8-2002

A Proposed Data Mining Methodology and its Application to Industrial Engineering

Jose Solarte
University of Tennessee - Knoxville

To the Graduate Council:

I am submitting herewith a thesis written by Jose Solarte entitled "A Proposed Data Mining Methodology and its Application to Industrial Engineering." I have examined the final electronic copy of this thesis for form and content and recommend that it be accepted in partial fulfillment of the requirements for the degree of Master of Science, with a major in Industrial Engineering.

Denise F. Jackson, Major Professor

We have read this thesis and recommend its acceptance:

Robert E. Ford, Tyler A. Kress

Accepted for the Council:
Carolyn R. Hodges

Vice Provost and Dean of the Graduate School

(Original signatures are on file with official student records.)

To the Graduate Council:

I am submitting herewith a thesis written by Jose Solarte entitled “A Proposed Data Mining Methodology and its Application to Industrial Engineering.” I have examined the final electronic copy of this thesis for form and content and recommend that it be accepted in partial fulfillment of the requirements for the degree of Master of Science, with a major in Industrial Engineering.

Denise F. Jackson
Major Professor

We have read this thesis
And recommend its acceptance:

Robert E. Ford

Tyler A. Kress

Accepted for the Council:

Dr. Anne Mayhew
Vice Provost and
Dean of Graduate Studies

(Original signatures are on file with official student records.)

A Proposed Data Mining Methodology and its Application to Industrial Engineering

A Thesis

Presented for the

Master of Science

Degree

The University of Tennessee, Knoxville

Jose Solarte

August 2002

DEDICATION

This thesis is dedicated to my parents, Jose Solarte and Maria Rueda, great role models and friends, who have always assisted me and helped me any way they could; and to the rest of my family, for always believing in me, inspiring me, and encouraging me to reach higher goals. To all of them I want to say “Thank you”, for being the best family that I could ever have.

ACKNOWLEDGMENTS

I would like to thank my instructors and advisors Dr. Denise Jackson, Dr. Robert Ford and Dr. Tyler Kress of the University of Tennessee, Knoxville, for all the help, support, guidance and encouragement they have provided me. I would also like to thank Dr. Adedeji Badiru, Dr. John Hungerford, Dr. Patricia Fisher, Dr. Richard Pollard, Dr. Charles Aikens, Dr. Kenneth Kirby, Dr. Rapinder Sawhney, Dr. Hampton Liggett, and Dr. Wayne Claycombe, for their instruction and the privilege of being in their classes.

I thank all of those who have helped me in the realization of this project, thanks to Rachel Anne Dresbeck, Cornelia Reichel, Rob Wise, Linda Van der Spek, Gary Miner, Sheila D. Ferguson, Rob Van der Veer, Mark Yuhn, Barry Shepherd, Irina Sered, Mehmet Goker, Bop Pattison, Deborah Arnold, Michelle Bula, John Thompson, Estelle Brand, Allison Nipper, Guy Daniello, Holly Larocque, Donna Bartko, Clemens van Brunschot, Luke Taylor, Lise Reid, Charles Huot, Alison Foley, Vanessa Westwood, John Haines, Kimberly Covington, François Halfen, and specially thanks to Bob Muenchen for all the time and assistance that they have given to me.

Finally, I would like to thank my good friends Cristina Zaharia, Archana Niranjan, and Fernando Parrado, for all their encouragement and their invaluable friendship.

ABSTRACT

Data mining is the process of discovering correlations, patterns, trends or relationships by searching through a large amount of data stored in repositories, corporate databases, and data warehouses. Industrial engineering is a broad field and has many tools and techniques in its problem-solving arsenal. The purpose of this study is to improve the effectiveness of industrial engineering solutions through the application of data mining. To achieve this objective, an adaptation of the engineering design process is used to develop a methodology for effective application of data mining to databases and data repositories specifically designed for industrial engineering operations. This paper concludes by describing some of the advantages and disadvantages of the application of data mining techniques and tools to industrial engineering; it mentions some possible problems or issues in its implementation; and finally, it provides recommendations for future research in the application of data mining to facilitate decisions relevant to industrial engineering.

TABLE OF CONTENTS

LIST OF FIGURES

ABBREVIATIONS

CAD:	Computer-aided Design.
CART:	Classification and Regression Trees.
CHAID:	Chi squared Automatic Interaction Detection.
CRM:	Customer Relationship Management.
DFD:	Data Flow Diagram.
DSS:	Decision Support Systems.
ERD:	Entities Relationship Diagram.
MARR:	Minimum Attractive Rate of Return.
MNIS:	Manning and Napier Information Services.
ODBC:	Open Database Connectivity.
ODBMS:	Object-oriented Database-management System
ODMG:	Object Database Management Group.
OLAP:	On-line Analytical Processing.
OLE-DB	Open Linking and Embedding for Databases.
OQL:	Object Query Language.
PCA:	Principal Component Analysis.
RDBMS:	Relational Database-management System.
SEMMA:	Sample, Explore, Modify, Model and Assess.
SQC:	Statistical Quality Control.
SQL:	Structured Query Language.
SPC:	Statistical Process Control.

CHAPTER 1
INTRODUCTION

Introduction to this Research

Data mining has recently become one of the most progressive and promising fields for the extraction and manipulation of data to produce useful information. Thousands of businesses are using data mining applications every day in order to manipulate, identify, and extract useful information from the records stored in their databases, data repositories, and data warehouses.

With this kind of information, companies have been able to improve their businesses by applying the patterns, relationships, and trends that have lain hidden or undiscovered within colossal amounts of data. For example, data mining has produced information that enables companies to create profiles of current and prospective customers to help in gaining and retaining their customers. Other uses of data mining include development of cross-selling and marketing strategies, exposure of possible crimes or frauds, finding patterns in the access of users to their web sites, and process improvement.

The power of data mining is yet to be fully exploited by industry. Manufacturing, for example, is one of the new fields in which data mining tools and techniques are beginning to be used successfully. Process optimization, job shop scheduling, quality control, and human factors are some of the areas in which data mining tools such as neural networks, genetic algorithms, decision trees, and data visualization can be implemented with great results.

However, implementation of these data mining techniques is inconsistent in practice. Why? Because software vendors propose different

and proprietary approaches that focus on specific business applications. These approaches even use different sets of analysis tools. To develop good data mining strategies, industrial engineers require an application-neutral methodology. Moreover, too often, data mining approaches fail to keep the goals of an organization in mind, so that the results of the data mining project are irrelevant. In addition, a systems perspective is not maintained; thus essential components of the organizational system are overlooked and, again, the data mining results are not as effective as they ought to be. What is needed is a guide through the maze of tools and approaches to the myriad of applications.

Background

Data mining is often described as the process of discovering correlations, patterns, trends or relationships by searching through a large amount of data stored in repositories, corporate databases, and data warehouses. The kinds of relationships that exist are believed to be sometimes unclear to information analysts because the amounts of information are too large or the kinds of relationships are too difficult to imagine. Humans, in that sense, are limited by information overload; thus, new tools and techniques are being developed to solve this problem through automation.

Data mining uses a series of pattern recognition technologies and statistical and mathematical techniques to discover the possible rules or relationships that govern the data in the databases. Data mining must also be considered as an iterative process that requires goals and objectives to be specified [1]. Once the intended goals are completely defined, it is necessary to determine what data is available or can be collected. Sometimes the data

is available in data warehouses, but before it can be used, some filtering is performed to transform it into information.

Data mining also involves a methodology for implementation. The methodology, or structured approach, usually varies from vendor to vendor. SAS Institute [2], for example, promotes SEMMA (sample, explore, modify, model and assess). Another methodology is CRISP-DM by SPSS, Inc. Each methodology strives to help users obtain the best data to provide the most responsive information to address their needs.

The recognition that effective decisions are based on appropriate information from accurate and current data is not new. The evolution and development of finding the right data for decision-making began 30 years ago, and it has progressed through several stages of development [7]. These are shown in Figure 1, and their descriptions follow.

