplete	Nominal
29	BLOOD_TRANS	If the transfer blood	Yes/no	Nominal
30	MOTHER_EXAM	Mother Exam before discharge	Yes/no	Nominal
31	MOTHER_RESULT	The of result of exam	Normal, Abnormal	Nominal
32	CATALYST_NAME_EN	Name of catalyst	Spontaneou, Augumentation by ARM, Augmentation by oxytocin, Induction by Oxytocin, Induction by prostaglandin E2, none, Induction by prostaglandin E1 More, Augmentation by , ARM and oxytocin	Nominal
33	ADMISSION_MOTHER _EXAM_COUNT	Count of exam	Number	Numeric
34	PT_NAME_AR	Payment Type	insurance, cash payment	Nominal

4.4 Data Preparation and Preprocessing

Data preprocessing is a stage where the data set is prepared to be useful for data mining purposes. The following steps are undertaken in the preprocessing stage; data cleaning, attribute and feature selection, and data transformation. The overriding objective is to predict LOS.

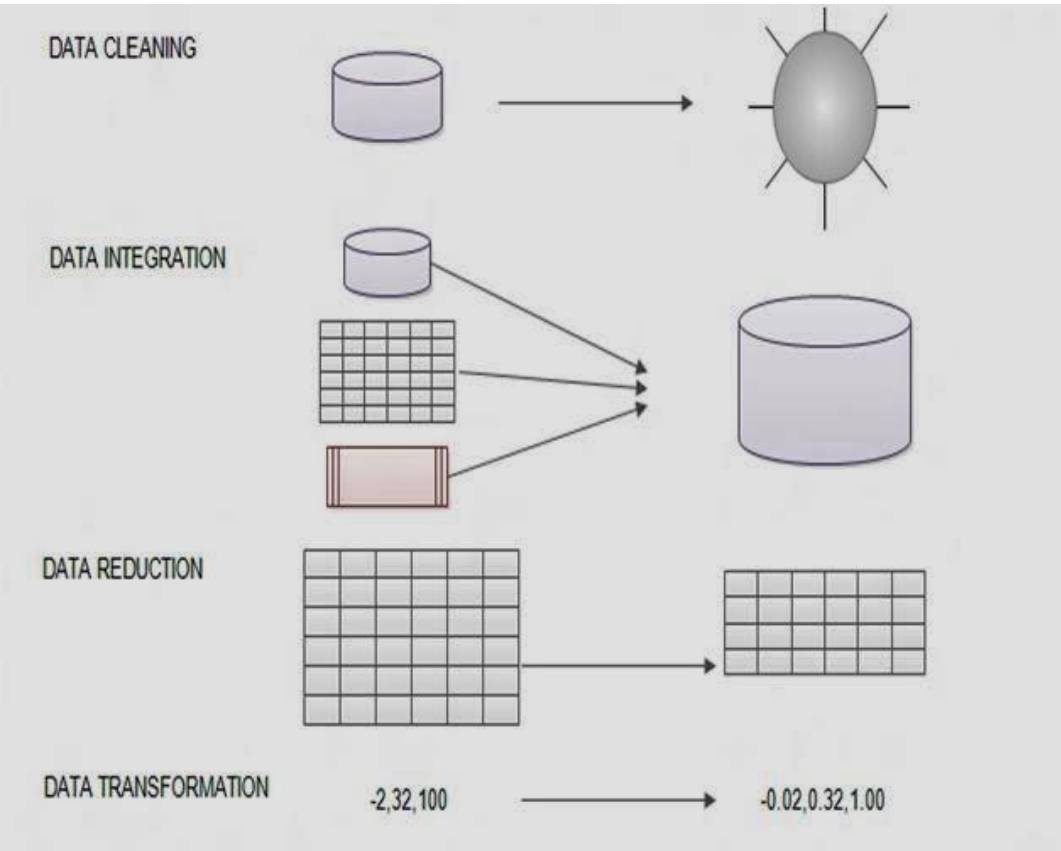


Figure: 4.1 Data Preparation and Preprocessing

4.4.1 Data Selection

The authors further stated that some algorithms may be confused by irrelevant or nosily attributes and construct poor classifiers. Therefore, eliminating some attributes which are assumed to be irrelevant to build the model can increase the accuracy of the classifier, save the computational time, and simplify results obtained. Some of the data or attributes in the initial dataset was not pertinent to the data mining goal and were ignored.

4.4.1.1 Data Attribute 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 (Two Crows, 2005). Therefore, the literatures consulted and communications with domain experts has given the researcher the knowledge of attributes and significant factors that affect Length of Stay. 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. Before this feature selection algorithm is applied on the dataset all the useless attributes are removed. According to output of the feature selection algorithm chisquer used, the attributes such as: OC_NAME_AR, DELIVERY_NAME_AR, POST_RISK_FACTOR, REASON_NAME_AR, ADMISSION_TWINS, BI_WEIGHT_GM, BOC_NAME_AR, PRE_NAME_AR, BORN_EXAM, and EXM_NAME_AR 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 as shown in Figure 4.2

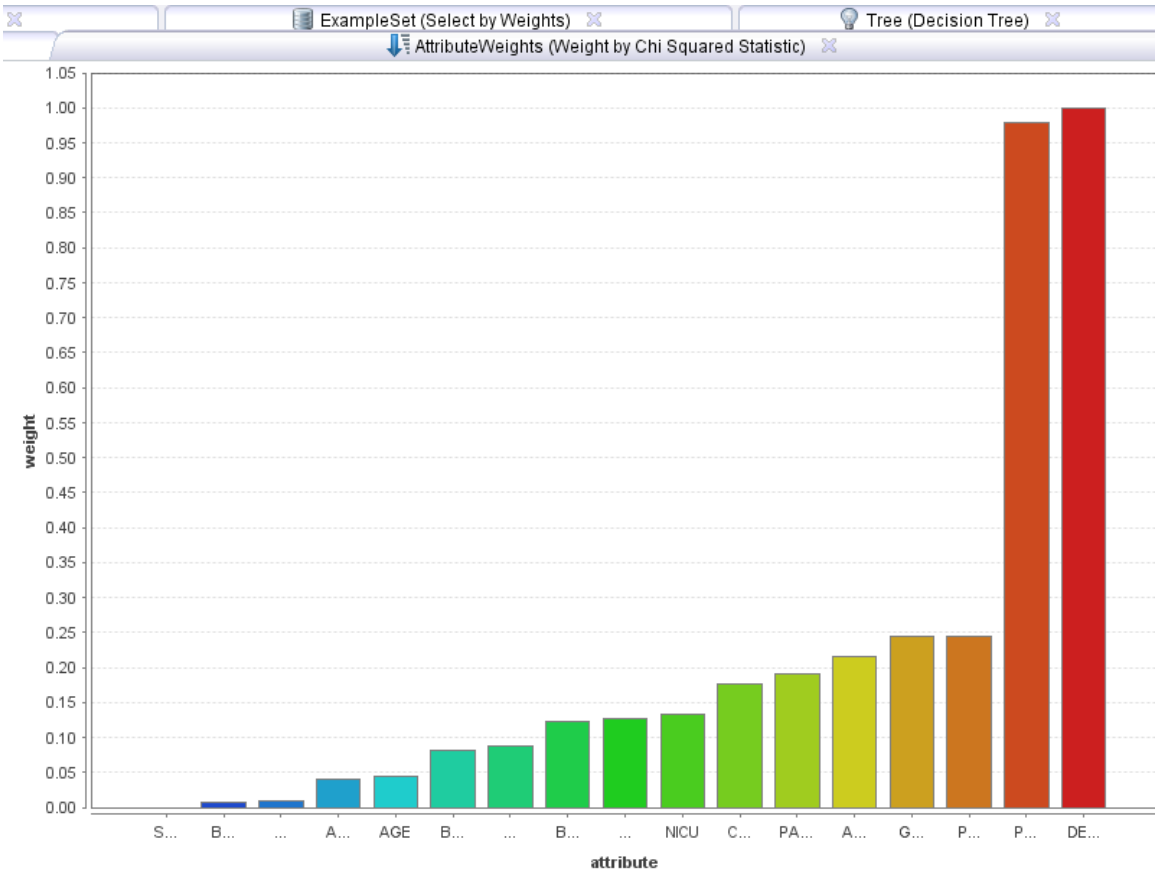


Figure 4.2 Attribute weights by chi squared statistic

4.4.1.2 Statistical Summary of the Attributes (features)

Here the selected attributes used for model building are statistically described in details. This statistical summary of the attributes is helpful for understanding of the data set for experimentation.

Length of Stay:

Length of stay is important medical variables and is computed as number of hours from difference between admission date and discharge date. And it is output attribute and will predict. This attribute is categorized into three parts as shown below table 4.2, and figure 4.3. LOS1≤24, 24<LOS2≤72, LOS3>72

Table 4.2: Frequency of length of stay Attribute

Length of Stay: Nominal		
Distinct Values	**Frequency**	**Percent (%)**
LOS1(≤24)	20095	89.47
LOS2 (≤72)	1551	6.90
LOS3 (>72)	815	3.63
Missing	0	0.0
Total	**22461**	**100**

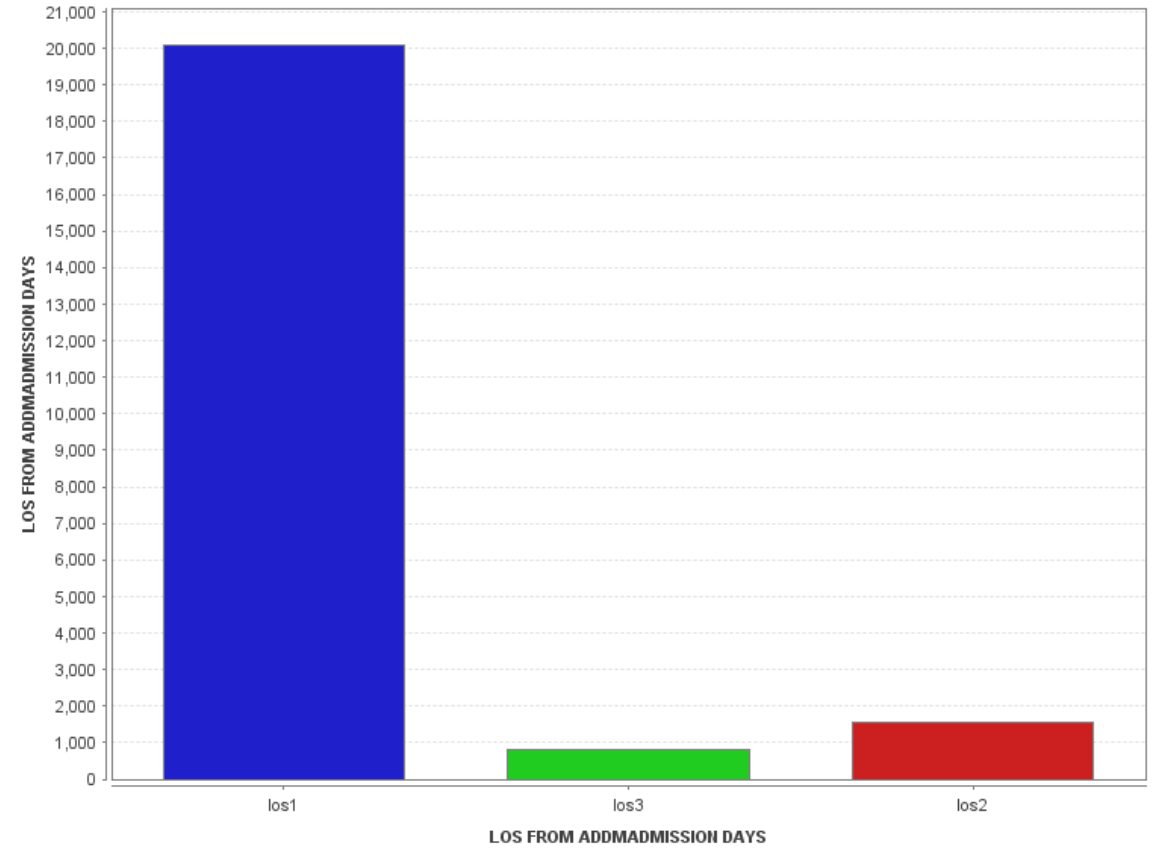


Figure 4.3 Frequency of Length of Stay

To summarize the relationship between two categorical variables, a cross-tabulation is used (also called a *contingency table*). A cross-tabulation (or *crosstab* for short) is a table that depicts the number of times each of the possible category combinations occurred in the sample data.

In the following, general descriptions of the fields and related work are presented.

Age:

Mother's age is important demographic variables and is the primary basis of demographic classification in vital statistics, censuses, and surveys. The age of mothers is classified by five year age groups. This attribute is categorized into seven parts figure 4.3.

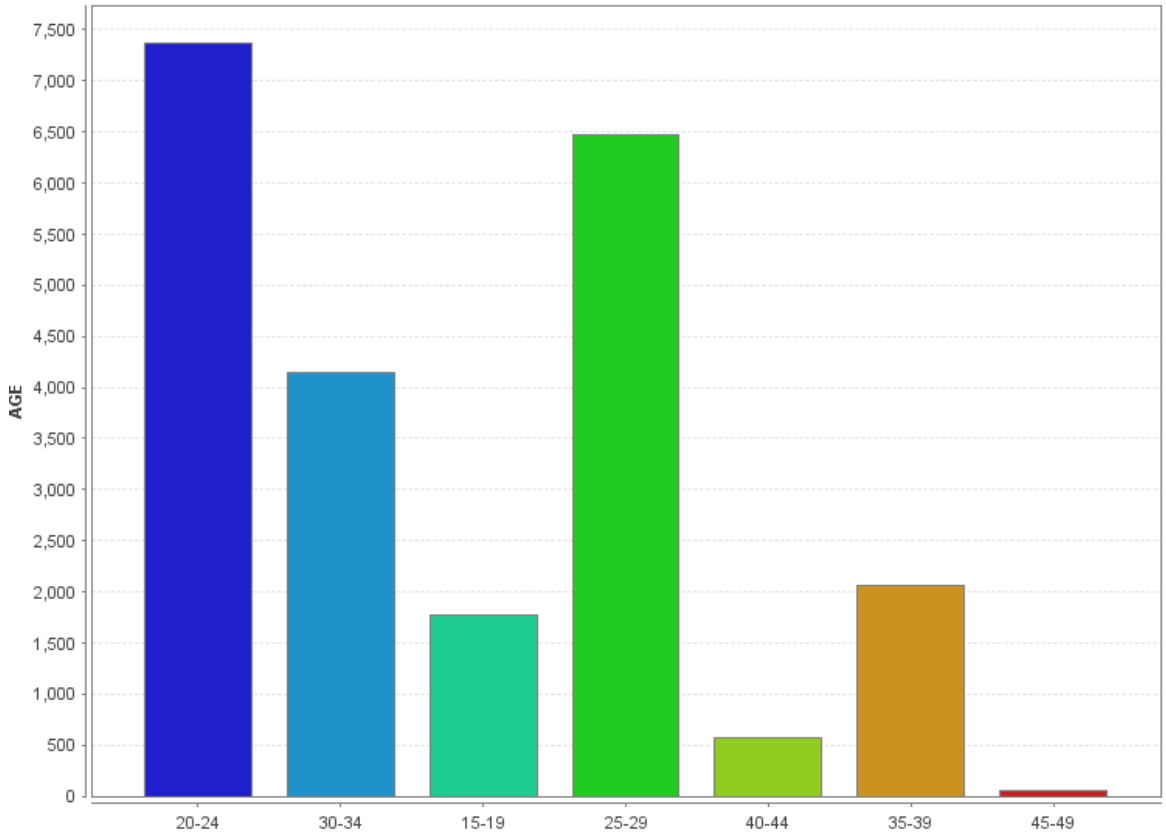


Figure 4.4 Frequency of Mother's Age Attribute

The results of cross tabulation between age and LOS showed that total percentage of women 15-19 age in the LOS2 amounted to 24.3%, while the percentage of women at 45-49 age 49.1%. Therefore increasing in age will increase the LOS, the result of cross tabulation in the following table 4.3

Table 4.3: Cross tabulation of AGE & LOS

AGE_CAT * LOS_CAT Cross tabulation						
			LOS_CAT			Total
			LOS1	LOS2	LOS3	
AGE CAT	15-19	Count	1280	430	58	1768
		% within AGE_CAT	72.4%	24.3%	3.3%	100.0%
		% within LOS_CAT	8.7%	6.8%	4.2%	7.9%
		% of Total	5.7%	1.9%	0.3%	7.9%
	20-24	Count	5178	1878	318	7374
		% within AGE_CAT	70.2%	25.5%	4.3%	100.0%
		% within LOS_CAT	35.2%	29.6%	23.0%	32.8%
		% of Total	23.1%	8.4%	1.4%	32.8%
	25-29	Count	4303	1802	373	6478
		% within AGE_CAT	66.4%	27.8%	5.8%	100.0%
		% within LOS_CAT	29.2%	28.4%	27.0%	28.8%
		% of Total	19.2%	8.0%	1.7%	28.8%
	30-34	Count	2551	1271	329	4151
		% within AGE_CAT	61.5%	30.6%	7.9%	100.0%
		% within LOS_CAT	17.3%	20.0%	23.8%	18.5%
		% of Total	11.4%	5.7%	1.5%	18.5%

	35-39	Count	1124	724	211	2059
		% within AGE_CAT	54.6%	35.2%	10.2%	100.0%
		% within AGE_CAT	7.6%	11.4%	15.3%	9.2%
		% of Total	5.0%	3.2%	0.9%	9.2%
	40-44	Count	276	220	82	578
		% within AGE_CAT	47.8%	38.1%	14.2%	100.0%
		% within LOS_CAT	1.9%	3.5%	5.9%	2.6%
		% of Total	1.2%	1.0%	0.4%	2.6%
	45-49	Count	16	26	11	53
		% within AGE_CAT	30.2%	49.1%	20.8%	100.0%
		% within LOS_CAT	0.1%	0.4%	0.8%	0.2%
		% of Total	0.1%	0.1%	0.0%	0.2%
Total		Count	14728	6351	1382	22461
		% within AGE_CAT	65.6%	28.3%	6.2%	100.0%
		% within LOS_CAT	100.0%	100.0%	100.0%	100.0%
		% of Total	65.6%	28.3%	6.2%	100.0%

OC_NAME_AR:

OC_NAME_AR is important medical variable that illustrates the status of discharge from the hospital as shown in table 4.4, and figure 4.5.