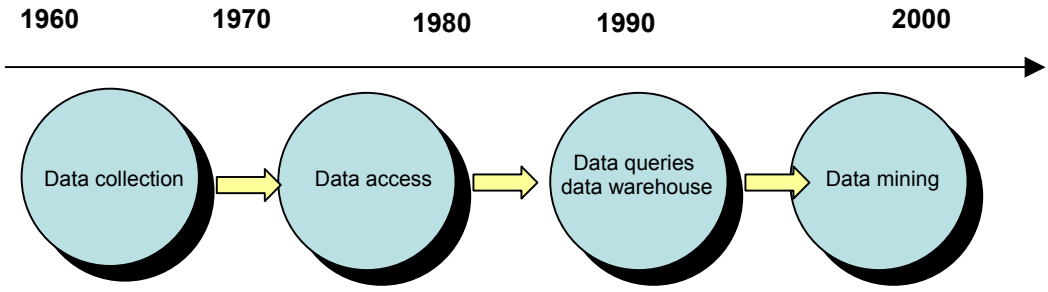


Figure 1. Data Mining Evolution

The evolutionary stages of data mining are as follows:

1. Data Collection.

During the late ‘60’s, simple reports of pre-formatted information were created from data stored in databases. These databases stored the data, while applications retrieved and manipulated it to produce structured reports containing information to meet specific decision-making needs.

2. Data Access.

In the 1980s, users began to want information more frequently and they wanted it to be more individualized. Thus, they began to make queries, or informational requests, of the databases. These were performed to obtain ad hoc information at a lower level of detail than the structured reports. The system developers generally defined these queries during system design and built them into the system.

3. Data Queries.

Later, in the 1990s, users required immediate access to more detailed information that responded to “on the fly” questions. They wanted information to be “just-in-time” to correlate with their production and decision-making processes. That meant that not all of the users’ informational needs could be preprogrammed into the system. At this stage, users began to write their own queries to extract the information that they needed from the database.

4. Data Mining.

In the last few years, users began to realize the need for more tools and techniques in order to identify and find relationships in data so that the information obtained was more meaningful for their applications. Additionally, companies recognized that they had accumulated volumes of data; and, as a result, they needed new tools to sort through it all and meet their informational needs. Such tools enabled the system to search for possible hidden relationships in the data, without the direct intervention of the end users. Data mining tools were first developed to help scientists find meaningful relationships or patterns from huge amounts of data that, if done in a traditional way, would require much time and many resources to find. The next step is to exploit these tools for meaningful applications.

Limitations of the Study

This study focuses specifically on applying data mining to problems generally addressed in industrial engineering. It emphasizes the application of a systems analysis and design perspective to develop a data mining methodology suitable for those applications. Only the information necessary to illustrate the concepts described in this document have been included. For that reason, requirements necessary for the application of data mining in other fields have been omitted.

The methodology proposed in this study is an abstract and functional framework. It is a conceptual model, it has not been implemented yet, and therefore it has not been executed or tested. That task remains for the future.

Problem Statement

Data mining not only involves a collection of systems, solutions or technologies, but also includes a structured process in which human interaction is important. Humans decide if the patterns discovered have some relevance to the problem at hand or if they justify further study and exploration. With this in mind, data mining approaches have been integrated with the needs and interests of specific businesses.

Data mining techniques can be used in many different fields and have many applications. They range from the biomedical and DNA analysis to financial analysis, and fraud detection. They can also be used to track customer preferences and for cross-selling products. More and more applications are being found every day.

In order for data mining techniques to provide the intended results--full exploitation of all available data--it is very important that the data is correctly prepared and collected for its specific applications. If there is no existing technique that matches, users must manipulate available ones to find the best fit. With so many choices on the market, users need assistance in deciding the various tools offered by the many vendors in the market.

Additionally, data mining applications continue to be developed. There are, however, few that support decision-making in industrial engineering. Thus, applications of data mining in areas such as quality control, process control, human factors, material handling and maintenance and reliability in production systems should be studied and addressed in more detail.

To address the lack of industrial engineering applications and guidelines for using data mining in existing applications, this research proposes to do the following:

- Develop a convenient methodology for the application of data mining in industrial engineering.
- Analyze and compare the different and successfully applied tasks and techniques used in data mining.
- Identify the main advantages and disadvantages for the application of data mining techniques and tools in industrial engineering.
- Identify possible problem areas or issues for the application of data mining in industrial engineering.
- Identify possible applications of data mining in the field of industrial engineering.

In order to accomplish these goals, existing approaches to data mining and current data mining applications are analyzed and reviewed. The results are then used to develop a proposed methodology for applying data mining to the informational needs of industrial engineering.

Structure of this Paper

This thesis is divided into five chapters, including this introductory chapter. Chapter 2, “Literature Review,” reviews some of the most important cases of data mining projects in areas such as quality control, scheduling, process control, process optimization, safety, cost reduction, maintenance and reliability, and product development. Chapter 3, “Research Methodology,” gives a general description of the research methodology used to address the problems which this study identifies earlier in this chapter. Chapter 4, “A Proposed Methodology” presents the proposed solution, a methodology for the application of data mining in the field of industrial engineering. Chapter 5, “Conclusions and Further Studies,” summarizes the major conclusions of this document. It presents the main advantages and disadvantages for the application of data mining techniques and tools in the field of industrial

engineering, as well as possible problem areas or issues for its implementation. Finally, this chapter also states possible areas of further research.

CHAPTER 2
LITERATURE REVIEW

Introduction

The application of data mining techniques to industrial engineering is an area that holds promise, but that is currently underdeveloped. Data mining, can, however, be strategically applied to industrial engineering processes such as scheduling, quality control, cost reduction, safety, and others. This chapter outlines some of the data mining techniques and applications that can be utilized by industrial engineers, as well as some of the existing ways that industrial engineers employ data mining.

Data Mining Techniques

There are a number of techniques used in data mining, but not all of them can be applied to all types of data. Neural network algorithms, for example, can be used to quantify data (numerical data), but they cannot qualify data precisely (categorical data); therefore, categorical data is usually broken up into multiple dichotomous variables, each of them with values of 1 (“yes”)or 0 (“no”) [34]. For that reason, one single technique cannot be used to perform a complete data mining study and each technique has its own scope of applications. Some of the techniques applied in data mining include traditional statistics, induction, neural networks, and data visualization. These are described in the following sections.

Traditional Statistics

Some of the traditional statistical methods that can be used for data mining are the following [18]:

- cluster analysis, also called segmentation.
- discriminant analysis.
- logistic regression.
- time series forecasting.

Cluster analysis (or segmentation) is one of the most frequently used data mining techniques; it involves separating sets of data into groups that include a series of consistent patterns. After the data reveals a consistent pattern, it is then sorted into subsets that are easier to analyze. This information is also used to identify subgroups of a population for supplementary studies, as well as to generate profiles for target marketing. Kohonen feature maps and K-means are some of the most important algorithms applied for cluster analysis [34].

Discriminant analysis is one of the oldest classification techniques. It finds hyper planes that separate classes so that users can then apply them to determine the side of the hyper plane in which to catalogue the data. Discriminant analysis has limitations, however. It assumes that all predictor variables are normally distributed--but this is not always true. Moreover, unordered categorical values cannot be classified, and boundaries are restricted to linear forms. New versions of discriminant analysis are being developed to handle these limitations by using quadratic boundaries, estimates of real distributions, and bins defined by the categorical variables [34].

Logistic regression is a generalization of linear regression. It is primarily used for predicting binary variables and, less frequently, multi-class variables. Models of logistic regression predict the logarithm of the odds of the occurrences of discrete variables. The main assumption of the logistic regression model is that the logarithm of the odds is linear in the coefficients of the predictor variables [34]. Analysts using this technique require experience and skill in order to select the right variables, choose the functional relationship with the response variable, and account for possible interactions.