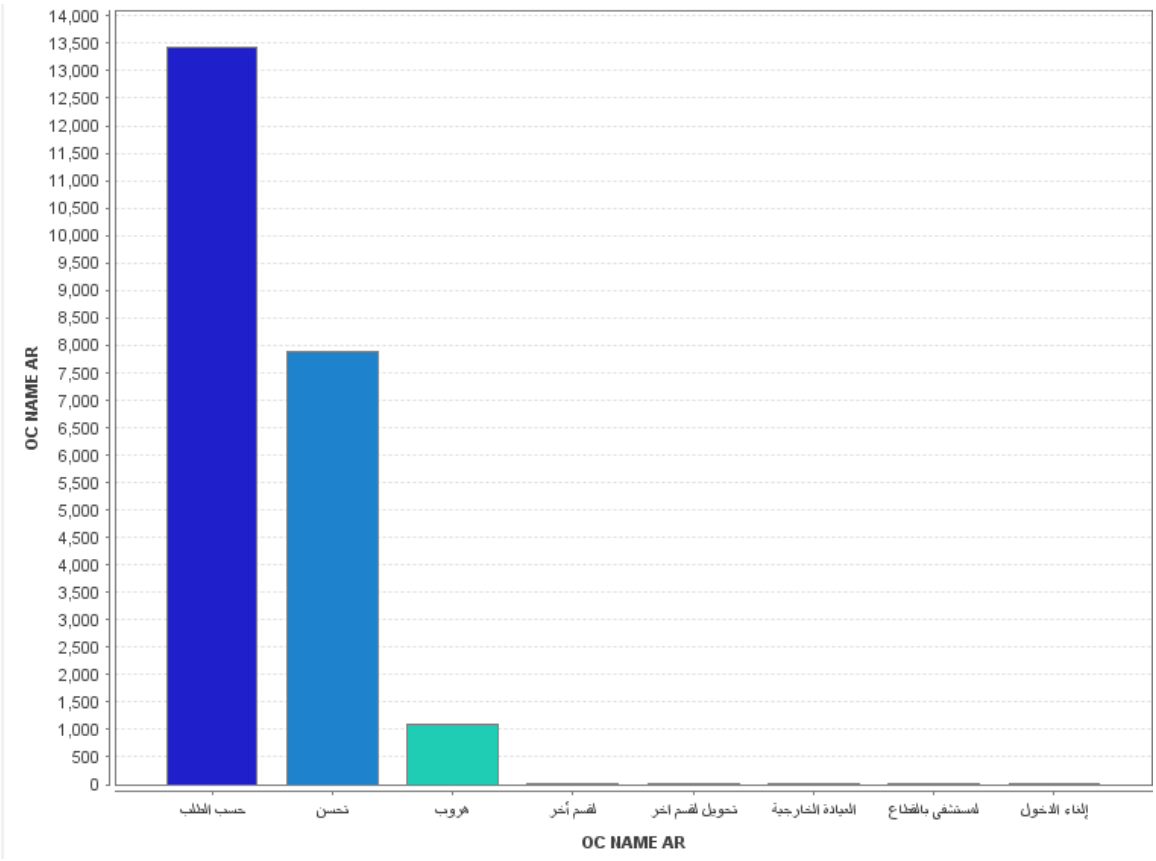


Figure 4.5 Summary of OC_NAME_AR Attribute

The highest frequency in under request discharge, this means that the LOS less than time required to stay in the hospital after birth and discharge of the hospital is under request and the responsibility of women, therefore is important to study the causes that lead to this situation and also study the cause of ran away situation. The following table 4.4 shows the relationship between discharge status and LOS. The Percentage of discharge in good status is distribution as the following 53.6% for LOS1, 34.9% LOS2, 11.5%LOS3.

Table 4.4: Cross tabulation of status of discharge & LOS

OC_NAME_AR * LOS_CAT Cross tabulation						
			LOS_CAT			Total
			LOS1	LOS2	LOS3	
OC_NAME_AR	Admission Canceled	Count	6	0	0	6
		% within OC_NAME_AR	100.0%	0.0%	0.0%	100.0%
		% within LOS_CAT	0.0%	0.0%	0.0%	0.0%
		% of Total	0.0%	0.0%	0.0%	0.0%
	outpatient clinincs	Count	0	3	5	8
		% within OC_NAME_AR	0.0%	37.5%	62.5%	100.0%
		% within LOS_CAT	0.0%	0.0%	0.4%	0.0%
		% of Total	0.0%	0.0%	0.0%	0.0%
	Good	Count	4223	2751	907	7881
		% within OC_NAME_AR	53.6%	34.9%	11.5%	100.0%
		% within LOS_CAT	28.7%	43.3%	65.6%	35.1%
		% of Total	18.8%	12.2%	4.0%	35.1%
	referred to another section	Count	3	0	1	4
		% within OC_NAME_AR	75.0%	0.0%	25.0%	100.0%
		% within LOS_CAT	0.0%	0.0%	0.1%	0.0%
		% of Total	0.0%	0.0%	0.0%	0.0%
	under request	Count	9542	3446	447	13435
		% within OC_NAME_AR	71.0%	25.6%	3.3%	100.0%
		% within LOS_CAT	64.8%	54.3%	32.3%	59.8%
		% of Total	42.5%	15.3%	2.0%	59.8%

OC_NAME_AR * LOS_CAT Cross tabulation						
			LOS_CAT			Total
			LOS1	LOS2	LOS3	
	to another section	Count	0	9	4	13
		% within OC_NAME_AR	0.0%	69.2%	30.8%	100.0%
		% within LOS_CAT	0.0%	0.1%	0.3%	0.1%
		% of Total	0.0%	0.0%	0.0%	0.1%
	referred to hospital out side gazastrip	Count	0	2	6	8
		% within OC_NAME_AR	0.0%	25.0%	75.0%	100.0%
		% within LOS_CAT	0.0%	0.0%	0.4%	0.0%
		% of Total	0.0%	0.0%	0.0%	0.0%
	ran away	Count	954	140	12	1106
		% within OC_NAME_AR	86.3%	12.7%	1.1%	100.0%
		% within LOS_CAT	6.5%	2.2%	0.9%	4.9%
		% of Total	4.2%	0.6%	0.1%	4.9%
Total		Count	14728	6351	1382	22461
		% within OC_NAME_AR	65.6%	28.3%	6.2%	100.0%
		% within LOS_CAT	100.0%	100.0%	100.0%	100.0%
		% of Total	65.6%	28.3%	6.2%	100.0%

DELIVERY_NAME_AR:

Most babies are born in a vaginal delivery. But in some cases, other types of delivery occur by choice or because of an emergency; therefore this attribute illustrates the type of delivery as shown in table 4.5, and figure 4.6.

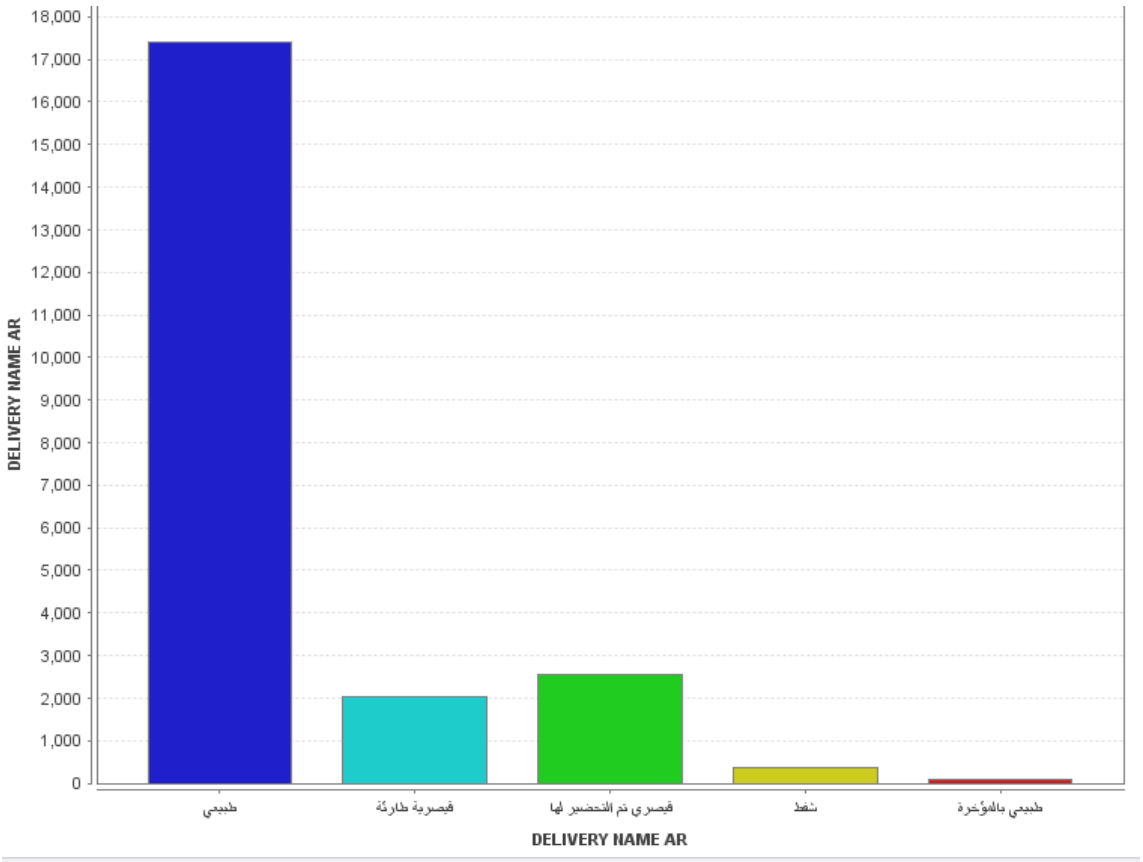


Figure 4.6 Frequency of DELIVERY_NAME_AR Attribute

The results of cross tabulation between delivery type and LOS showed that the LOS is affected by the type of birth, where there is an increase in LOS in cases of caesarean section. The elective caesarean section is planned operation and the admission is before one day to prepare for the operation, while the emergency caesarean as a result of a complex situation for the mother. In addition the results show that 66% of elective caesarean section LOS is between 24 - 72 hours (LOS2). As shown in table 4.5

Table 4.5: Cross tabulation of DELIVERY_NAME_AR & LOS

DELIVERY_NAME * LOS_CAT Cross tabulation						
			LOS_CAT			
			LOS1	LOS2	LOS3	Total
DELIVERY_NAME	Vacum extractor	Count	246	95	8	349
		% within DELIVERY_NAME	70.5%	27.2%	2.3%	100.0%
		% within LOS_CAT	1.7%	1.5%	0.6%	1.6%
		% of Total	1.1%	0.4%	0.0%	1.6%
	Spontanuos Vaginal	Count	14117	3048	258	17423
		% within DELIVERY_NAME	81.0%	17.5%	1.5%	100.0%
		% within LOS_CAT	95.9%	48.0%	18.7%	77.6%
		% of Total	62.9%	13.6%	1.1%	77.6%
	Vaginal breech	Count	75	9	2	86
		% within DELIVERY_NAME	87.2%	10.5%	2.3%	100.0%
		% within LOS_CAT	0.5%	0.1%	0.1%	0.4%
		% of Total	0.3%	0.0%	0.0%	0.4%
	Elective CS	Count	66	1713	787	2566
		% within DELIVERY_NAME	2.6%	66.8%	30.7%	100.0%
		% within LOS_CAT	0.4%	27.0%	56.9%	11.4%
		% of Total	0.3%	7.6%	3.5%	11.4%
	Emergency CS	Count	224	1486	327	2037
		% within DELIVERY_NAME	11.0%	73.0%	16.1%	100.0%
		% within LOS_CAT	1.5%	23.4%	23.7%	9.1%
		% of Total	1.0%	6.6%	1.5%	9.1%
Total		Count	14728	6351	1382	22461
		% within DELIVERY_NAME	65.6%	28.3%	6.2%	100.0%
		% within LOS_CAT	100.0%	100.0%	100.0%	100.0%
		% of Total	65.6%	28.3%	6.2%	100.0%

PRE_RISK_FACTOR:

This factor indicates to risk status for pregnant when admission to hospital. A pregnancy is considered high-risk when there are potential complications that could affect the mother, the baby, or both. High-risk pregnancies require management by a specialist to help ensure the best outcome for the mother and baby as shown in table 4.6, and figure 4.7.

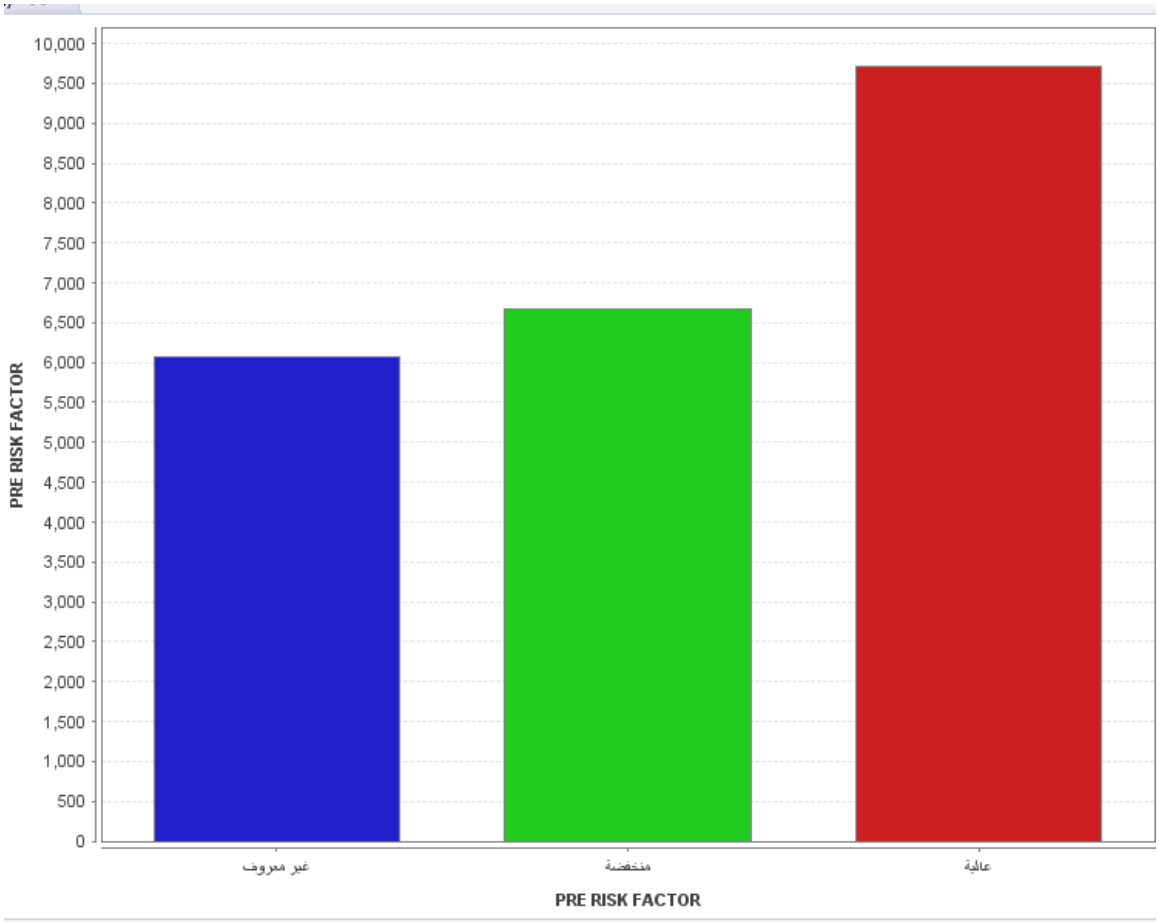


Figure 4.7 Frequency of PRE_RISK_FACTOR Attribute

The results of cross tabulation between delivery type and LOS showed that the LOS is affected by the risk factor, where there is an increase in LOS in cases of high risk factor. The results show that 35.3% of high risk factor, LOS is between 24 - 72 hours (LOS2), and 8.2% LOS 2 in low risk factor, on other hand the percentage of frequency in

unknown cases is 27 % therefore is sensational to determine the risk status since LOS affected by risk factor.

Table 4.6: Cross tabulation of PRE_RISK_FACTOR & LOS

PRE_RISK_FACTOR * LOS_CAT Cross tabulation						
			LOS_CAT			
			LOS1	LOS2	LOS3	Total
PRE_RISK_FACTOR	high	Count	5424	3428	866	9718
		% within PRE_RISK_FACTOR	55.8%	35.3%	8.9%	100.0%
		% within LOS_CAT	36.8%	54.0%	62.7%	43.3%
		% of Total	24.1%	15.3%	3.9%	43.3%
	unknown	Count	3343	2373	352	6068
		% within PRE_RISK_FACTOR	55.1%	39.1%	5.8%	100.0%
		% within LOS_CAT	22.7%	37.4%	25.5%	27.0%
		% of Total	14.9%	10.6%	1.6%	27.0%
	low	Count	5961	550	164	6675
		% within PRE_RISK_FACTOR	89.3%	8.2%	2.5%	100.0%
		% within LOS_CAT	40.5%	8.7%	11.9%	29.7%
		% of Total	26.5%	2.4%	0.7%	29.7%
Total		Count	14728	6351	1382	22461
		% within PRE_RISK_FACTOR	65.6%	28.3%	6.2%	100.0%
		% within LOS_CAT	100.0%	100.0%	100.0%	100.0%
		% of Total	65.6%	28.3%	6.2%	100.0%

ADMISSION_TWINS:

This feature indicates if one or more fetus in a single pregnancy.