Finally, time series forecasting predicts “unknown future values, based on time varying series of predictors” [34]. Time series databases contain series of sequences of values and events that change over time. The trends of those values can be used to construct functions of the form Y=f(t), so attributes can be predicted in time or based on other process values. For example, downtimes can be predicted as a function of setups. With this information, preventive maintenance programs can then be implemented, scheduled, and adjusted in real time. Yet with this technique, important factors such as hierarchy of periods, seasonality, calendar effects, and date arithmetic may influence the results. Thus, these factors should be accounted for when time series forecasting is used.

Induction and Decision Trees

Induction techniques try to uncover associations in the data. They search for similarities within the existing records and try to infer the rules that express those relationships. The specific occurrences of the events in the data are then applied to establish a “confidence factor of the rule” [1]. Decision trees are flow charts--tree structures in which nodes represent tests or attributes, branches represent test outcomes, and leaf nodes represent

classes or class distributions. Using decision trees, unknown events such as types of defects can be classified, testing the values of important attributes against the values of each node. By following this process, a path can be traced from the root node to the leaf node that identifies the class prediction for that event. Rules can be constructed very easily using decision trees, and they usually follow a form such as “If x =’y’ and z=’d’ and p =’0,5’, then defect = ‘yes’.”

Induction techniques have been applied to market-basket and cross-selling analyses, where they have helped to determine the kind of products that usually sell together. They have also been used in fraud analysis detection, where the associations can reveal unusual relations and coupling of procedures, and in testing the efficacy of medical treatments, where they analyze the results of combinations of multiples procedures and their outcomes.

Inductive techniques also include classification and regression trees (tree-based models) [34]. Two of the more important are the CART (Classification and Regression Trees) and the CHAID (Chi Squared Automatic Interaction Detection). These techniques construct trees based on the patterns or relationships detected [18].

Neural Networks

The neural networks approach includes a series of mathematical models that have the ability to “learn” and thus adapt their actions according to results that have been previously obtained. This technique is based on research in neurophysiology, which studies how the human brain works, replicating it with computers. Neural networks can analyze imprecise, incomplete, and complex information and deduct or find important

relationships or patterns from this information. Usually the patterns involved in this kind of analysis are so intricate that they are not easily detected by humans or by other types of computer-based analysis.

Neural networks are thought to behave as experts do in a specific field. They employ a series of processing nodes in the same way as the neurons work in a human brain. The nodes are interconnected and have the ability to influence each other, and the level or degree of influence can be adjusted according to the circumstances that the nodes encounter. These interactions allow them to respond or react to certain patterns or conditions present within the data of the analysis and they can therefore help to detect or identify other possible relationships.

Although the tree-based models and the neural networks are good for detecting non-linear models, they work in different ways for variable predictions. Tree-based models work better in selecting relevant variables and “work well when many of the predictors are irrelevant” [34]. In contrast, neural networks are better at merging many input parameters, so they work well when there is more redundancy in the predictors of the study.

Data Visualization

Data visualization is also useful for data mining. Through using visual tools, analysts can reach a better understanding of the data because they can focus their attention on some of the patterns found by other method. Using variations of color, dimensions, and depth, it is possible to find new associations and improve the differentiation between them.

Data visualization is a very useful technique for the identification of patterns, relationships, and missing and exceptional values. However, its

greatest limitation is that visualization must collapse many different dimensions into a two- or three-dimensional screen. Moreover, tools developed for data visualization usually require considerable training and are not suitable for people who are colorblind or who have difficulty with spatial analysis [34].

Data Mining Tasks

Data mining can be used in many different ways [18]. Some of the tasks most commonly found are:

- description and summarization
- concept descriptions
- segmentation
- classification and case-based reasoning
- prediction
- dependency analysis.

Description and summarization involves the study of data in order to find its major and most important characteristics. This task enables analysts to better understand the general features of the data and provides an outline of its overall structure. The most common techniques applied for description and summarization are the basic descriptive statistical models and data visualization (histograms, box plots, scatter plots).

The main goal of concept descriptions is to describe data classes or subgroups and to point out important concepts, characteristics and parts that may facilitate the process of understanding them. Clustering and induction methods are usually employed in concept description.

Segmentation is mainly used for sorting data into a series of unknown different classes or subgroups that share the same characteristics, but that are different from each other. The techniques frequently used in segmentation include clustering, neural networks, and data visualization.

Classification is a task that is very similar to segmentation. The major difference between them is that classification assumes classes and subgroups are known. Each class has a class label value that can be discrete or symbolic; it is used to assign all the data elements to a corresponding class. Classification builds models that search all the data and classifies it according to its attributes and class labels. Furthermore, classification can also be used to identify variations or unexpected data attributes. It employs techniques such as discriminant analysis, induction and decision trees, neural networks, and genetic algorithms.

Prediction models try to find or forecast an unknown continuous value corresponding to a specific class. Prediction models are usually built using techniques such as neural networks, regression analyses, regression trees, and genetic algorithms.

Dependency analysis describes all the important and significant dependencies among the data elements. Dependency analysis is used, for example, in shopping basket analysis to study products that are usually sold together. Two special cases of dependency analysis are particularly valuable for data mining: association and sequential patterns. Associations describe and find similarities or events that occur together within the data, while sequential patterns characterize dependencies that occur in the data in a sequential order. Some of the common techniques for dependency analysis include correlation analysis, regression analysis, association rules, bayesian networks, and data visualization.

Traditional Applications

With each day, more and more data mining applications are being discovered and implemented; they are helping many companies to manage and allocate their resources in a more effective and efficient way, reducing costs, and improving the quality of products and services that they offer. One of the more frequently used data mining applications is cross-selling or expanding the products that are being sold to customers [36]. Thanks to data mining, it is possible to identify groups of customers who have a special set of characteristics or preferences and are therefore more likely to respond to some kind of targeted publicity or marketing strategy. Data mining techniques have been used to predict and detect fraudulent transactions or claims, to combine medical procedures that may produce better treatments, and to implement quality control in the manufacture of a variety of products.

Data mining can also establish what motivations or factors influence customer behavior and which groups are more likely to change from one company to another in a given time. This approach has been widely applied, with very good results, to mailing lists, catalogues, and the distribution arrangement of products in stores and supermarkets. For that reason, data mining techniques are being introduced into applications such as customer relationship management (CRM) solutions, which are decision support systems (DSS) capable of managing and studying customers' data, behaviors, and preferences, in order to increase and maximize the profits in businesses [18].

Data mining technologies can also be used for mining web contents and web linkage structures [1], as well as applied in usage mining, to determine possible patterns in the users' accesses in the web logs. This

application is mostly utilized for the identification of possible customers for electronic commerce and to improve quality of service to the end users.

Industrial Engineering Decisions

Industrial engineers focus on the design, improvement, and installation of integrated systems of people, processes, materials, and equipment. As a result, there are many possible applications for data mining techniques in industrial engineering. Industrial engineers must decide and select the most effective ways for an organization to apply the basic factors of production for example, machines, materials, people, processes, information and energy to make or generate products and services. Industrial engineers also plan, design, implement, and manage integrated production and service delivery systems, and make decisions that ensure performance, reliability, maintainability, schedule adherence, and cost control [23].

Data Mining Applications in Industrial Engineering

Because data mining techniques search through large amount of data in order to discover correlations, patterns, rules or relationships, they can be applied in many different fields. Data mining solutions have been focused thus far on applications such as customer retention, customer profile analysis, fraud detection, cross-selling, marketing expansion, medical treatments, and the creation of user access profiles over the internet.

Yet these are not the only fields in which data mining can be applied. Industrial engineers can indeed use data mining to understand complex systems. While the use of data mining in industrial engineering is not widespread, several successful applications of data mining in fields related to

industrial engineering have been reported. Examples of data mining applications in manufacturing processes are the computation of job shop schedules [31], process improvement in circuit and semiconductor manufacturing [24], and the design and selection of new materials[10]. These and other examples are explained in more detail in the following section.

Quality Control

Data mining has been applied in some Statistical Quality Control (SQC) software packages as an integral part of decision support tools used in the analysis of process behavior [11]. These SQC systems are usually employed to analyze data collected by Statistical Process Control (SPC) systems, which monitor production processes in real time through the use of online sensors. SQC is usually applied using statistical techniques also included in data mining, but these techniques are also capable of analyzing parameters, with the same understandable effect on the process of the one given by SPC systems [11].

Additionally, data mining techniques have been applied to analyze and detect possible defects and their corresponding causes in the fabrication of semiconductors [8, 15]. When data mining analysis is applied to data regarding physical and chemical conditions of wafer processing, it is possible to determine whether a defect has been produced in the plasma process for the manufacture of semiconductor wafers.

Data mining has also been applied in predicting defects for the paper-making industry [25]. Prediction models have been developed to predict and avoid deviation and corrugation defects using data mining techniques to generate rules based on process data and fault information from historical records. Furthermore, companies such as Daimler Chrysler have used data

mining to evaluate warranty claims in order to identify high quality patterns, as well as the key factors that give rise to claims, improving customer satisfaction and the reliability of its products [3].

Scheduling

Scheduling has been an important area for applying data mining techniques. For example, schedules for job shop operations have been created using rules extracted with data mining analysis over schedules generated by genetic algorithms [22]. Additionally, quality tests have also been scheduled with a data mining approach. Using decision tree models and mining the data provided by a MRP system in a factory of hydraulic pumps, new and improved schedules have been generated [31].

In general, the number of operators and work stations assigned to a specific order or task could be improved through the use of rules and models generated by data mining applied to historical data such as throughput, operations performance, and completed orders.

Process Optimization

In integrated circuit manufacturing, yield improvement has been considered a suitable application for data mining techniques to address the problem of low yield analysis [24]. In this case, it is possible to analyze data concerning samples of low-yield wafers to identify priorities for process improvement. Furthermore, in a fluid catalytic cranking process, data mining tools have analyzed historical data in order to minimize the time during changeovers. Data mining can also be use to reduce rework in process. For example, in exploring data of a specific production process or product line, it is possible to find rules identifying the best settings and conditions to achieve more throughput, to reduce cost, or to reduce waste.

Process Control

Process control, monitoring, and diagnosis are other important areas in which data mining analysis can be effectively applied. For example, long performance deterioration in processes can be studied using historical data to identify its major factors [35]. Historical process logs can be analyzed to monitor the process at different stages. When monitoring processes, the models created with data mining tools can determine whether or not the current process state can generate satisfactory outcomes; if corrections are needed, then programs can also notify operators or recommend changes.

Safety

Regarding safety, data mining studies in road traffic accidents have already created classification models and identified influential factors for accident severity [30]. Hazardous elements such gasses or radiation levels have also been monitored to protect and extend human lives. Moreover, data mining techniques used to analyze occupational accidents and disease records have revealed important patterns that can be used to reduce occupational risks [4,5].

Cost Reduction

Data mining can be effectively be applied to cost reduction. A good example of an application of data mining techniques to reduce cost in products with high customization, for example, is analyzing sales and product options to identify the ones with greater demand [3]. The products with same and most common options can be manufactured together to reduce cost and inventories.

Maintenance and Reliability

Data mining techniques can also be applied to identify combinations of plants, machines, workstations and products that have higher breakdown or malfunction rates, or to find repairs that are likely to occur together or in close time proximity, or to report problems that often precede specific repairs. They can also be used to create rules and models that identify the source of problems or to identify additional patterns in parts or equipment failures [17]. With this information, preventive maintenance can be performed in parts or components that are identified as having a similar time between failures, reducing downtimes for repairs and their corresponding costs.

Product Development

Other applications of data mining include product development and design [10]. Data mining can be used to extract relationships between design requirements and manufacturing specifications to explore the different tradeoffs between overlapping activities and coordinating costs. Data mining can also be performed in historical data to reduce inventory costs for new products, to analyze suppliers and delivery times, and to select materials for the manufacturing process.

Problems in Making Effective decisions in Data Mining

Some authors, such as Koonce and Fang [21], acknowledge that industrial engineers can use data mining to explain the behavior of complex systems. They also have established the need for further studies related to applications of data mining to job shop scheduling systems [22]. Furthermore, Bertina and Catania also found in their study of data mining applications for wafer manufacturing [8] that it is difficult to select specific data mining techniques. They recognized the need for general methodologies and

guidelines that could support users in the development of data mining applications.

Selection is not the only difficulty with applying data mining. Other problems in making effective decisions are that companies and organizations store great amounts of data and information that are very difficult and time-consuming to analyze by traditional means. Moreover, there are many elements to consider in selecting data mining tools [14]. There are many options, software vendors, and techniques; and is difficult to decide how to chose a data mining tool.

Finally, data mining is the result of the confluence of multiple disciplines [18]. It thus requires much specialized knowledge to make the right decisions. Many different fields (e.g., statistics, databases, artificial intelligence, and software development) contribute to data mining. Figure 2 identifies the variety of knowledge areas required for data mining.

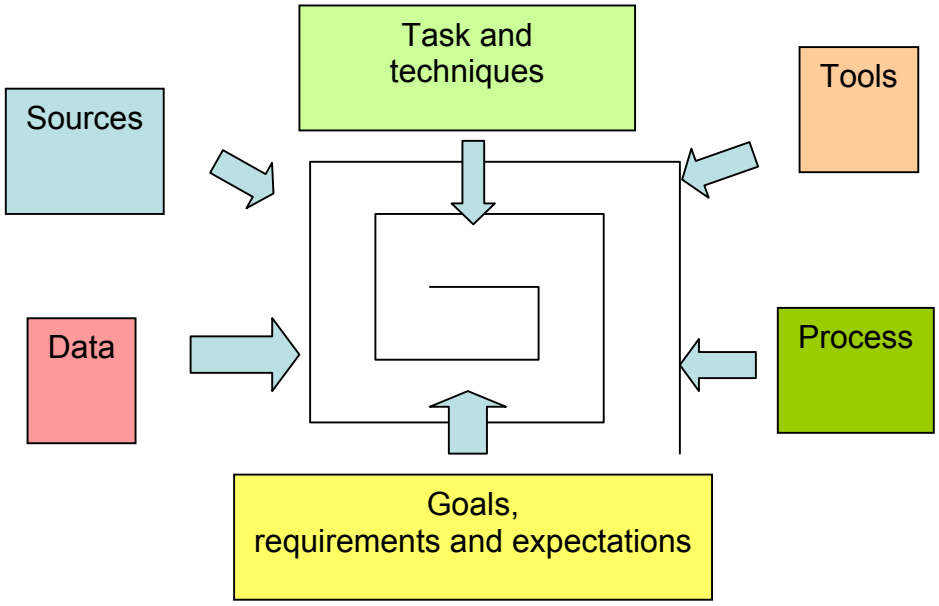


Figure 2. The Data Mining Labyrinth of Knowledge

CHAPTER 3
RESEARCH METHODOLOGY

Introduction

The engineering design process is based on the scientific approach to problem solving. The distinguishing characteristic of engineering, however, is that it uses a systems perspective; that is, it studies a problem environment in order to implement corrective solutions that take the form of new or improved systems. The engineering design process, as described by Landis [23], was used in the execution of this study. This engineering design process is depicted in Figure 3 and its six steps are detailed below.