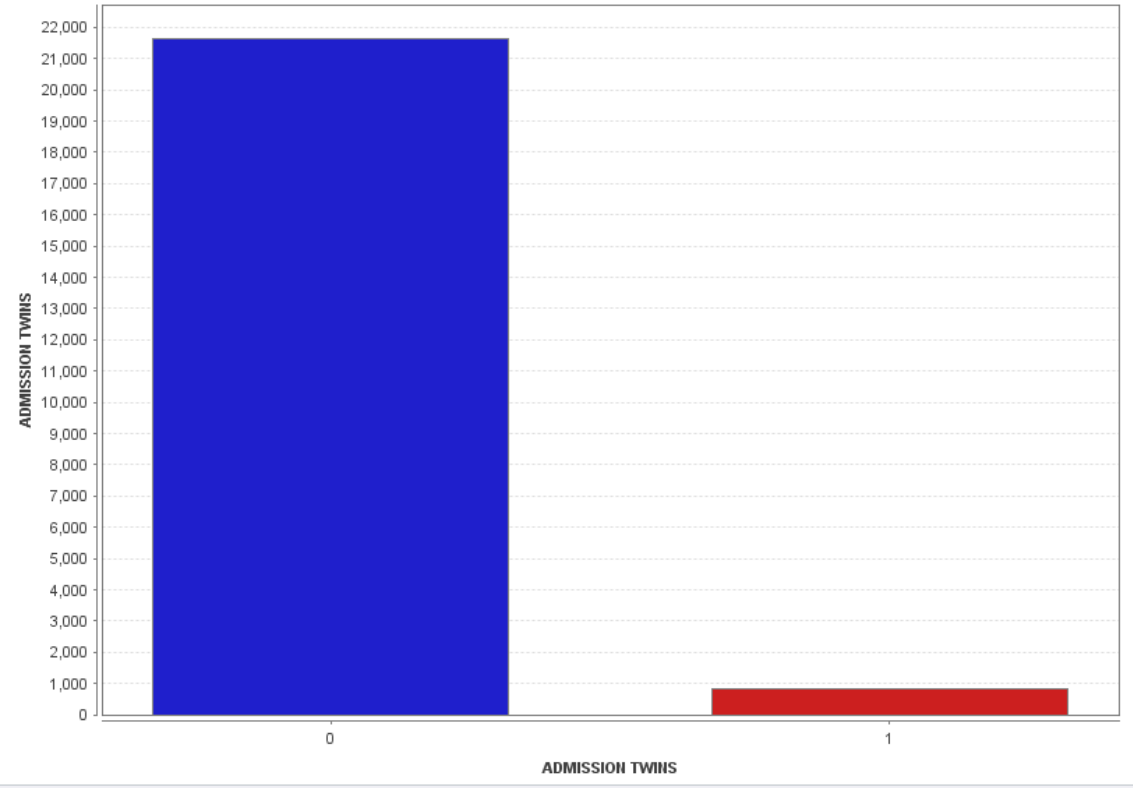


Figure 4.8 Frequency of ADMISSION_TWINS Attribute

The results of cross tabulation between delivery type and LOS showed that the LOS is affected by admission twins; where there is an increase in LOS in cases of have twins in single pregnancy. The results show that 50.2% of cases of have twins the LOS is between 24 - 72 hours (LOS2) and 16.1% the LOS > 72 hours. As shown in table 4.7

Table 4.7: Cross tabulation of ADMISSION_TWINS & LOS

ADMISSION_TWINS * LOS_CAT Cross tabulation						
			LOS_CAT			Total
			LOS1	LOS2	LOS3	
ADMISSION_TWINS	NO	Count	14453	5942	1251	21646
		% within ADMISSION_TWINS	66.8%	27.5%	5.8%	100.0%
		% within LOS_CAT	98.1%	93.6%	90.5%	96.4%
		% of Total	64.3%	26.5%	5.6%	96.4%

ADMISSION_TWINS * LOS_CAT Cross tabulation						
			LOS_CAT			
			LOS1	LOS2	LOS3	Total
	YES	Count	275	409	131	815
		% within ADMISSION_TWINS	33.7%	50.2%	16.1%	100.0%
		% within LOS_CAT	1.9%	6.4%	9.5%	3.6%
		% of Total	1.2%	1.8%	0.6%	3.6%
Total		Count	14728	6351	1382	22461
		% within ADMISSION_TWINS	65.6%	28.3%	6.2%	100.0%
		% within LOS_CAT	100.0%	100.0%	100.0%	100.0%
		% of Total	65.6%	28.3%	6.2%	100.0%

BI WEIGHT GM:

This attribute indicates newborn baby birth weight in grams. At full term, the average baby will be about 20 inches (51 cm) long and will weigh approximately 6 to 9 pounds (2700 to 4000 grams).

Average: 3262.59

Deviations: 589.28

Max: 0

Min: 8700.0

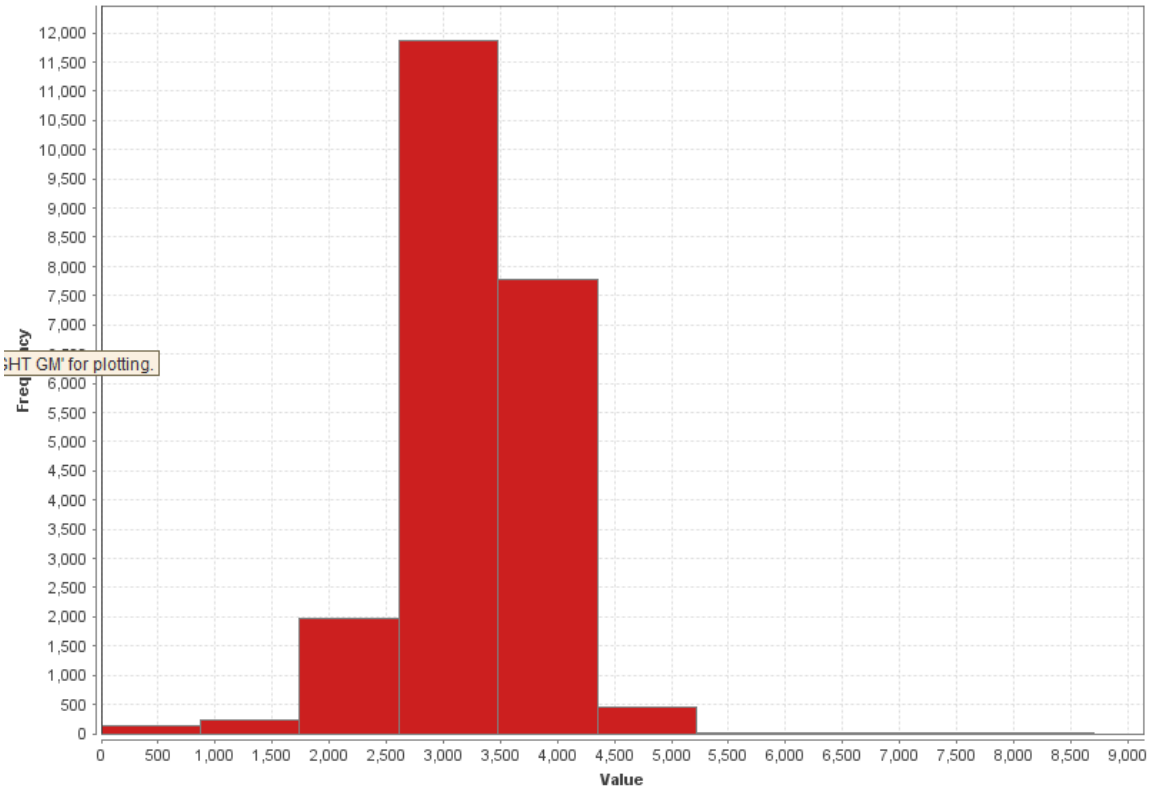


Figure 4.9 Summary of ADMISSION_TWINS Attribute

BOC_NAME_AR:

This feature indicates status of baby in childbirth. As shown in table 4.8, and figure 4.10.

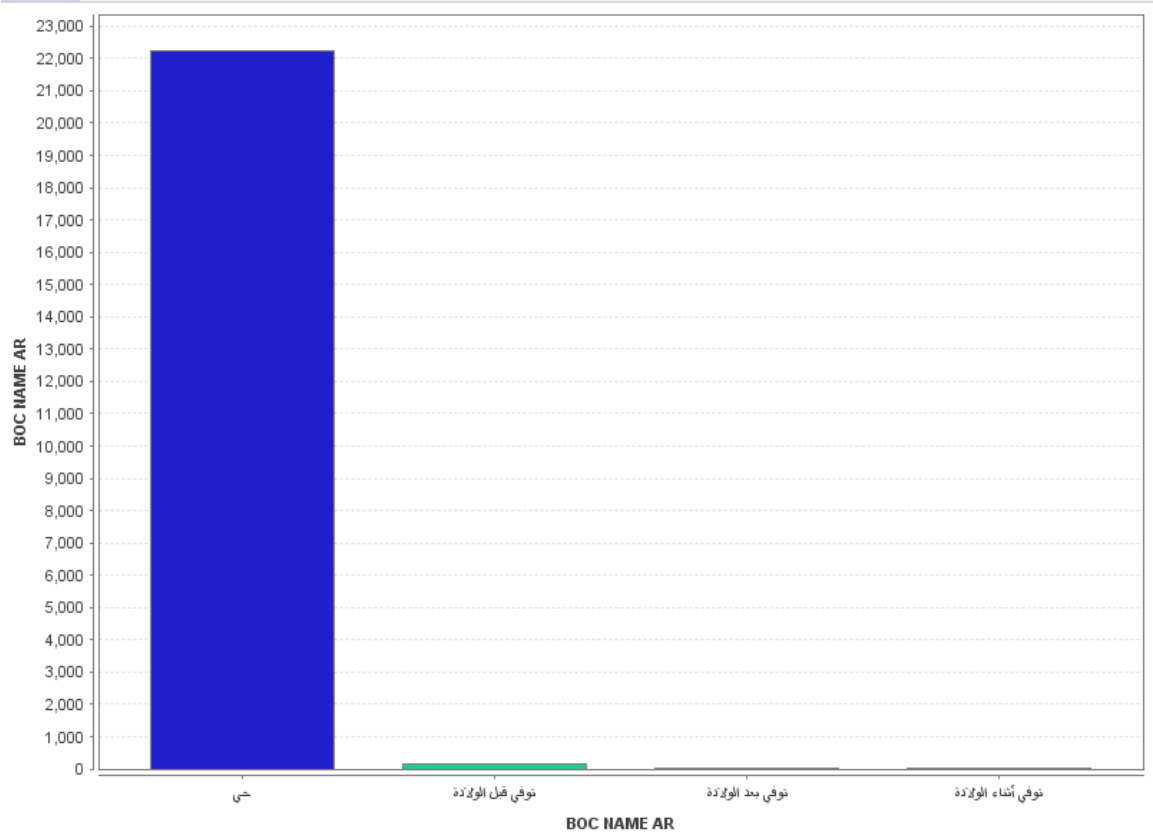


Figure 4.10 Frequency of BOC_NAME_AR Attribute

The results of cross tabulation between status of baby and LOS showed that the LOS is affected by admission twins; where there is an increase in LOS in cases of have twins in single pregnancy. The results show that the highest percentage of LOS is in dead after birth cases as 51.3% in LOS2 (24-72) hours and 28.2% in LOS3. As shown in table 4.8

Table 4.8: Cross tabulation of BOC_NAME_AR & LOS

BOC_NAME_AR * LOS_CAT Cross tabulation						
			LOS_CAT			
			LOS1	LOS2	LOS3	Total
BOC_NAME_AR	stillbirth intrapartum	Count	2	1	0	3
		% within BOC_NAME_AR	66.7%	33.3%	0.0%	100.0%
		% within LOS_CAT	0.0%	0.0%	0.0%	0.0%
		% of Total	0.0%	0.0%	0.0%	0.0%
	dead after	Count	8	20	11	39

	birth	% within BOC_NAME_AR	20.5%	51.3%	28.2%	100.0%
		% within LOS_CAT	0.1%	0.3%	0.8%	0.2%
		% of Total	0.0%	0.1%	0.0%	0.2%
	stillbirth antepartum	Count	87	59	21	167
		% within BOC_NAME_AR	52.1%	35.3%	12.6%	100.0%
		% within LOS_CAT	0.6%	0.9%	1.5%	0.7%
		% of Total	0.4%	0.3%	0.1%	0.7%
	Live	Count	14631	6271	1350	22252
		% within BOC_NAME_AR	65.8%	28.2%	6.1%	100.0%
		% within LOS_CAT	99.3%	98.7%	97.7%	99.1%
		% of Total	65.1%	27.9%	6.0%	99.1%
Total		Count	14728	6351	1382	22461
		% within BOC_NAME_AR	65.6%	28.3%	6.2%	100.0%
		% within LOS_CAT	100.0%	100.0%	100.0%	100.0%
		% of Total	65.6%	28.3%	6.2%	100.0%

PRE_NAME_AR:

Most babies will move into delivery position a few weeks prior to birth, with the head moving closer to the birth canal. When this fails to happen, the baby's buttocks and/or feet will be positioned to be delivered first. PRE_NAME_AR refers to this position. As shown in table 4.9 and figure 4.11

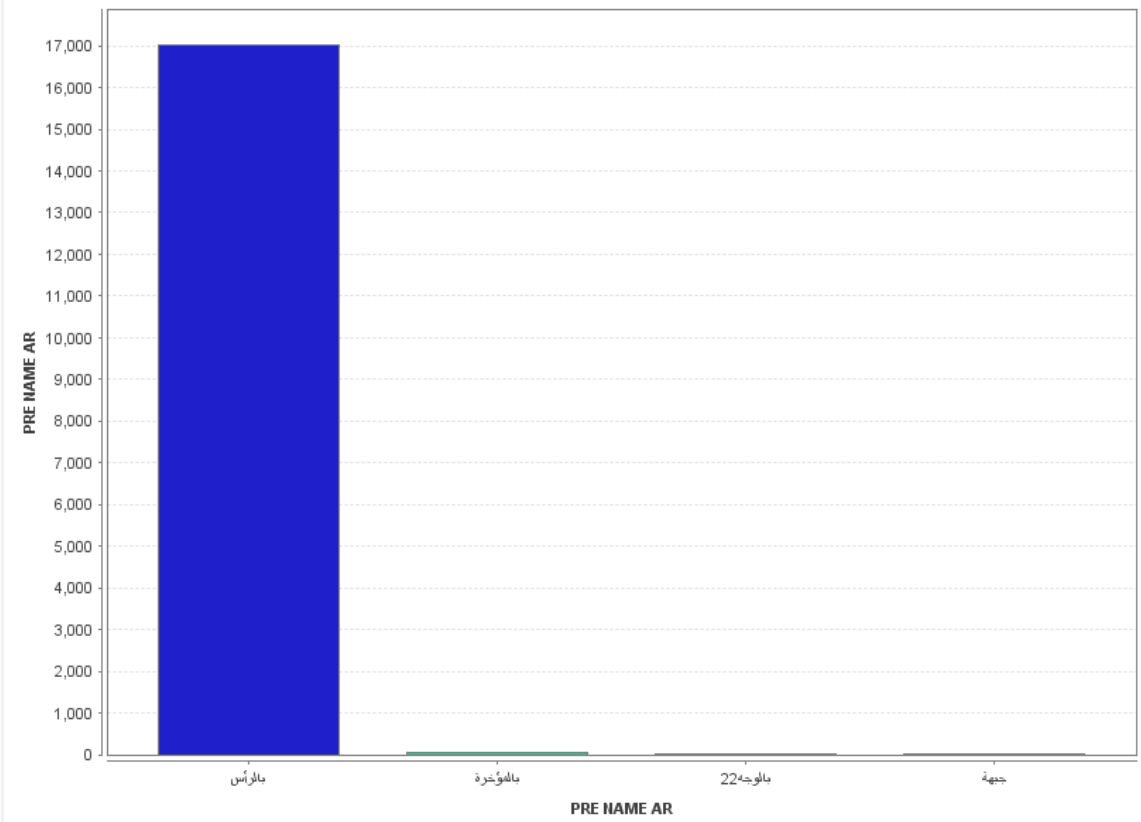


Figure 4.11 Frequency of PRE_NAME_AR Attribute

The results of cross tabulation between delivery position and LOS showed that the LOS is affected by delivery position; where there is an increase in LOS in cases of breech, face, cesren section in childbirth. The results show that the highest percentage of LOS is in cesren section cases as 69.5%% in LOS2 (24-72) hours and 24.2%% in LOS3 (>72) hours. As shown in table 4.9.

Table 4.9: Cross tabulation of PRE_NAME_AR & LOS

PRE_NAME_AR_RESULT * LOS_CAT Cross tabulation						
			LOS_CAT			
			LOS1	LOS2	LOS3	Total
PRE_NAME_AR_RESULT	cephalic	Count	14390	3144	265	17799
		% within PRE_NAME_AR_RESULT	80.8%	17.7%	1.5%	100.0%
		% within LOS_CAT	97.7%	49.5%	19.2%	79.2%
		% of Total	64.1%	14.0%	1.2%	79.2%

	breech	Count	46	7	3	56
		% within PRE_NAME_AR_RESULT	82.1%	12.5%	5.4%	100.0%
		% within LOS_CAT	0.3%	0.1%	0.2%	0.2%
		% of Total	0.2%	0.0%	0.0%	0.2%
	face	Count	1	1	0	2
		% within PRE_NAME_AR_RESULT	50.0%	50.0%	0.0%	100.0%
		% within LOS_CAT	0.0%	0.0%	0.0%	0.0%
		% of Total	0.0%	0.0%	0.0%	0.0%
	brow	Count	1	0	0	1
		% within PRE_NAME_AR_RESULT	100.0%	0.0%	0.0%	100.0%
		% within LOS_CAT	0.0%	0.0%	0.0%	0.0%
		% of Total	0.0%	0.0%	0.0%	0.0%
	Cesren section	Count	290	3199	1114	4603
		% within PRE_NAME_AR_RESULT	6.3%	69.5%	24.2%	100.0%
		% within LOS_CAT	2.0%	50.4%	80.6%	20.5%
		% of Total	1.3%	14.2%	5.0%	20.5%
Total		Count	14728	6351	1382	22461
		% within PRE_NAME_AR_RESULT	65.6%	28.3%	6.2%	100.0%
		% within LOS_CAT	100.0%	100.0%	100.0%	100.0%
		% of Total	65.6%	28.3%	6.2%	100.0%

BORN_EXAM_ RESULT:

This attribute explores result of examination for baby. The baby should be examined briefly immediately after birth. This should be confined to quick assessment of respiration, circulation, temperature, neurological status, and screening for anomalies or disease that might mandate emergency treatment. As shown in table 4.9 and figure 12.4.

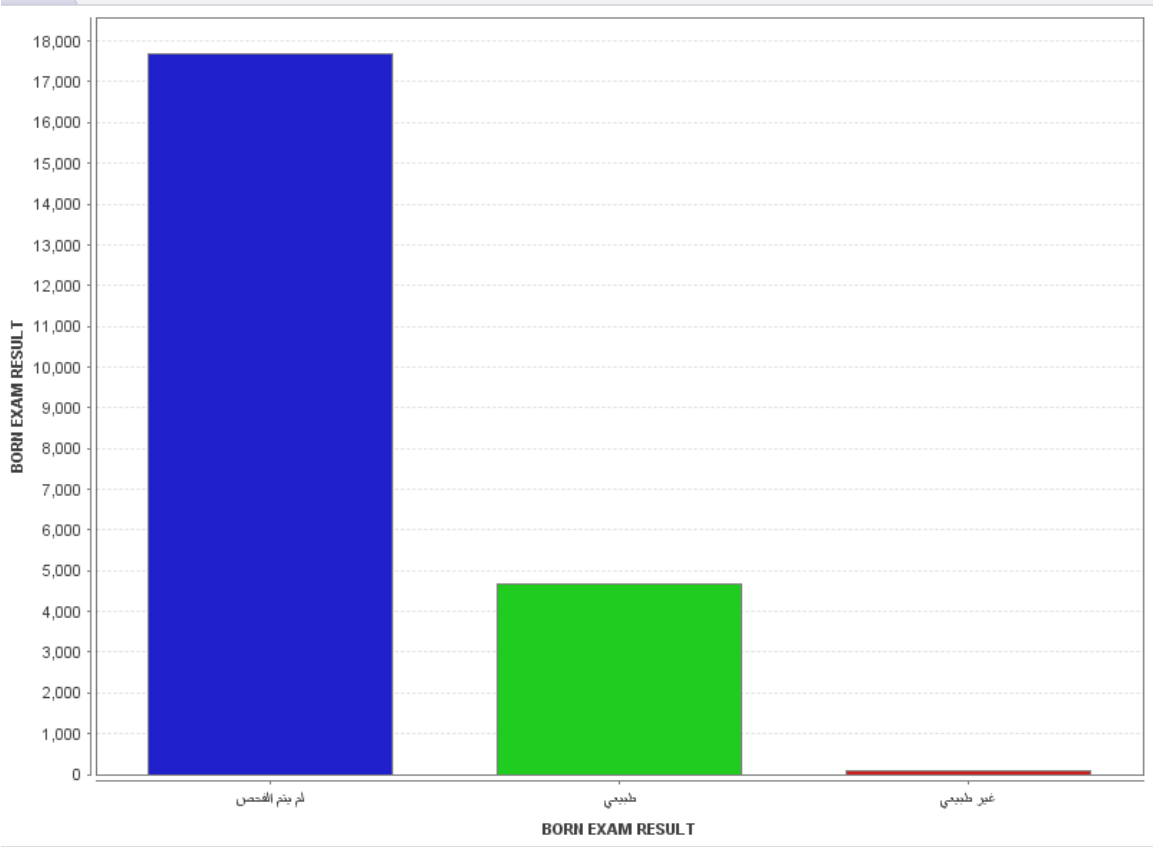


Figure 4.12 Summary of BORN_EXAM_RESULT Attribute

The results of cross tabulation between delivery position and LOS showed that the LOS is affected by delivery position; where there is an increase in LOS in cases of breech, face, caesarean section in childbirth. The results show that the percentage of LOS is in unknown cases as 71.0% in LOS1 (<24) hours and 24.7% in LOS2 (24 -72) hours. As shown in table 4.10.

Table 4.10: Cross tabulation of BORN_EXAM_RESULT & LOS

BORN_EXAM_RESULT * LOS_CAT Cross tabulation						
			LOS_CAT			Total
			LOS1	LOS2	LOS3	
BORN_EXAM_RESULT	Normal	Count	2123	1942	598	4663
		% within BORN_EXAM_RESULT	45.5%	41.6%	12.8%	100.0%
		% within LOS_CAT	14.4%	30.6%	43.3%	20.8%

		% of Total	9.5%	8.6%	2.7%	20.8%
	Abnormal	Count	33	42	22	97
		% within BORN_EXAM_RESULT	34.0%	43.3%	22.7%	100.0%
		% within LOS FROM ADDM BYHOURS CAT	0.2%	0.7%	1.6%	0.4%
		% of Total	0.1%	0.2%	0.1%	0.4%
	Unknown	Count	12572	4367	762	17701
		% within BORN_EXAM_RESULT	71.0%	24.7%	4.3%	100.0%
		% within LOS FROM ADDM BYHOURS CAT	85.4%	68.8%	55.1%	78.8%
		% of Total	56.0%	19.4%	3.4%	78.8%
Total		Count	14728	6351	1382	22461
		% within BORN_EXAM_RESULT	65.6%	28.3%	6.2%	100.0%
		% within LOS FROM ADDM BYHOURS CAT	100.0%	100.0%	100.0%	100.0%
		% of Total	65.6%	28.3%	6.2%	100.0%

PAIN_RELIEF_NAME_EN:

This attribute explores types of pain relief medications.

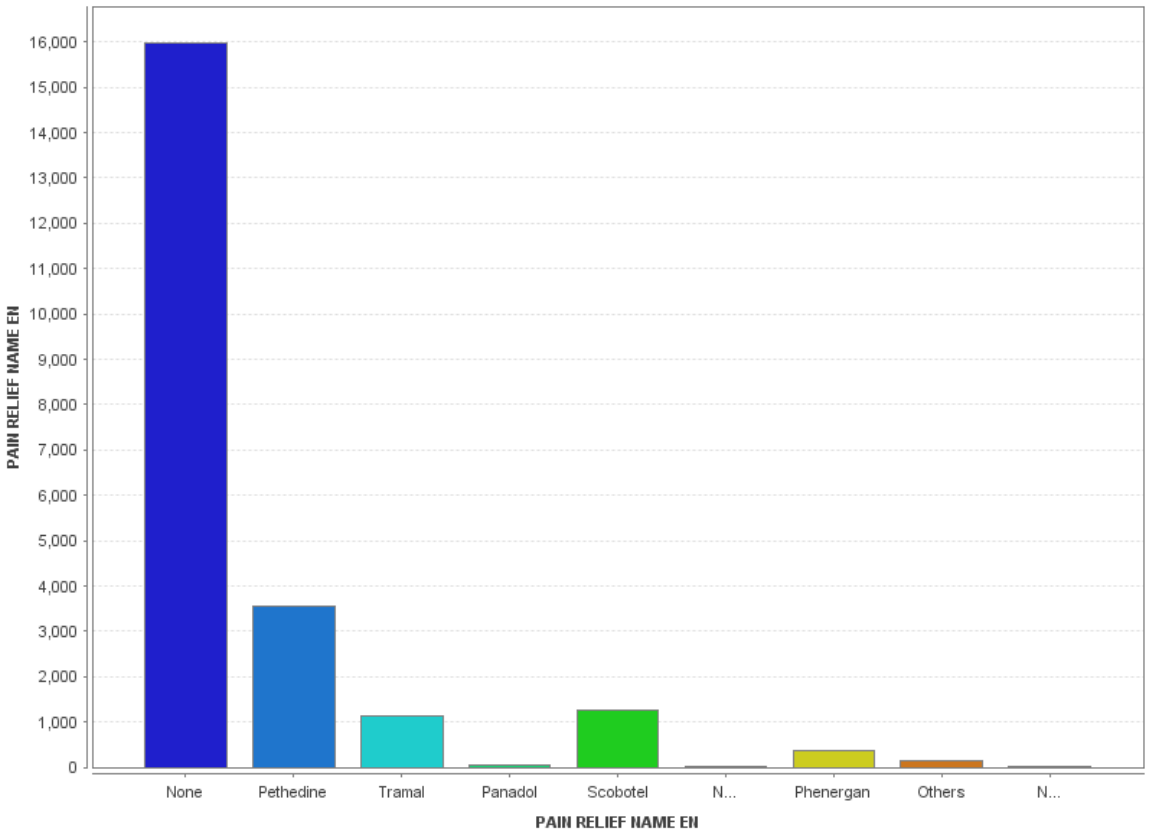


Figure 4.13 Summary of PAIN_RELIEF_NAME_EN Attribute

The results of cross tabulation between PAIN_RELIEF_NAME_EN and LOS showed that the LOS is affected by PAIN_RELIEF_NAME_EN; where there is an increase in LOS in cases of have pain relief section in childbirth. The results show that the percentage of LOS is in Tramal cases as 68.9% in LOS2 (24-72) hours and 16.7% in LOS3 (>72) hours. As shown in table 4.11.

Table 4.11: Cross tabulation of PAIN_RELIEF_NAME_EN &LOS

PAIN_RELIEF_NAME_EN * LOS_CAT Cross tabulation						
			LOS_CAT			Total
			LOS1	LOS2	LOS3	
PAIN_RELIEF_NAME_EN	Non pharmaceutical pain relief	Count	5	5	0	10
		% within PAIN_RELIEF_NAME_EN	50.0%	50.0%	0.0%	100.0%
		% within LOS_CAT	0.0%	0.1%	0.0%	0.0%
		% of Total	0.0%	0.0%	0.0%	0.0%
	None	Count	11382	3746	863	15991
		% within PAIN_RELIEF_NAME_EN	71.2%	23.4%	5.4%	100.0%
		% within LOS_CAT	77.3%	59.0%	62.4%	71.2%
		% of Total	50.7%	16.7%	3.8%	71.2%
	Others	Count	19	86	35	140
		% within PAIN_RELIEF_NAME_EN	13.6%	61.4%	25.0%	100.0%
		% within LOS_CAT	0.1%	1.4%	2.5%	0.6%
		% of Total	0.1%	0.4%	0.2%	0.6%
	Panadol	Count	25	14	0	39
		% within PAIN_RELIEF_NAME_EN	64.1%	35.9%	0.0%	100.0%
		% within LOS_CAT	0.2%	0.2%	0.0%	0.2%

		% of Total	0.1%	0.1%	0.0%	0.2%
	Pethedine	Count	1953	1327	261	3541
		% within PAIN_RELIEF_NAME_EN	55.2%	37.5%	7.4%	100.0%
		% within LOS_CAT	13.3%	20.9%	18.9%	15.8%
		% of Total	8.7%	5.9%	1.2%	15.8%
	Phenergan	Count	244	106	7	357
		% within PAIN_RELIEF_NAME_EN	68.3%	29.7%	2.0%	100.0%
		% within LOS_CAT	1.7%	1.7%	0.5%	1.6%
		% of Total	1.1%	0.5%	0.0%	1.6%
	Scobotel	Count	937	289	28	1254
		% within PAIN_RELIEF_NAME_EN	74.7%	23.0%	2.2%	100.0%
		% within LOS_CAT	6.4%	4.6%	2.0%	5.6%
		% of Total	4.2%	1.3%	0.1%	5.6%
	Tramal	Count	163	778	188	1129
		% within PAIN_RELIEF_NAME_EN	14.4%	68.9%	16.7%	100.0%
		% within LOS_CAT	1.1%	12.3%	13.6%	5.0%
		% of Total	0.7%	3.5%	0.8%	5.0%
Total		Count	14728	6351	1382	22461

	% within PAIN_RELIEF_NAME_EN	65.6%	28.3%	6.2%	100.0%
	% within LOS_CAT	100.0%	100.0%	100.0%	100.0%
	% of Total	65.6%	28.3%	6.2%	100.0%

NICU (neonatal intensive-care unit):

This attribute explores if the baby admission to NICU or not**.** A neonatal intensive-care unit (NICU), also known as an intensive care nursery (ICN), is an intensive-care unit specializing in the care of ill or premature newborn infants.

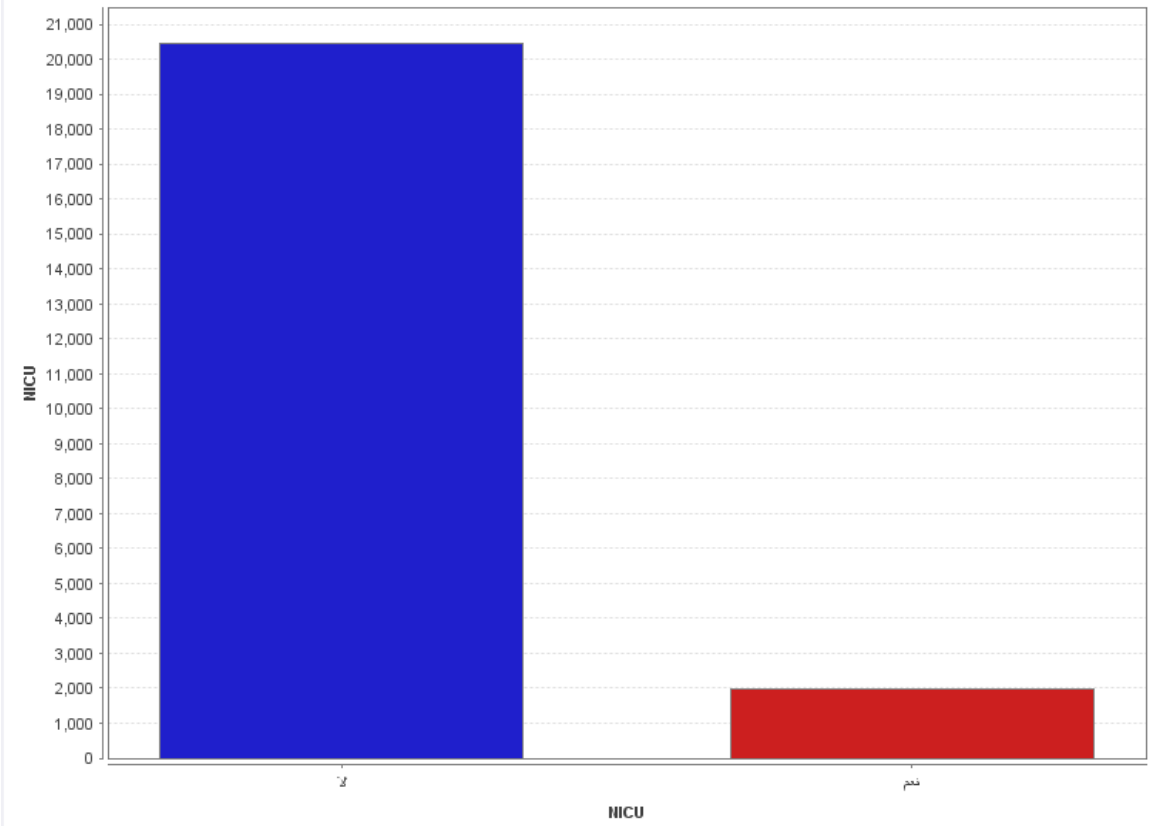


Figure 4.14 Summary of NICU Attribute

The results of cross tabulation between NICU and LOS showed that the LOS is affected by NICU; where there is an increase in LOS in cases baby admission to NICU in childbirth.

The results show that the percentage of LOS is in admission to NICU cases as 48.9% in LOS2 (24-72) hours and 19.0% in LOS3 (>72) hours. As shown in table 4.12.

Table 4.12: Cross tabulation of NICU & LOS

NICU * LOS_CAT Cross tabulation						
			LOS_CAT			Total
			LOS1	LOS2	LOS3	
NICU	NO	Count	14091	5381	1006	20478
		% within NICU	68.8%	26.3%	4.9%	100.0%
		% within LOS_CAT	95.7%	84.7%	72.8%	91.2%
		% of Total	62.7%	24.0%	4.5%	91.2%
	YES	Count	637	970	376	1983
		% within NICU	32.1%	48.9%	19.0%	100.0%
		% within LOS_CAT	4.3%	15.3%	27.2%	8.8%
		% of Total	2.8%	4.3%	1.7%	8.8%
Total		Count	14728	6351	1382	22461
		% within NICU	65.6%	28.3%	6.2%	100.0%
		% within LOS_CAT	100.0%	100.0%	100.0%	100.0%
		% of Total	65.6%	28.3%	6.2%	100.0%

GENERATOR_NAME_AR:

This attribute explores type of practitioner. The main difference between doctors and midwives is that, while midwives are trained to deal with women who are having normal, uncomplicated, low-risk pregnancies, doctors are trained to handle any complications.