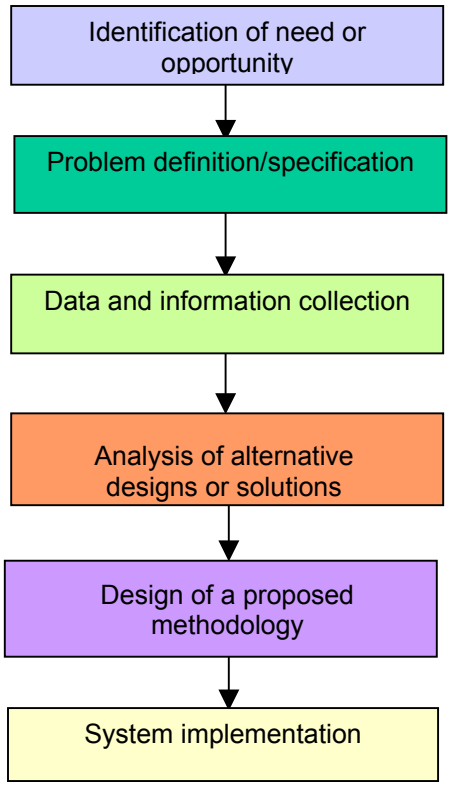


Figure 3. The engineering design process applied

Identification of a need or opportunity

The first step in problem-solving is the identification of a need or opportunity. For industrial engineering, these needs and opportunities are extensive and varied. Industrial engineering is a broad area of specialty among the engineering disciplines. Those sitting for its exam for the Fundamentals of Engineering (FE) Exam need mastery in twenty topical areas. This is at least twice as many as other disciplines. Thus, industrial engineers have a larger-than-average demand for information in doing their jobs. Much of this information is available; the challenge is getting to it. The ideal tool to assist in that effort is data mining.

But while data mining can help industrial engineers process and analyze information, deciding on the most effective data mining techniques and systems can be complicated. As noted above, there are many different software vendors with many different data mining software applications. Each promotes its own data mining methodology. Data mining, like industrial engineering, is the result of the confluence of multiple disciplines and for that reason the process of implementing data mining process in industrial engineering is difficult and requires much specialized knowledge.

Problem Definition

There are so many options, tasks, techniques, tools, formats, and approaches to data mining that industrial engineers find it very difficult to design and implement projects. Although methodologies already exist, they are designed for specific software packages. Most of these methodologies use a traditional statistical approach. It is still not clear that this approach to data mining is sufficient for obtaining the vast array of data needed for

industrial engineering applications. Thus, a data mining methodology to meet the specific requirements of industrial engineering is needed. Such a methodology should assist industrial engineers in selecting appropriate data mining tools and implementing data mining projects from a systems perspective.

Data and information Collection

In order to accomplish this study, surveys, analysis, reviews, and comparisons of data mining applications were conducted. These were based on vendors' information and case studies available in literature and research publications. The survey was sent to more than 80 different companies of data mining software over the Internet. There were 30 responses (see Appendix). The survey asked companies whether their product had been or could be used in industrial engineering applications. It also asked whether they had applied or sold their data mining products for the implementation of projects related to industrial engineering areas such as quality control, scheduling, manufacturing, safety, or ergonomics. Other questions were related to hardware requirements and prices. The most relevant results of this survey are shown in Figures 4 and 5.

Figure 4 shows that approximately 60% of the companies have either sold their product for industrial engineering applications or believe their product is applicable for industrial engineering. The survey also asked about costs, because the cost of data mining may prevent some companies from using it even though it could benefit them. Figure 5 shows that the average cost is approximately $5,000, a figure that might be prohibitive for smaller companies. Thus, design restrictions and cost appear to be key factors that affect the use of data mining in industrial engineering.

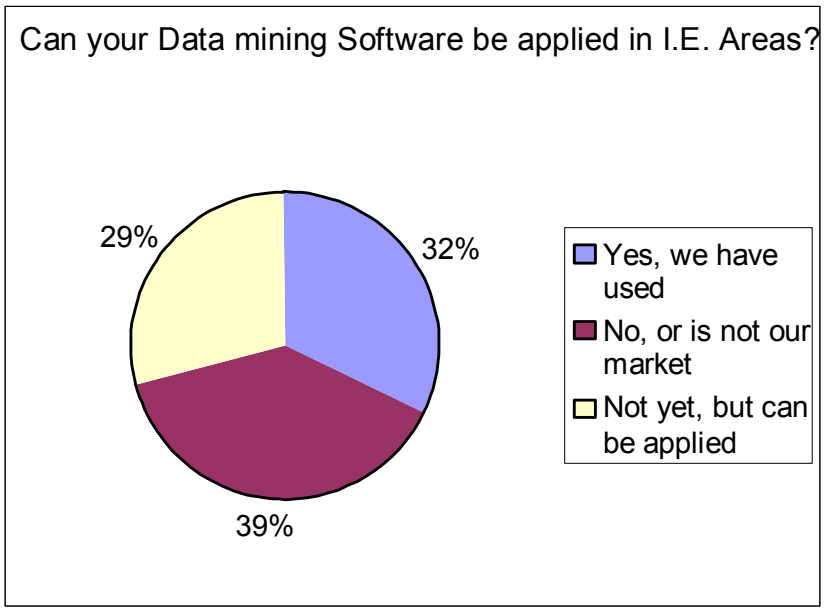


Figure 4. Application of data mining software to IE areas

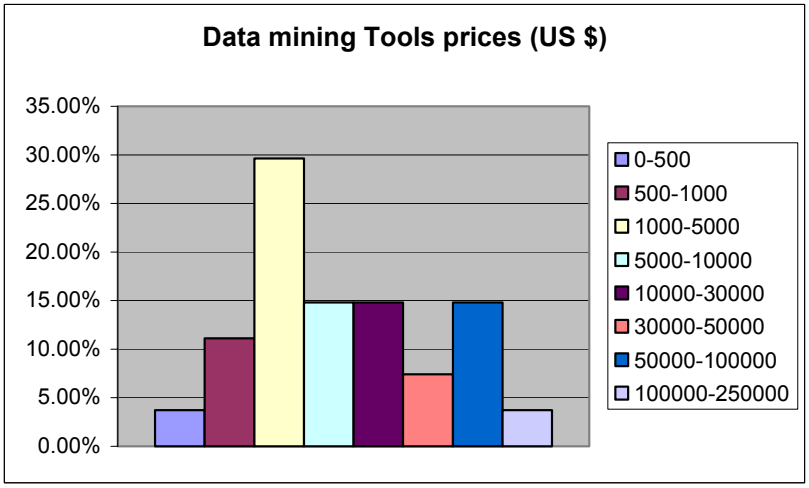


Figure 5. Data mining software price distribution, year 2002

Analysis of Alternatives

There are several different data mining methodologies, but there is no one standard methodology for applying data mining. Consequently, several vendors have created their own proprietary methodologies. These have some drawbacks. Software vendors have designed approaches that are strongly correlated with the design of their own solutions and software packages. A related methodological issue is that data mining has been considered as a kind of art in which each analyst may follow his or her own "recipe" or form [36]. However, this statement is only partially true. While individual preferences and intuition may contribute to ingenious data mining methodologies, at the same time, there are essential steps that data mining methodologies must include and elegant, efficient ways to combine methodological elements to obtain superior results.

Two popular methodologies are SEMMA and CRISP-DM. They are described in the following sections.

SEMMA

SEMMA is the methodology for data mining processes proposed by the SAS Institute--one of the most important companies that develop statistical software applications--with the software package Enterprise Miner [2]. In SEMMA, SAS offers a data mining process that consists of five steps: sample, explore, modify, model, and assess. This methodology begins by analyzing a small portion of a large data set. The next step is to explore the data and the information by looking for trends and anomalies in the data with the purpose of gaining some information about the data. In the third phase, data is modified to create, select, and transform the variables for the study. A valid model is then created using the software tools, which search

automatically for combinations of rules and patterns that reliably predict the observed results. Finally, the last step of the SEMMA methodology consists of evaluating the usefulness and reliability of the findings.

Although the SEMMA methodology contains some of the essential elements of any data-mining project, it concerns only the statistical, the modeling, and the data manipulation parts of the data-mining process. It lacks some of the fundamental parts of any information systems project, including analysis, design, and implementation phases.

But even more important, the SEMMA methodology does not consider the roles of the organization and the stakeholders during the project; it does not see data mining as an integral element within a systems perspective. SEMMA is specifically designed to work with the Enterprise Miner software, the data mining software of the SAS institute. It cannot be applied outside the limitations of that system.