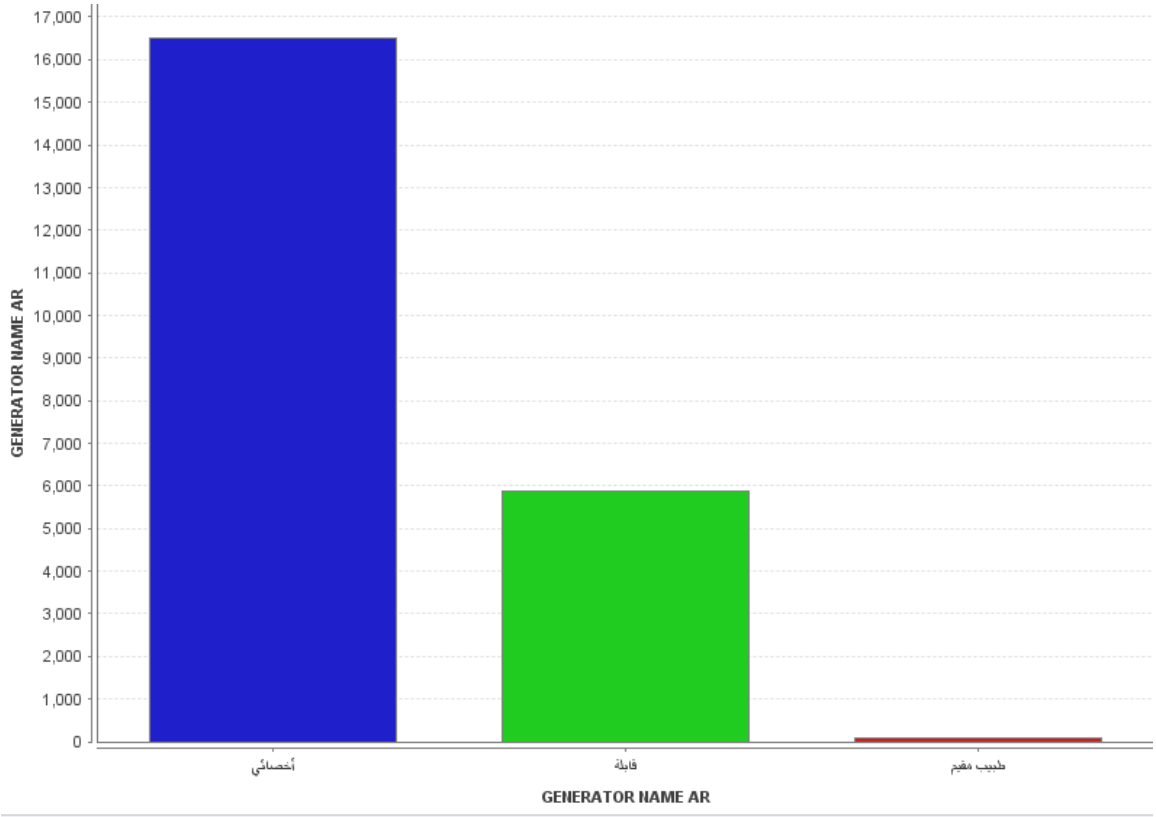


Figure 4.15 Summary of GENERATOR_NAME_AR Attribute

The results of cross tabulation between type of practitioner and LOS showed that the LOS is affected by type of practitioner; where there is an increase in LOS in cases type of practitioner in childbirth. The results show that the percentage of LOS is in doctor practitioner cases as 35.7% in LOS2 (24-72) hours and 8.0% in LOS3 (>72) hours. As shown in table 4.13.

Table 4.13 Cross tabulation of GENERATOR_NAME_AR & LOS

GENERATOR_NAME_AR * LOS_CAT Crosstabulation						
			LOS_CAT			
			LOS1	LOS2	LOS3	Total
GENERATOR_NAME_AR	doctor	Count	9303	5892	1319	16514
		% within GENERATOR_NAME_AR	56.3%	35.7%	8.0%	100.0%
		% within LOS_CAT	63.2%	92.8%	95.4%	73.5%
		% of Total	41.4%	26.2%	5.9%	73.5%
	resident doctor	Count	48	19	3	70
		% within GENERATOR_NAME_AR	68.6%	27.1%	4.3%	100.0%
		% within LOS_CAT	0.3%	0.3%	0.2%	0.3%
		% of Total	0.2%	0.1%	0.0%	0.3%
	midwife	Count	5377	440	60	5877
		% within GENERATOR_NAME_AR	91.5%	7.5%	1.0%	100.0%
		% within LOS_CAT	36.5%	6.9%	4.3%	26.2%
		% of Total	23.9%	2.0%	0.3%	26.2%
Total		Count	14728	6351	1382	22461
		% within GENERATOR_NAME_AR	65.6%	28.3%	6.2%	100.0%
		% within LOS_CAT	100.0%	100.0%	100.0%	100.0%
		% of Total	65.6%	28.3%	6.2%	100.0%

BLOOD_TRANS:

This attribute explores if it was a blood transfusion for pregnant or not.

A blood transfusion is a frequently performed procedure where you receive blood through an intravenous (IV) line into one of your blood vessels. However, there are two conditions that may warrant a blood transfusion during pregnancy; Iron-Deficient Anemia and Hemorrhage.

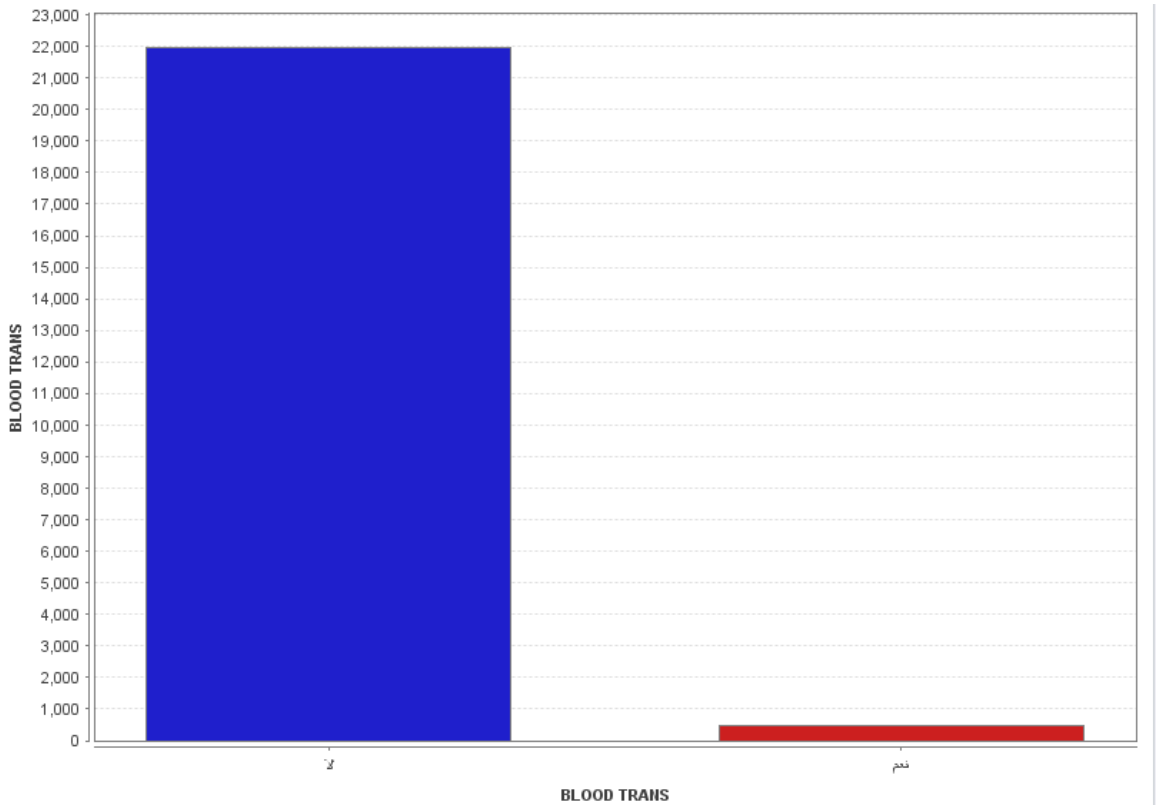


Figure 4.16 Frequency of BLOOD_TRANS Attribute

The results of cross tabulation blood transfusion for pregnant and LOS showed that the LOS is affected by blood transfusion for pregnant; where there is an increase in LOS in cases have blood transfusion in childbirth. The results show that the percentage of LOS is in blood transfusion in childbirth cases as 52.7% in LOS2 (24-72) hours and 27.7%% in LOS2 (24-72) hours other cases. As shown in table 4.14.

Table 4.14: Cross tabulation of BLOOD_TRANS AGE & LOS

BLOOD_TRANS * LOS_CAT Cross tabulation						
			LOS_CAT			Total
			LOS1	LOS2	LOS3	
BLOOD_TRANS	NO	Count	14649	6095	1231	21975
		% within BLOOD TRANS	66.7%	27.7%	5.6%	100.0%
		% within LOS_CAT	99.5%	96.0%	89.1%	97.8%
		% of Total	65.2%	27.1%	5.5%	97.8%
	YES	Count	79	256	151	486
		% within BLOOD TRANS	16.3%	52.7%	31.1%	100.0%
		% within LOS_CAT	0.5%	4.0%	10.9%	2.2%
		% of Total	0.4%	1.1%	0.7%	2.2%
Total		Count	14728	6351	1382	22461
		% within BLOOD TRANS	65.6%	28.3%	6.2%	100.0%
		% within LOS_CAT	100.0%	100.0%	100.0%	100.0%
		% of Total	65.6%	28.3%	6.2%	100.0%

CATALYST_NAME_EN:

This feature indicates if the women have any labor induction. Labor induction - also known as inducing labor - is a procedure used to stimulate uterine contractions during pregnancy before labor begins on its own. Successful labor induction leads to a vaginal birth. A health care provider might recommend labor induction for various reasons, primarily when there's concern for a mother's health or a baby's health. as shown in table 4.16, and figure 4.17.

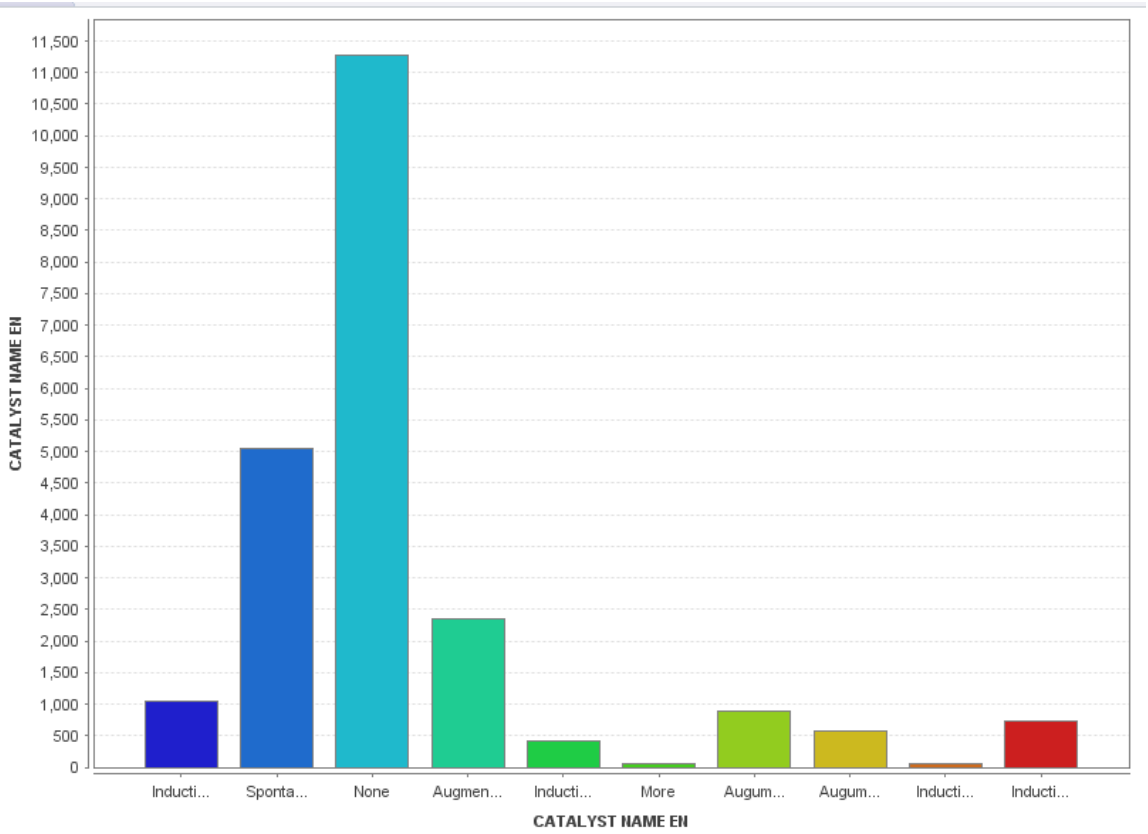


Figure 4.17 Summary of CATALYST_NAME_EN Attribute

The results indicate through data analysis that almost 50% of women have labor induction. In addition the percentage of LOS is less than 24 hours about 60% for none catalyst.

Table 4.15: Cross tabulation of CATALYST_NAME_EN & LOS

CATALYST_NAME_EN * LOS_CAT Cross tabulation						
			LOS_CAT			Total
			LOS1	LOS2	LOS3	
CATALYST_NAME_EN	Augmentation by oxytocin	Count	1701	619	33	2353
		% within CATALYST_NAME_EN	72.3%	26.3%	1.4%	100.0%
		% within LOS_CAT	11.5%	9.7%	2.4%	10.5%
		% of Total	7.6%	2.8%	0.1%	10.5%

	Augumentation by ARM	Count	766	114	13	893
		% within CATALYST_NAME_EN	85.8%	12.8%	1.5%	100.0%
		% within LOS_CAT	5.2%	1.8%	0.9%	4.0%
		% of Total	3.4%	0.5%	0.1%	4.0%
	Augumentation by ARM and oxytocin	Count	439	120	13	572
		% within CATALYST_NAME_EN	76.7%	21.0%	2.3%	100.0%
		% within LOS_CAT	3.0%	1.9%	0.9%	2.5%
		% of Total	2.0%	0.5%	0.1%	2.5%
	Induction by Oxytocin	Count	710	318	15	1043
		% within CATALYST_NAME_EN	68.1%	30.5%	1.4%	100.0%
		% within LOS_CAT	4.8%	5.0%	1.1%	4.6%
		% of Total	3.2%	1.4%	0.1%	4.6%
	Induction by prostaglandin	Count	197	209	21	427
		% within CATALYST_NAME_EN	46.1%	48.9%	4.9%	100.0%
		% within LOS_CAT	1.3%	3.3%	1.5%	1.9%
		% of Total	0.9%	0.9%	0.1%	1.9%
	Induction by prostaglandin E1	Count	28	26	3	57
		% within CATALYST_NAME_EN	49.1%	45.6%	5.3%	100.0%

		% within LOS_CAT	0.2%	0.4%	0.2%	0.3%
		% of Total	0.1%	0.1%	0.0%	0.3%
	Induction by prostaglandin E2	Count	151	498	77	726
		% within CATALYST_NAME_EN	20.8%	68.6%	10.6%	100.0%
		% within LOS_CAT	1.0%	7.8%	5.6%	3.2%
		% of Total	0.7%	2.2%	0.3%	3.2%
	More	Count	22	30	3	55
		% within CATALYST_NAME_EN	40.0%	54.5%	5.5%	100.0%
		% within LOS_CAT	0.1%	0.5%	0.2%	0.2%
		% of Total	0.1%	0.1%	0.0%	0.2%
	None	Count	6746	3453	1087	11286
		% within CATALYST_NAME_EN	59.8%	30.6%	9.6%	100.0%
		% within LOS_CAT	45.8%	54.4%	78.7%	50.2%
		% of Total	30.0%	15.4%	4.8%	50.2%
	Spontaneous	Count	3968	964	117	5049
		% within CATALYST_NAME_EN	78.6%	19.1%	2.3%	100.0%
		% within LOS_CAT	26.9%	15.2%	8.5%	22.5%
		% of Total	17.7%	4.3%	0.5%	22.5%
Total		Count	14728	6351	1382	22461

	% within CATALYST_NAME _EN	65.6%	28.3%	6.2%	100.0%
	% within LOS_CAT	100.0%	100.0%	100.0%	100.0%
	% of Total	65.6%	28.3%	6.2%	100.0%

4.4.1.3 Selection of Instances

Building a predictive model of Length of Stay for delivery (Normal- cesarean - Abortion) requires selection of instances with no additional type of department admissions. Thus, in addition to the removal of irrelevant attributes which was done based on the attributes irrelevance to the prediction, just instances that have admission delivery date were selected from the database. 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 (Han & Kamber, 2001).The remaining dataset will then have 22461records.

4.4.2 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. After ignoring attributes that have no data mining value, the remaining attributes were checked for missing values, inconsistencies and other interpretable observations. The data collected had a huge number of variables with missing values. Table 4.17 summarizes variables and percentage (%) of missing values associated with each attribute.

Table 4.16: Attributes listed List of Variables with their Missing Values

No	Attribute name	Percentage of missing values	Missing >10%
1	MRP_DOB	0%	
2	R_NAME_AR	0%	
3	C_NAME_AR	0.08%	
4	ADMISSION_ENTER_DATE	0%	
5	ADMISSION_OUT_DATE	0%	
6	ADMISSION_DELIVERY_DATE	0%	
7	OC_NAME_AR	0%	
8	DELIVERY_NAME_AR	0.02%	
9	PRE_RISK_FACTOR	0%	
10	POST_RISK_FACTOR	83.24%	×
11	REASON_NAME_AR	0%	
12	ADMISSION_TWINS	0%	
13	BI_WEIGHT_GM	0.5 %	
14	BOC_NAME_AR	0.45 %	
15	PRE_NAME_AR	23.47	×
16	BORN_EXAM	0.45	
17	EXM_NAME_AR	65.11%	×
18	CA_NAME_AR	100%	×
19	BI_APAGAR_1	14.57	×
20	BI_APAGAR_5	14.57	×
21	PAIN_RELIEF_NAME_EN	0.02%	
22	ICD_CD	0%	
23	ICD_NAME_EN	0%	
24	NICU	0.45	
25	PARTOGRAM	11.63	×
26	GENERATOR_NAME_AR	0.02%	
27	STATUS_NAME_AR	22%	×
28	BLOOD_TRANS	0.02%	
29	MOTHER_EXAM	0.02%	

No	Attribute name	Percentage of missing values	Missing >10%
30	MOTHER_RESULT	7.29	
31	CATALYST_NAME_EN	0.02%	
32	ADMISSION_MOTHER_EXAM_COUNT	17.28%	X
33	PT_NAME_AR	0%	

4.4.2.1 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 (Cios, Pedrycz, & Swiniarski, 1998).

After identifying percentage of missing instances, attributes with a higher percentage of missing values have been removed from the dataset due to the fact that they may compromise the research goal. The removed attributes have more than 10% missing values.

4.4.3 Data Transformation

Data transformation is necessary for two purposes to fix problems with the data such as missing values and categorical variables that take on too many values, and to bring information to the surface by creating new variables to represent trends and other ratios and combinations.

Table 4.17: Attributes listed List of Variables Data Transformation

No	Original attributes	Derived attributes	Values
1	MRP_DOB(calclate when addmision)	AGE	numeric

2	ADMISSION_ENTER_DATE – ADMISSION_OUT_DATE	LOS_FROM_ADDMADMISSION_HOURS	Numeric
3	BORN_EXAM + EXM_NAME_AR	EXAM_BABY_RESULTS	Nominal
4	MOTHER_EXAM+ MOTHER_RESULT	EXAM_MOTHER_RESULT	Nominal

4.5 Predictive Model Building Using Decision Tree

In this study an attempt was made to design a model that enables to predict the length of stay in government maternal hospitals. To this end, decision tree classifier is experimented on. Childbirth database is consulted to extract the dataset required for training and evaluating the models created by the classifiers. For creating prediction model a total size of 22461 datasets are used for training and testing. The researcher used 10-foldcross validation to measure the various performances of the classifiers. In 10-fold cross validation, the data is divided in10 segments, where 9 segments are used in the training phase, and the remaining segment is used in the test phase to measure the performance of the model. This process is repeated 10 times, each time with a different set of 9 segments as the training set, and the remaining segment as the test set. The overall performance measures of the model are then averaged out over the 10 different runs.

The type of classification selected for the experimentation was decision tree. The classifier has operation that is completely interactive and they benefit from powerful visualization features. The experiment on the decision tree predictive model building was based on the childbirth data from MOH maternity hospitals that has been preprocessed

and introduced to the RapidMiner software. In the experiments, variable "LOS" was set as dependent variable and the remaining other attributes were set as independent variables. The classification tree was built using all the default parameters suggested by the RapidMiner software as shown in figure (Figure 4.19).

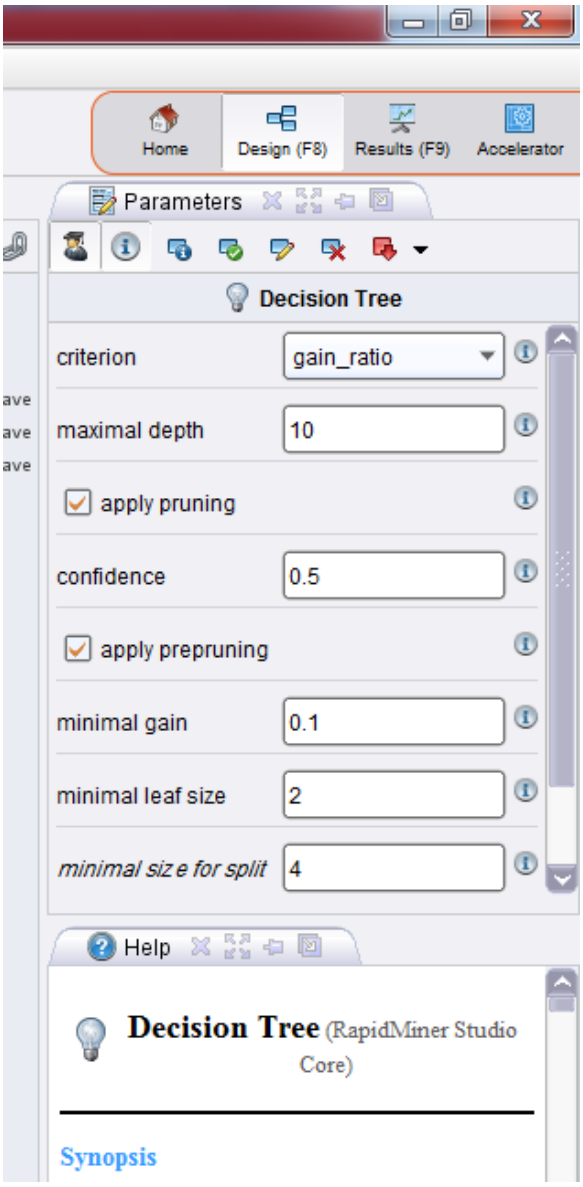


Figure 4.18 Decision tree parameters

The experiments were done with the decision tree using different with a 10-fold cross-validation mode.

Figure 4.20 shows an example decision tree using default parameters with rapidminer software. The numbers at the leaves are the predicted class. The visualization of the regression tree informs patients which attributes have the highest predictive powers for hospitalization period as well as identifies the interactions between attributes.

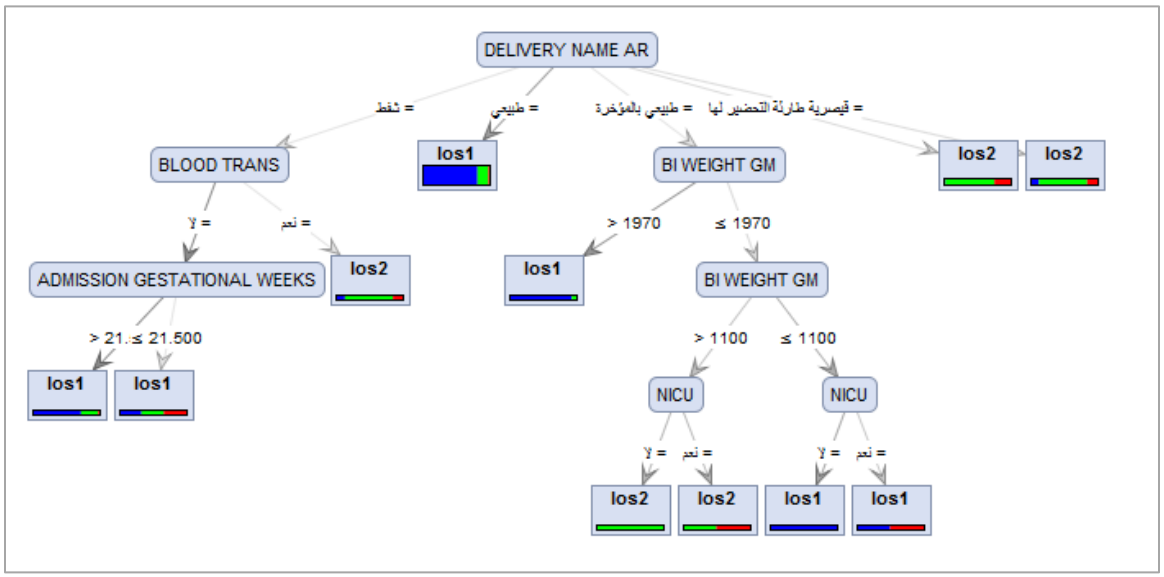


Figure 4.19 Decision tree for our model

According to Figure 4.20, delivery type is the leading factor for the length of stay. Hospitalization period are likely to be lower if delivery type is normal. The next influential factor is the blood transfer. A patient who was a blood transfusion he stayed for a longer period. For patients who stay more than 24 hours and less than 72 hours in the hospital, the weight of fetus is the next important factor for their hospital charges. If the weight of fetus less than 1100 grams and baby admission to Neonatal Intensive Care Unit (NICU),the length of stay will be increase.

In addition, the insurance type factor effects on hospitalization period ((Benbelkacem et al., 2014)), but in this study the insurance type factor not has any effect on length of stay, because the most of populations have health insurance, therefor

difficult economic conditions in Gaza strip and preferred to deliver in government hospital.

The visualization of decision trees, compared to most predictive modeling techniques, provides a more intuitive understanding of length of stay drivers and their interaction to empower healthcare consumers.

4.6 Model Evaluation

The valuation of models is carried out using 10-fold cross validation on the Childbirth dataset. It is based on the partition of the original sample into ten subsamples, using nine as training data and one for testing data. The cross validation process is then repeated ten times with each of the ten samples, averaging the results from the ten folds to produce a single estimation (De Toledo et al., 2009) Figure 3 shows the performance of the prediction model obtained with each classifier in terms of accuracy, precision, recall, kappa statistic.

accuracy: 79.99% +/- 0.37% (mikro: 79.99%)

	true los1	true los2	true los3	class precision
pred. los1	14205	2851	256	82.05%
pred. los2	356	3339	926	72.26%
pred. los3	0	1	0	0.00%
class recall	97.56%	53.93%	0.00%	

Figure 4.20 Performance evaluation of LOS prediction Models

The overall accuracy score equal to 79.99 is obtained by decision tree. The highest precision for classes score again belongs to LOS (< 24h). Also, the recall was slightly better with class 1 (LOS < 24 h). As shown by results, this study can be benefit to

maternity hospital manager to predict LOS. This information can be used to make the decision more proactive.

Chapter 5

Conclusion and Future Works

Chapter 5 CONCLUSION AND RECOMMENDATIONS

5.1 Summary and Conclusion

5.1.1 Summary

Data mining is extracting meaningful patterns and rules from large quantities of data. It is clearly useful in any field where there are large quantities of data and something worth learning. In this respect, widespread use of medical information systems and explosive growth of medical databases require traditional manual data analysis to be coupled with methods for efficient computer-assisted analysis.

This research investigated the potential applicability of data mining technology in improving healthcare resource management and developing a model to predict the occurrence of length of stay (LOS) in Ministry of Health (MOH) hospitals.

This investigation was conducted according to the CRISP-DM process model. The data was collected from childbirth database organized from 01/01/2013 to 30/12/2013 for the research purpose. Analyzing the large volume of childbirth data and extracting useful information and knowledge for decision making about length of stay was done. First the data was preprocessed for data cleaning, attribute and feature selection, and data transformation. This experimental research, which engaged a CRISP-DM methodological approach, made use predictive modeling techniques, decision tree to address the problem. The experiment result shows that decision tree suitable for predicting LOS. Hence decision tree with 79.99 % accuracy prediction model building was selected to extract interesting rules to cite.

5.1.2 Conclusion

Knowledge gained with the use of techniques of data mining can be used to make successful decisions that will improve success of healthcare organization and health of the patients.

In addition, results from this study have shown that the problem of predicting LOS could be supported by the use of data mining, in particular with decision trees technique. Moreover, further extensive experiments at district level and using various data sources available with those organizations working in related public health will enhance the result obtained in this study.

5.2 Recommendations

This research work was carried out for academic purpose and it has revealed the potential applicability of data mining technology to improve healthcare resource management through developing a model to predict LOS using childbirth data. Accordingly, based on the findings of this research work, the researcher forwards the following recommendations for future work particularly in relation to the possible application of data mining technology in supporting the effort to efficient resource management.

Academic researchers:

- Results of this research could be improved through extensive tests and use of other prediction techniques such as neural networks, support vector machine or a combination of them. So further experiments need to be done for better classification performance.

- The childbirth data has a large quantity of missing values and there is inconsistency in filling out the required information in the dataset. Therefore there is a need to investigate the possibility of designing a classifier that works better given incomplete data.
- There is a need different data mining research investigations based on clinical datasets from different health facilities.
- Based on the amount and type of data available, a decision tree with rules that are simple to follow provides the best insight into this data. Of course, the best recommendation is to obtain more data, use a validation data set, and have subject matter expertise applied to enhance this analysis.

Ministry of health:

- Hospital and health centers especially those give delivery services for mothers must keep records properly that includes all mothers and child information, because these records are useful to apply data mining techniques.
- Before embarking on data mining, however, an organization must formulate clear policies on the privacy and security of patient records.
- Audit data, validating to increase the accuracy of predictive models.
- Building team consisting of experts in the field of data mining and health department to get the best results.
- Although in this study encouraging results were obtained, maternity primary healthcare center should be done by integrating the various childbirth data sources.

Finally data mining and advanced data analysis courses are becoming more and more popular in business schools. Researcher asked colleagues about such electives in their universities,

and also did some web- searching. It looks like different courses range in topics and flavor, depending in many times on the instructor's field of expertise (statistics, operations research, information systems, marketing, machine learning, etc). But all courses typically revolve around real business applications. This course is suitable for MBA students with interests in IS, finance, marketing, operational management, and healthcare management

References

1. Ansari, S., Chetlur, S., Prabhu, S., Kini, G. N., Hegde, G., & Hyder, Y. (2013). An Overview of Clustering Analysis Techniques used in Data Miniing. *International Journal of Emerging Technology and Advanced Engineering, 3*(12), 284-286.
2. Apté, C., & Weiss, S. (1997). Data mining with decision trees and decision rules. *Future generation computer systems, 13*(2), 197-210.
3. Awad, E. M., & Ghaziri, H. M. (2004). Knowledge Management, 2004. *ed: Prentice-Hall, Upper Saddle River, New Jersey*.
4. Azari, A., Janeja, V. P., & Mohseni, A. (2012). *Predicting hospital length of stay (PHLOS): A multi-tiered data mining approach.* Paper presented at the Data Mining Workshops (ICDMW), 2012 IEEE 12th International Conference on.
5. Bath, P. A. (2004). Data mining in health and medical information. *Annual review of information science and technology, 38*(1), 331-369.
6. Bellinger, G., Castro, D., & Mills, A. (2004). Data, information, knowledge, and wisdom.
7. Benbelkacem, S., Kadri, F., Chaabane, S., & Atmani, B. (2014). *A DATA MINING-BASED APPROACH TO PREDICT STRAIN SITUATIONS IN HOSPITAL EMERGENCY DEPARTMENT SYSTEMS.* Paper presented at the 10ème Conférence Francophone de Modélisation, Optimisation et Simulation-MOSIM'14.
8. Bielza, C., & Larranaga, P. (2014). Discrete Bayesian network classifiers: a survey. *ACM Computing Surveys (CSUR), 47*(1), 5.
9. Carlo, V. (2009). Business intelligence: data mining and optimization for decision making. *Editorial John Wiley and Sons*.
10. Chaffey, D., & White, G. (2010). *Business information management: improving performance using information systems*: Pearson Education.
11. Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C., & Wirth, R. (2000). CRISP-DM 1.0 Step-by-step data mining guide.
12. Chen, F., Deng, P., Wan, J., Zhang, D., Vasilakos, A. V., & Rong, X. (2015). Data mining for the internet of things: literature review and challenges. *International Journal of Distributed Sensor Networks, 501*, 431047.
13. Cios, K. J., Pedrycz, W., & Swiniarski, R. W. (1998). *Data Mining and Knowledge Discovery*: Springer.
14. Dao, T. K., Zabaneh, F., Holmes, J., Disrude, L., Price, M., & Gentry, L. (2008). A practical data mining method to link hospital microbiology and an infection control database. *American Journal of Infection Control, 36*(3), S18-S20.
15. De Toledo, P., Rios, P. M., Ledezma, A., Sanchis, A., Alen, J. F., & Lagares, A. (2009). Predicting the outcome of patients with subarachnoid hemorrhage using machine learning techniques. *Information Technology in Biomedicine, IEEE Transactions on, 13*(5), 794-801.
16. Delen, D. (2014). *Real-World Data Mining: Applied Business Analytics and Decision Making*: FT Press.
17. Deshpande, S., & Thakare, D. V. (2010). Data mining system and applications: A review. *International Journal of Distributed and Parallel systems (IJDPS), 1*(1), 32-44.
18. Diwani, S., Mishol, S., Kayange, D. S., Machuve, D., & Sam, A. (2013). Overview Applications of Data Mining In Health Care: The Case Study of Arusha Region . *International Journal of Computational Engineering Research, 3*.

19. Dunham, M. H. (2006). *Data mining: Introductory and advanced topics*: Pearson Education India.
20. El-Sappagh, S., El-Masri, S., Riad, A., & Elmogy, M. (2013). Data mining and knowledge discovery: applications, techniques, challenges and process models in healthcare. *Int. J. Eng. Res. Appl, 3*(3), 900-906.
21. El-Sappagh, S. H., El-Masri, S., Elmogy, M., Riad, A., & Saddik, B. (2014). An ontological case base engineering methodology for diabetes management. *Journal of medical systems, 38*(8), 1-14.
22. Esfandiari, N., Babavalian, M. R., Moghadam, A.-M. E., & Tabar, V. K. (2014). Knowledge discovery in medicine: Current issue and future trend. *Expert Systems with Applications, 41*(9), 4434-4463.
23. Gosain, A., & Bhugra, M. (2013). *A comprehensive survey of association rules on quantitative data in data mining.* Paper presented at the Information & Communication Technologies (ICT), 2013 IEEE Conference on.
24. Hájek, P., Holeňa, M., & Rauch, J. (2010). The GUHA method and its meaning for data mining. *Journal of Computer and System Sciences, 76*(1), 34-48.
25. Han, J., & Kamber, M. (2001). Data mining: concepts and techniques: Morgan Kaufmann San Francisco, Calif, USA.
26. Han, J., Kamber, M., & Pei, J. (2011). *Data mining: concepts and techniques: concepts and techniques*: Elsevier.
27. Hand, D. J. (2007). Principles of data mining. *Drug safety, 30*(7), 621-622.
28. Health Information Center - Palestinian Ministry of Health. (2014). General Administration of hospitals.
29. Hwang, W.-J., & Wen, K.-W. (1998). Fast kNN classification algorithm based on partial distance search. *Electron Lett, 34*(21), 2062-2063.
30. Jashapara, A. (2004). *Knowledge management: an integral approach*: Pearson Education.
31. Jeng-Shyang, P., Yu-Long, Q., & Sheng-He, S. (2004). A fast K nearest neighbors classification algorithm. *IEICE Transactions on Fundamentals of Electronics, Communications and Computer Sciences, 87*(4), 961-963.
32. Jessup, L. M., & Valacich, J. S. (2002). *Information systems today*: Prentice Hall Professional Technical Reference.
33. Jiang, L., Zhang, H., & Cai, Z. (2009). A novel Bayes model: Hidden naive Bayes. *Knowledge and Data Engineering, IEEE Transactions on, 21*(10), 1361-1371.
34. Jovic, A., Brkic, K., & Bogunovic, N. (2014). *An overview of free software tools for general data mining.* Paper presented at the Information and Communication Technology, Electronics and Microelectronics (MIPRO), 2014 37th International Convention on.
35. Kesavaraj, G., & Sukumaran, S. (2013). *A study on classification techniques in data mining.* Paper presented at the Computing, Communications and Networking Technologies (ICCCNT), 2013 Fourth International Conference on.
36. Khajehei, M., & Etemady, F. (2010). *Data Mining and Medical Research Studies.* Paper presented at the Computational Intelligence, Modelling and Simulation (CIMSiM), 2010 Second International Conference on.
37. King, B. R., & Satyanarayana, A. (2013). *Teaching data mining in the era of big data.* Paper presented at the ASEE Annual Conference.
38. Koh, H. C., & Tan, G. (2011). Data mining applications in healthcare. *Journal of healthcare information management, 19*(2), 65.