CRISP-DM

Another data mining methodology is CRISP-DM [12] (cross-industry standard process for data mining). CRISP-DM was originally conceived in late 1996, but it was not completed until 1999; it is intended to be industry-, tool-, and application-neutral. It was developed by a consortium of data mining vendors and companies through an effort funded by the European Commission. The four partners of this project were NCR, Daimler Chrysler, OHRA, and Integral Solutions Limited (ISL), which became part of SPSS in 1998.

The CRISP-DM 1.0[12] methodology comprises a hierarchical breakdown in which the data mining process is divided into four levels of

abstraction: phases, generic tasks, specialized tasks, and process instances. CRIPS-DM 1.0 also recognizes four different dimensions of data mining context that drive the generic and specialized levels of the CRISP-DM. The four dimensions are 1) application domain, 2) problem type, 3) technical aspect, and 4) tools and techniques.

While this approach has broader applications than SEMMA, one of its major drawbacks is that it combines tools (software packets) and techniques in the same category. If tools and techniques are combined and selected simultaneously, techniques may be chosen because they are supported by specific data mining tools, and not because they are the most relevant to the purpose of the study, or because they are needed. This also may cause the organization's goals and requirements to be under-analyzed and biases the study. In data mining projects, it is important to analyze the organization's needs, requirements, goals, and strategies. Organizations, for example, may need to extract information from their data warehouses either on a one-time basis or on a recurrent one.

Techniques under CRISP-DM may be applied because they are incorporated in the tools available for the organization and not because they are really needed. Therefore, results from this approach may not properly correspond to the organization's main objectives, and the models generated this way may not truly represent the behavior of the entities for which the study was intended in the first place. This may specificity be true for industrial engineering applications such as those related to quality control, process monitoring, scheduling, process optimization, and many more; but they are not limited to industrial engineering alone and can be applied to many different data mining projects in other fields.

Still, the CRISP-DM methodology is useful. It describes a data-mining project as a 6-phase cycle in which the sequence of phases is not rigid. The phases that CRISP-DM considers are [12] business understanding, data understanding, data preparation, modeling, evaluation, and deployment. This approach includes in its first phase very important elements such the business's objectives, requirements, constraints, and resources available for the project, in order to establish the data mining goals. Good documentation is also promoted from the beginning of the plan.

Other essential elements that are considered are the collection of data during the data understanding phase, but only for analyzing available data (not necessarily the data required). Data is also analyzed and verified in order to ensure that the quality of data will allow the intended results of the modules.

CRISP data preparation comprises the selection, cleaning, construction, integration, and formatting that data requires in order to create any model. However, it also assumes that all the information required is already available and continues to be valid, so new data should not be collected.

Another problem with the approach suggested by CRISP is that the selection of the technique is delayed until the modeling phase; if the data required is not available or is in the wrong format, the model has to return to the data analysis phase again. CRISP-DM, indeed, stresses in the importance of assessing tools and techniques early in the process but also affirms that the selection of tools may influence the entire project [12]. This is a major impediment to data mining for industrial engineering because if the selection of tools influences the complete project, the results, rules, and patterns and predictions obtained with resulting models are not guaranteed to solve the problems that organizations are trying to solve.

Techniques should be selected according to an organization's goals and requirements and should not depend only on the data available. If the data available is not enough for the organization to perform a data mining project, new data and information should be collected; otherwise, the selection of the technique required may be influenced by only the data in existence, so the resulting models may be biased and will not correspond with the organization's actual intentions. A related problem is that although assumptions are clearly declared, they may not be sufficiently revised. Changes in data from the past to the future may cause assumptions about data that were once valid to be incorrect.

CRISP-DM methodology and other approaches also emphasize that, according to the technique selected, data must sometimes be divided into training and validation sets. After building the models with the training data, a validation test is then used to ensure that the obtained model behaves with adequate fidelity to the real system.

Finally, in CRISP methodology, after a satisfactory creation of models is done, the evaluation phase continues with an analysis of the results, a review of the process, and the final deployment phases. The CRISP deployment phase consists in the creation of a deployment plan, a monitoring and maintenance plan, a final report, and the final review of the project. Besides the difficulties that the CRISP-DM methodology presents, it is a good approach to the general process of data mining and the data mining cycle.

Design of a Proposed Methodology

While the two major data mining methodologies are useful in their ways, they may not be the most useful methodologies for industrial engineering purposes. Through the revision of the data mining methodologies

discussed in the previous section and the application of information systems analysis and design structure, a proposed methodology for the application of data mining in industrial engineering is described in the next chapter.

CHAPTER 4
A PROPOSED METHODOLOGY

Introduction

Engineers follow a structured approach to problem-solving. This enables them to duplicate results or determine where errors have occurred in the process. As a result, they may have confidence in the solutions they recommend. For that reason, this study offers a methodology for using data mining in solving problems related to industrial engineering. This structured approach should lead analysts through the steps required in obtaining the data needed to provide information required for problem-solving. This approach has a number of steps:

1. Analyze the organization.
2. Structure the work.
3. Develop the data model.
4. Implement the model.
5. Establish on-going support.

These steps are shown in Figure 6 and described in detail in the following sections.

Analyze the Organization

Organization Description

The first step of any data mining project is to understand the purpose of its existence. When a data mining project is conducted in an organization,

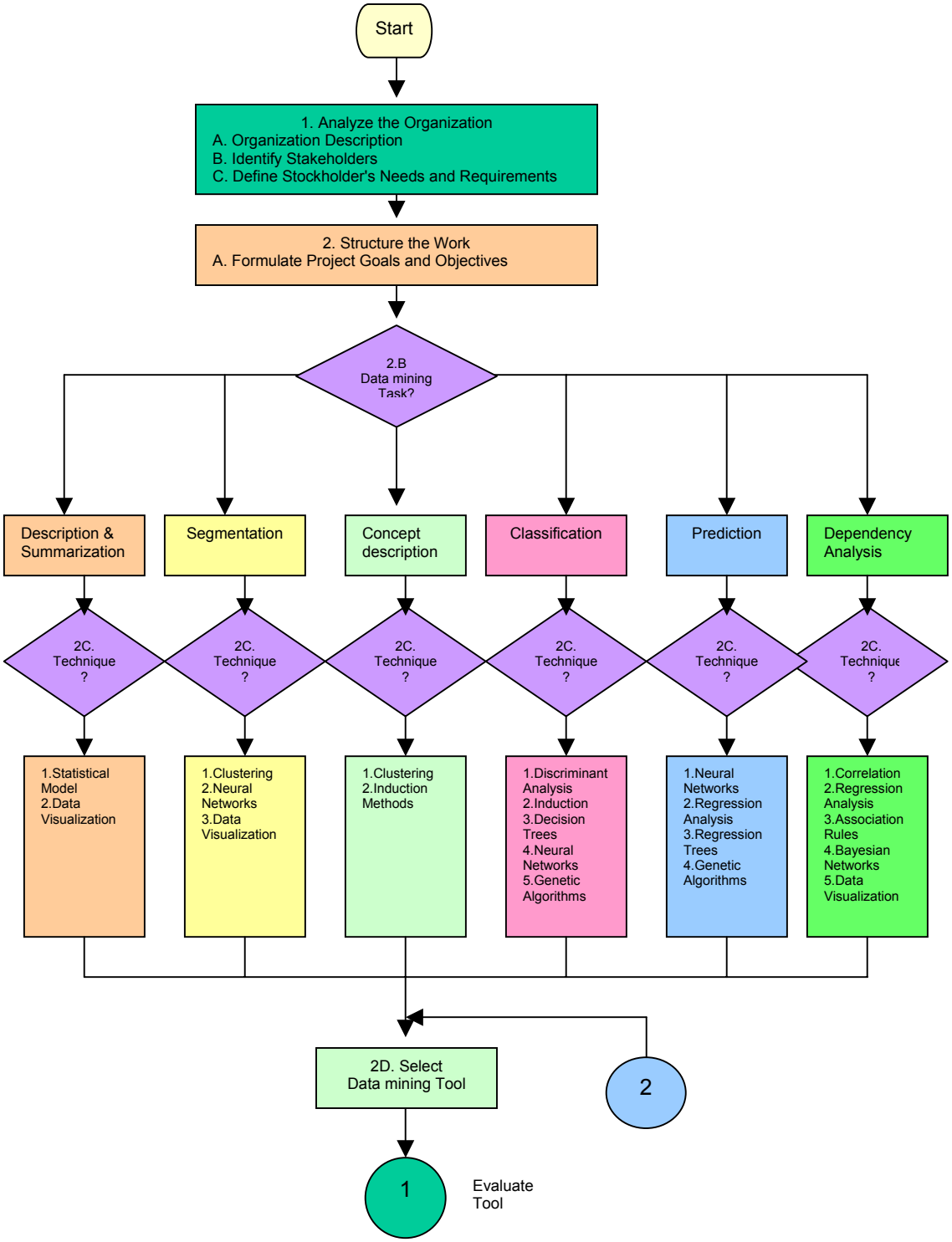


Figure 6. Proposed Methodology.