39. Lan, H., Frank, E., & Hall, M. (2011). Data mining: Practical machine learning tools and techniques: Morgan Kaufman, Boston.
40. Mahindrakar, P., & Hanumanthappa, D. M. (2013). Data Mining In Healthcare: A Survey of Techniques and Algorithms with Its Limitations and Challenges. *Int. Journal of Engineering Research and Applications, ISSN*, 2248-9622.
41. Maimon, O., & Rokach, L. (2008). Data mining with decision trees: theory and applications: USA: World Scientific Publishing.
42. Mierswa, I., Wurst, M., Klinkenberg, R., Scholz, M., & Euler, T. (2006). *Yale: Rapid prototyping for complex data mining tasks.* Paper presented at the Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining.
43. Milovic, B. (2012). Prediction and decision making in Health Care using Data Mining. *International Journal of Public Health Science (IJPHS), 1*(2), 69-78.
44. Morton, A., Marzban, E., Giannoulis, G., Patel, A., Aparasu, R., & Kakadiaris, I. (2014). *A Comparison of Supervised Machine Learning Techniques for Predicting Short-Term In-Hospital Length of Stay Among Diabetic Patients.* Paper presented at the Machine Learning and Applications (ICMLA), 2014 13th International Conference on.
45. Ngai, E., Hu, Y., Wong, Y., Chen, Y., & Sun, X. (2011). The application of data mining techniques in financial fraud detection: A classification framework and an academic review of literature. *Decision Support Systems, 50*(3), 559-569.
46. North, M. (2012). *Data mining for the masses*.
47. Palestinian Central Bureau of Statistics (PCBS). (2013). Special Statistical Bulletin On the 65th Anniversary of the Palestinian Nakba. Retrieved 15 Jan, 2015, from http://www.pcbs.gov.ps/site/512/default.aspx?tabID=512&lang=en&ItemID=788&mid=3171&wversion=Staging
48. Pang-Ning, T., Steinbach, M., & Kumar, V. (2006). *Introduction to data mining.* Paper presented at the Library of Congress.
49. Pareek, D. (2006). *Business Intelligence for telecommunications*: CRC Press.
50. Rowley, J. E. (2007). The wisdom hierarchy: representations of the DIKW hierarchy. *Journal of information science*.
51. Sharma, A., & Mansotra, V. (2014). *Emerging applications of data mining for healthcare management-A critical review.* Paper presented at the Computing for Sustainable Global Development (INDIACom), 2014 International Conference on.
52. Shelly, G., Dharminder, K., & Anand, S. (2011). Performance Analysis Of Various Data Mining Classification Techniques On Healthcare Data. *International Journal of Computer Science and Information Technology, 3*(4), 155-169. doi: 10.5121/ijcsit.2011.3413
53. Sun, S. (2011). Analysis and acceleration of data mining algorithms on high performance reconfigurable computing platforms.
54. Thirumuruganathan, S. (2010). A detailed introduction to K-nearest neighbor (KNN) algorithm. *Retrieved March, 20*, 2012.
55. Thompson, T. L., & Warren, J. J. (2009). Are they all data? Understanding the work of organizational knowledge. *Clinical Nurse Specialist, 23*(4), 185-186.
56. Tomar, D., & Agarwal, S. (2013). A survey on Data Mining approaches for Healthcare. *International Journal of Bio-Science and Bio-Technology, 5*(5), 241-266. doi: 10.14257/ijbsbt.2013.5.5.25
57. Turban, E., Rainer, R., & Potter, R. (2005). Introduction to information technology, chapter 2nd, information technologies: Concepts and management: John Wiley and Sons.

58. Xing, Y., Wang, J., Zhao, Z., & Gao, Y. (2007). *Combination data mining methods with new medical data to predicting outcome of coronary heart disease.* Paper presented at the Convergence Information Technology, 2007. International Conference on.

Appendix A

Experiment

Appendix A Experiment

A.1. Rapidminer Main Screen

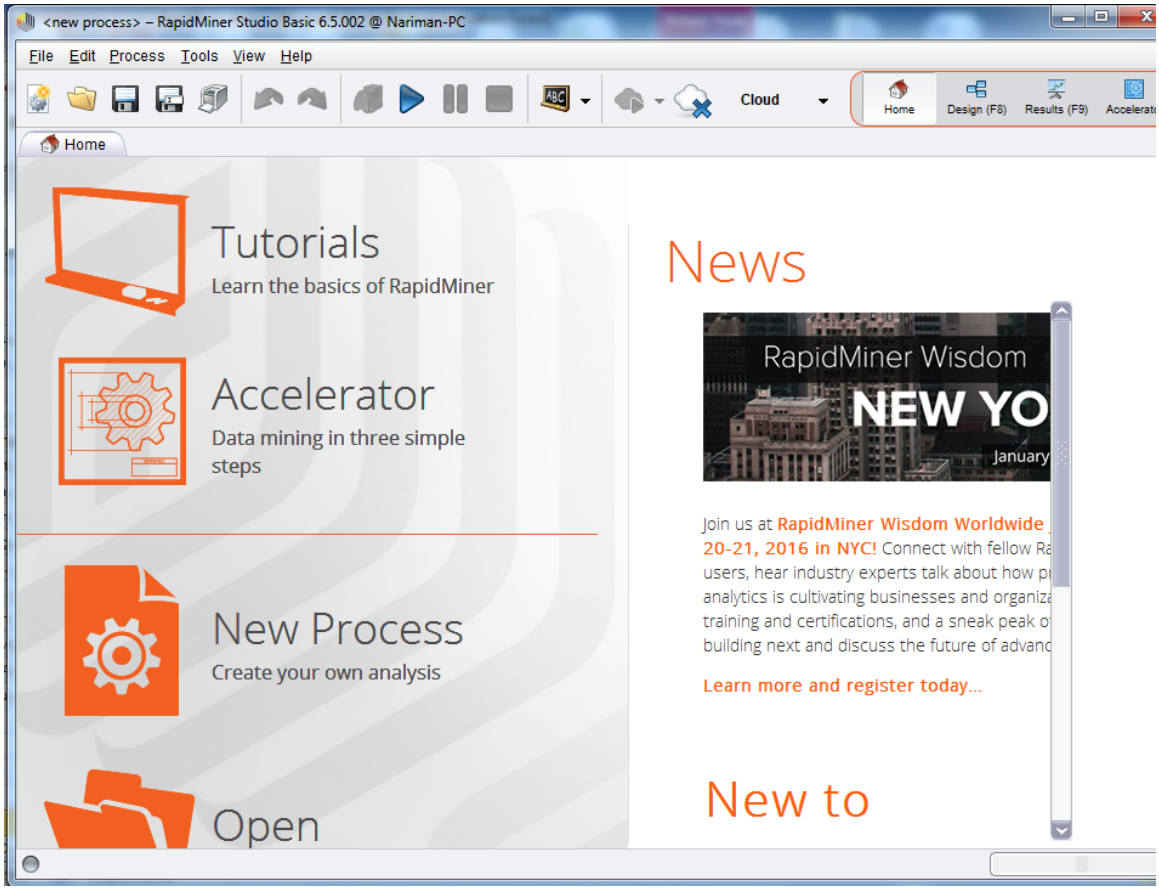


Figure A.1 Rapidminer Main Screen

A.2. Rapidminer Experiment Screens

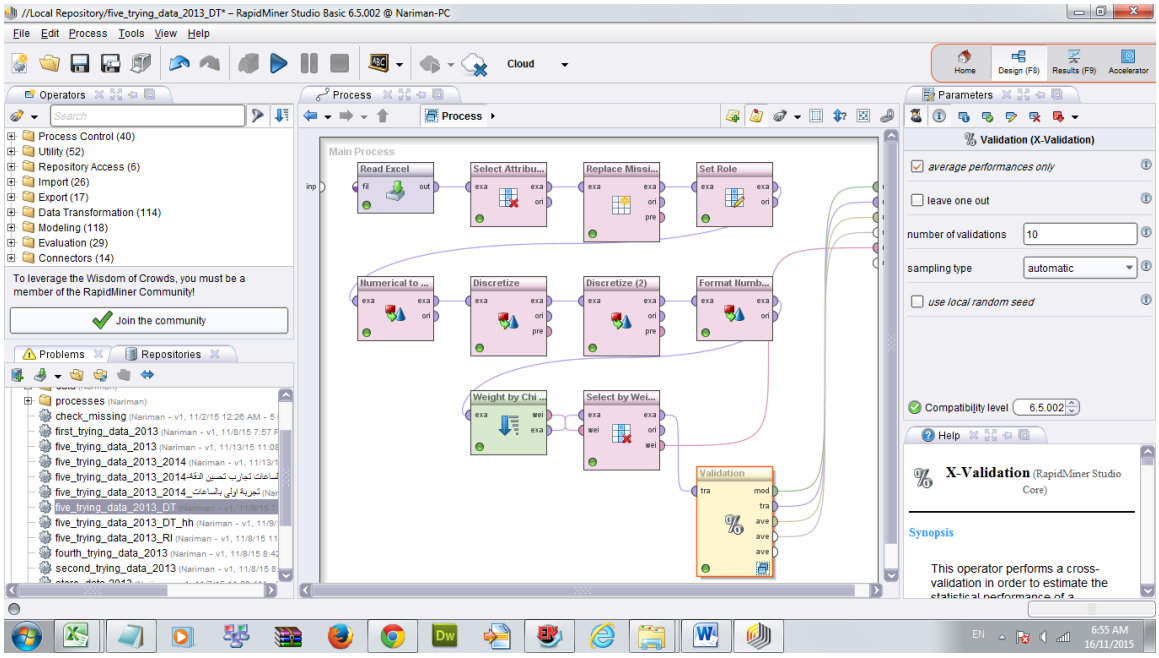


Figure A.2 Experiment Screen

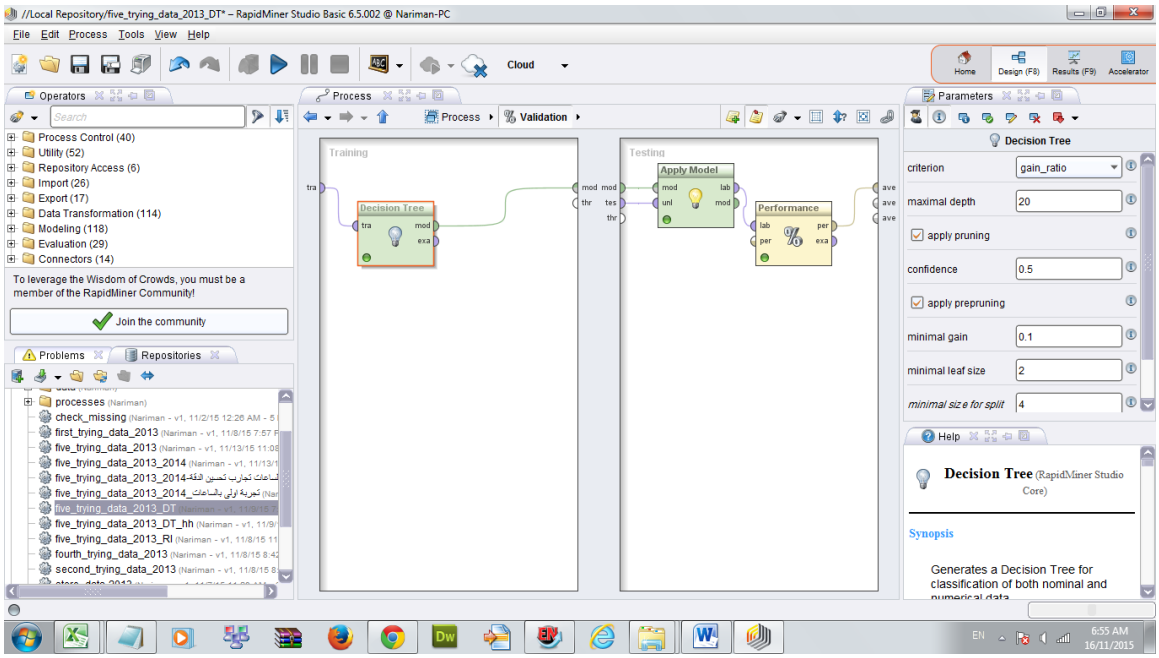


Figure A.3 Cross validation Screen

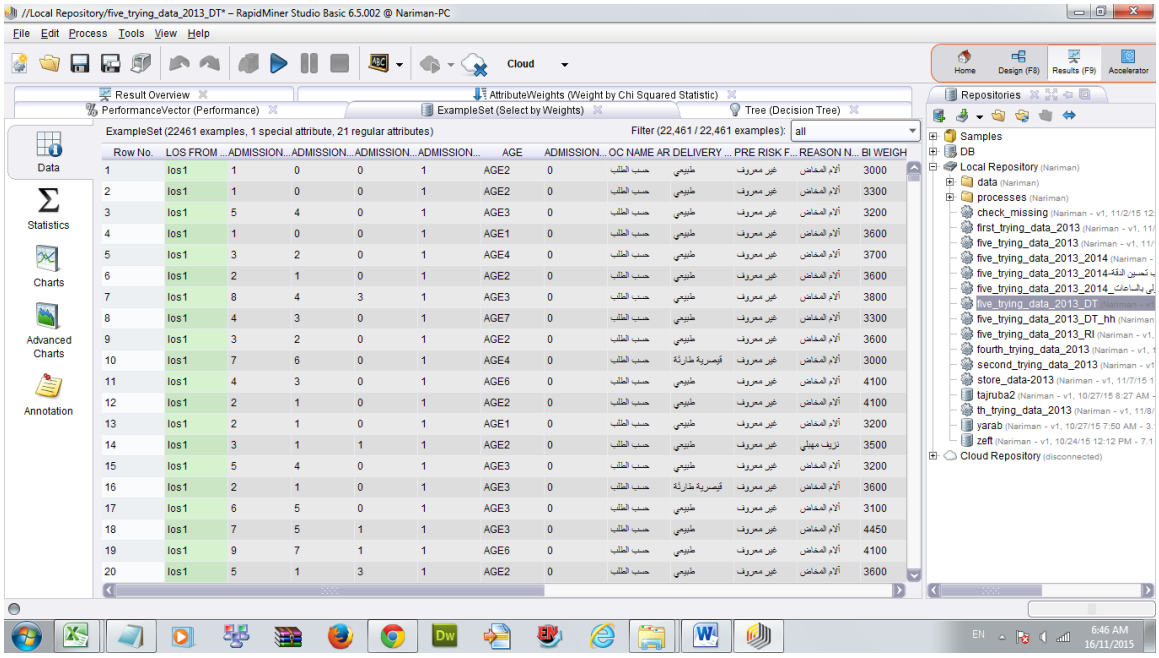


Figure A.4 Example set selecting by weights

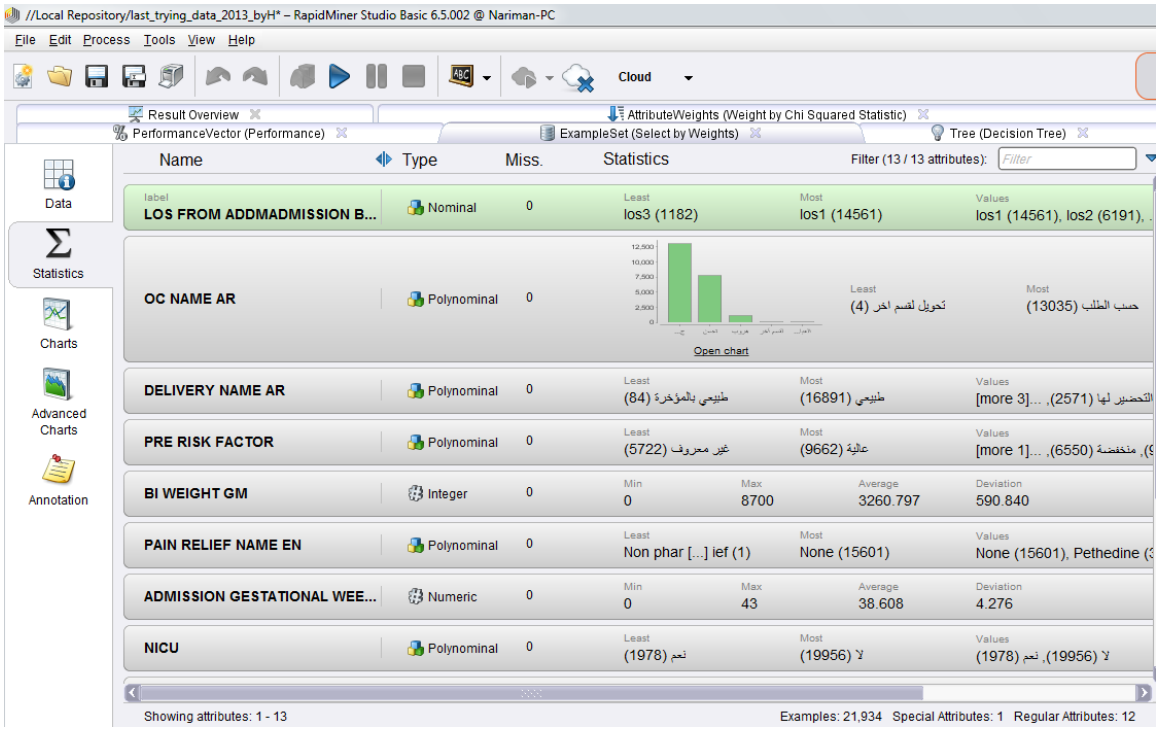


Figure A.5 Statistics of data set selecting by weights

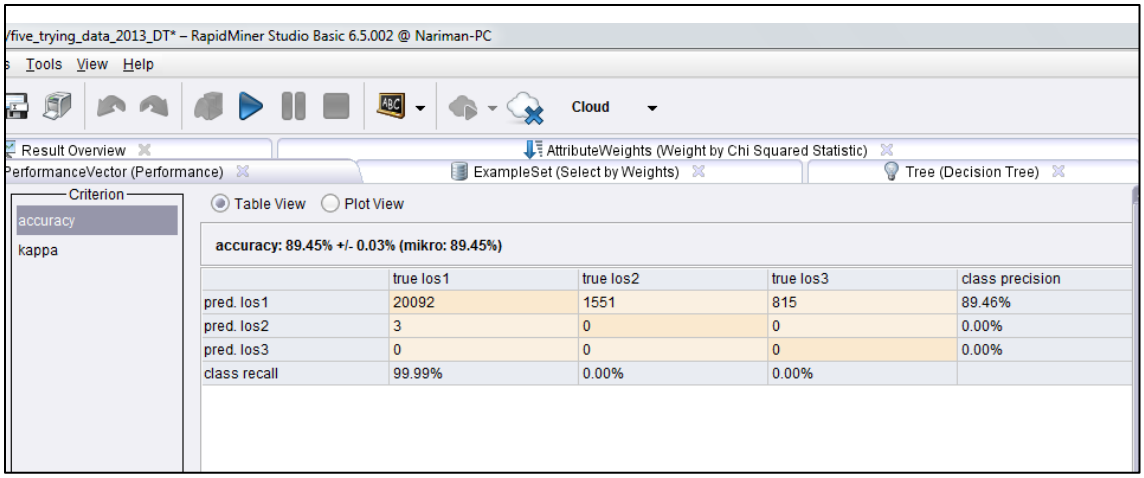


Figure A.6 Rapidminer confusion matrix

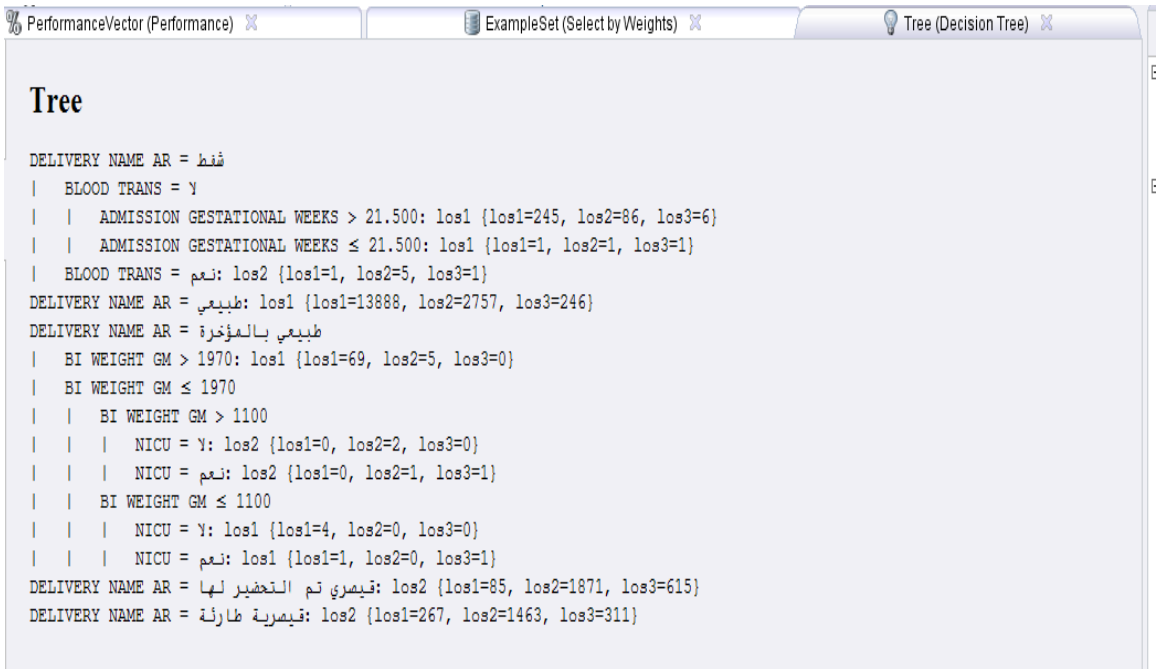


Figure A.7 Rapidminer text view

A.3. Rapidminer Weight by Chi Squared Statistic

Table 0.1: Weight by Chi Squared Statistic

Attribute	Weight
DELIVERY NAME AR	1.0
PRE NAME AR RESULT	0.979
PRE RISK FACTOR	0.244
GENERATOR_NAME_AR	0.243
ADMISSION GESTATIONAL WEEKS	0.21
PAIN_RELIEF_NAME_EN	0.19
CATALYST_NAME_EN	0.17
NICU	0.13
OC_NAME_AR	0.12
BORN EXAM RESULT	0.12
BI WEIGHT GM	0.08
BLOOD TRANS	0.08
AGE	0.04
ADMISSION TWINS	0.04
PT NAME AR	0.00
BOC_NAME_AR	0.00
STATUS NAME AR	0.00

Appendix B

Admission , Discharge of hospitals

Appendix B Admission , Discharge of hospitals

B.1. First Phase: Add Master Patient Index


MEDICAL RECORD

المسجل الطبي

CONFIDENTIAL

سـري

HOSPITAL:

مستشفى /

NOT TO BE HANDLED BY PATIENT

لا يسمح باعطاء السجل الطبي للمريض

NOT TO BE TAKEN OUT OF THE HOSPITAL

لا يسمح باخذ السجل الطبي خارج المستشفى

Patient's Name:

مجلد رقم VOLUME NO	1 I	2 II	3 III	4 IV	5 V
من OF	1 I	2 II	3 III	4 IV	5 V

يجب إعادة السجل الى قسم السجلات الطبية

PLEASE RETURN TO MEDICAL RECORDS DIVISION

حركة المنف ACTIVITY YEARS	Month																													
	Year																													
	Status																													

رقم السجل الطبي Medical Record Number	

Figure B.1 Medical Record Form

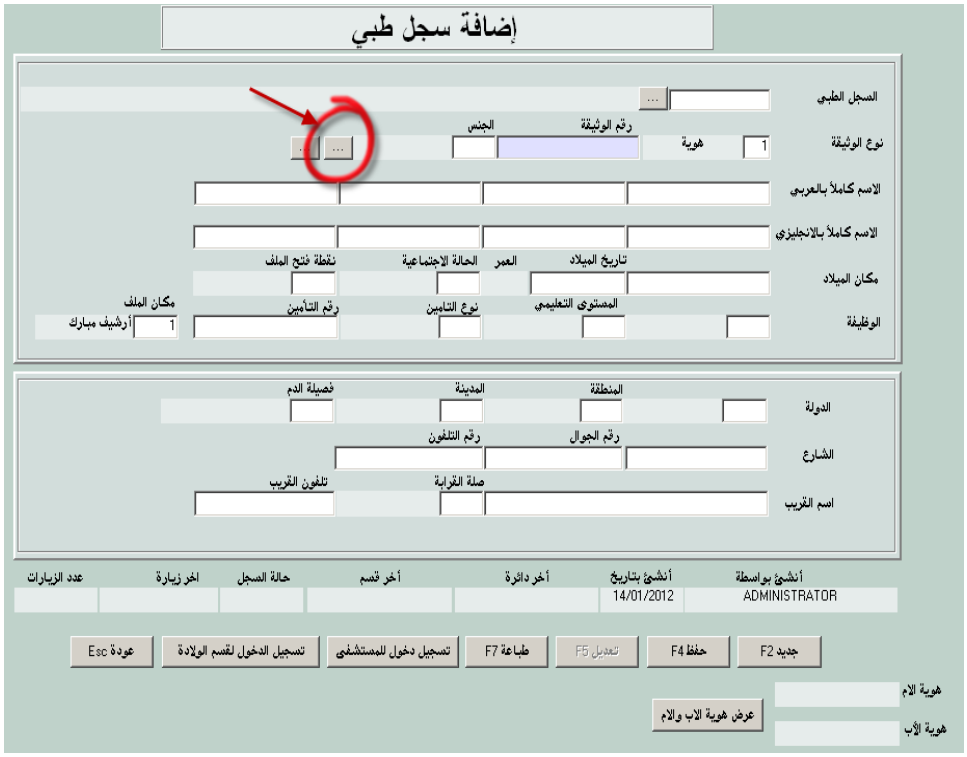


Figure B.2 Medical Record Screen

B.2. Second Phase: Admission to maternity hospital

Figure B.3 Admission to maternity button

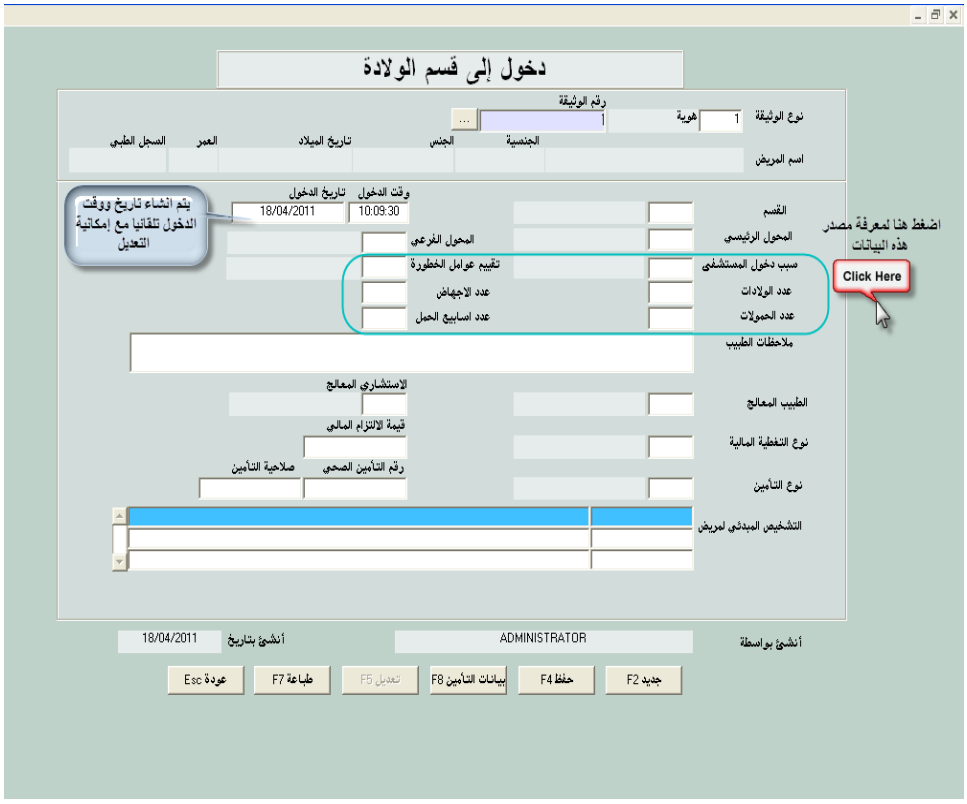
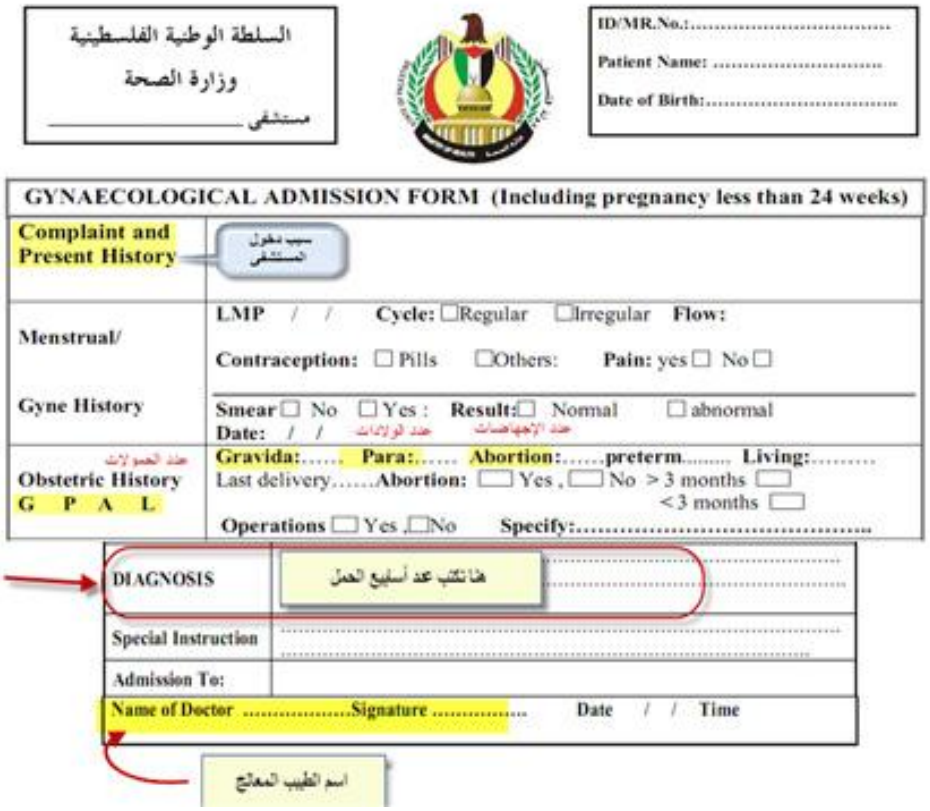


Figure B.4 Admission to maternity Screen


السلطة الوطنية الفلسطينية
وزارة الصحة
مستشفى _____

ID/MR.No.:.....
Patient Name: .....
Date of Birth:.....

GYNAECOLOGICAL ADMISSION FORM (Including pregnancy less than 24 weeks)

Complaint and Present History (سبب دخول المستشفى)	
Menstrual/ Gyne History	LMP / / Cycle: ☐Regular ☐Irregular Flow: Contraception: ☐ Pills ☐Others: Pain: yes ☐ No ☐ Smear ☐ No ☐ Yes : Result: ☐ Normal ☐ abnormal Date: / / (عدد الولادات) (عدد الإجهاضات)
Obstetric History (عدد الحمولات) G P A L	Gravida:..... Para:..... Abortion:.....preterm..... Living:..... Last delivery.....Abortion: ☐ Yes , ☐ No > 3 months ☐ < 3 months ☐ Operations ☐ Yes , ☐No Specify:.....
DIAGNOSIS (هنا تكتب عدد أسابيع الحمل)	.....
Special Instruction	.....
Admission To:	
Name of Doctor .....Signature	Date / / Time

(اسم الطبيب المعالج)

Figure B.5 Admission to maternity Form

B.3. Third Phase: Discharge From maternity hospital

The Data entered in this phase in next day of discharge of mothers and collected from many departments

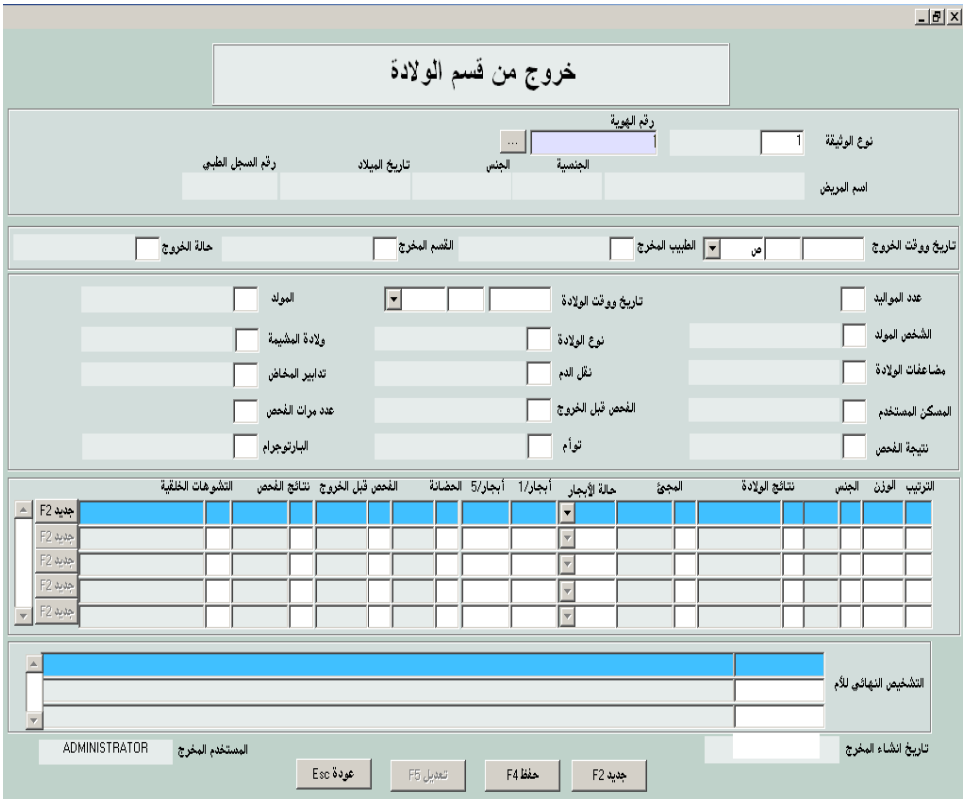


Figure B.6 Discharge From maternity hospital Screen