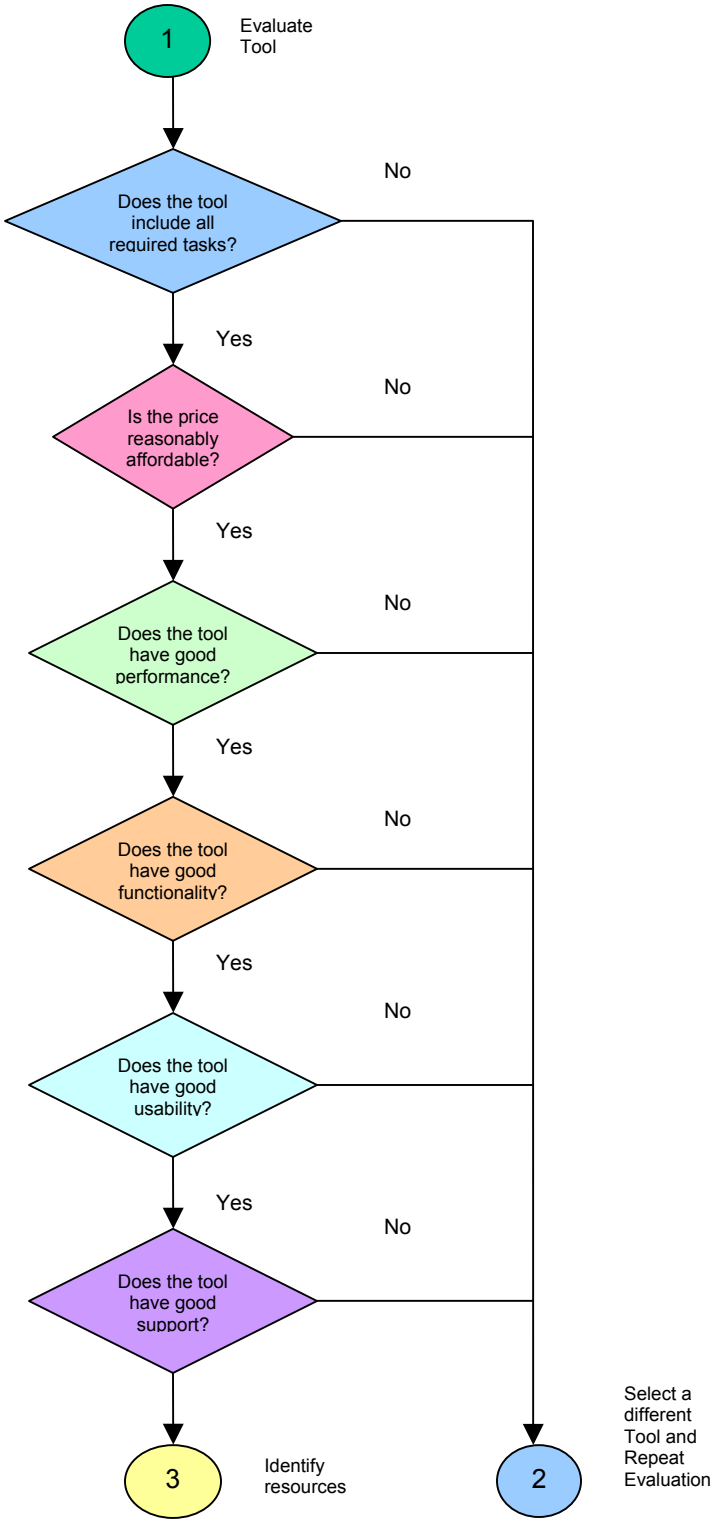


Figure 6 Continued

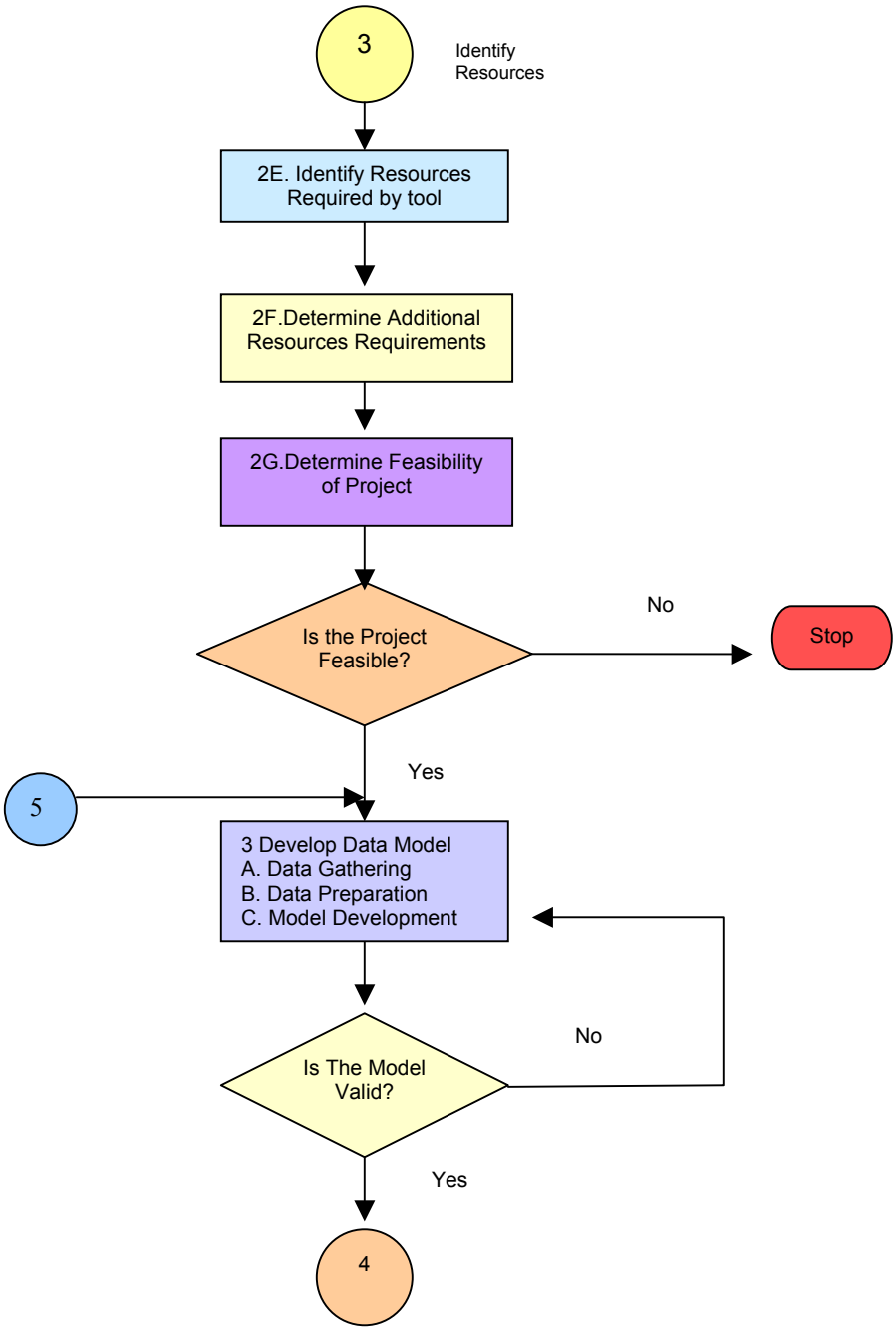


Figure 6 Continued

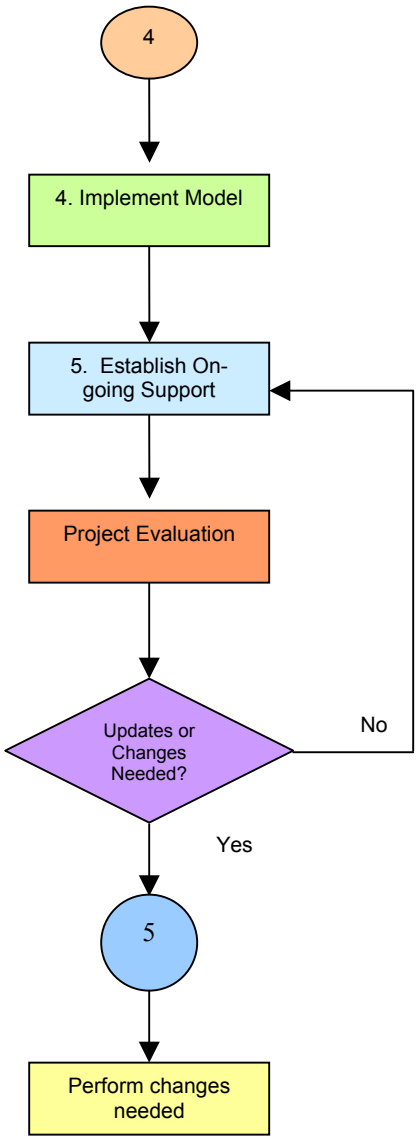


Figure 6 Continued

a study of the organization's goals, objectives, and strategies is required in order to understand the purpose of the project. This enables the analyst to determine the best way to execute the project so that it will empower and facilitate the achievement of the business's targets. Failing to understand the organization's needs before implementing the project may cause its results to be incompatible with or of no use at all to the organization.

However, understanding the main goals and guidelines of an organization is not enough. Because of the broad scope that data mining encompasses, a data mining project must be specifically defined and understood on its own terms; otherwise it is too easy to become lost in the infinite numbers, options, models, and results. Data mining projects, then, must be consistent with the business strategy of an organization and internally consistent as well.

Identify Stakeholders

To successfully implement a data mining project, it is very important to identify all the key elements involved in it, and stakeholders are an essential element in any information system analysis and design task. Stakeholders include all the major owners, users, analysts, designers, and developers of an information system, as well as the essential personnel on which the successful implementation of the project will depend. Identifying the stakeholders and their requirements will allow analysts to completely recognize the critical elements of the project, together with its true intentions and expected results.

Define Stakeholders' Requirements and Expectations

Before identifying the requirements and defining the goals and objectives of a complete data mining project, the requirements and

expectations of the stakeholders must be recognized. It's a well-known fact that the successful implementation of any information project depends in great part on the direct involvement of the staff and stakeholders, the compromises that they develop for the project, and the satisfactions of their own expectations. If users and stakeholders do not believe in the project's results, it is likely that models, patterns, or relationships will not be applied or implemented.

Structure the Work

Formulate Project Goals and Objectives

The goals and objectives of a data mining project must be clear and specific; they must be completely understood by all the participants involved. These goals and objectives are also dynamic because they must correspond to the needs and requirements of the business, factory, or organization. Goals and objectives should be periodically revised, updated, and they must be defined within a timeframe that corresponds to the business's or organization's perspective; otherwise, the data, rules, models and relationships structured and defined by the project will be outdated and useless.

The necessity for clear but dynamic goals corresponds to the way that products, business, organizations, and processes evolve. New markets, new customers, new products, and new processes may require different data, different tasks, or different tools. Any data mining project that tries to produce useful results under these conditions must account for change—an essential factor that cannot be ignored, no matter how inconvenient (see Figure 7).

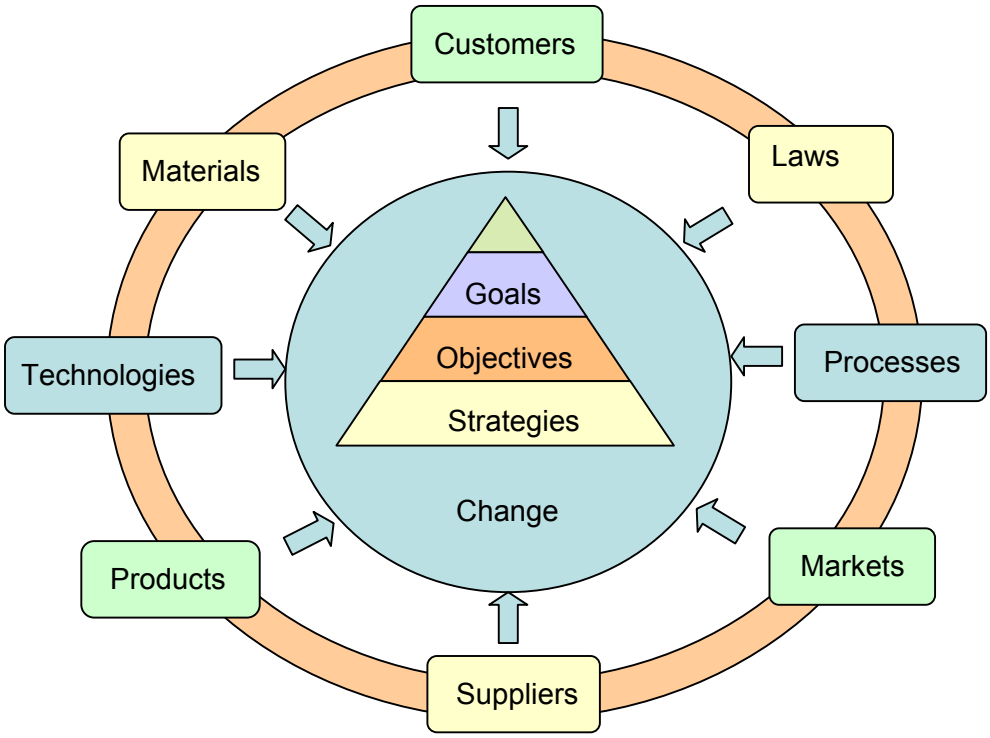


Figure 7. Factors of Change in Data Mining Projects

One way to account for change is to maintain a Gantt chart, which is a valuable tool that can be used to plan and schedule data mining projects. A Gantt chart represents project tasks as horizontal bars, using a calendar time line [37]. They can be easily prepared; they are easy to update, easy to understand, and are very useful in evaluating a project's progress.

Select Task, Techniques and Tools for the Project

Once the goals and objectives of the project have been defined, it is possible select the appropriate data mining tasks, techniques, and tools. However, the selection of tasks must depend primarily on the goals of the project, rather than solely on the techniques and tools available. It is unwise

to select the tasks after the selection of the techniques and tools because both tools and techniques may influence or limit the data mining tasks.

The selection of data mining techniques and tools must also correspond with the goals of the data mining project. In this thesis, data mining tools are defined as the specific software packages and solutions presently offered by a large number of different vendors to implement data mining projects. It is important to keep in mind that selecting tools only for convenience can severely limit the boundaries of the project.

Data mining techniques, moreover, must be chosen before tools are chosen, to avoid applying techniques that do not correspond with the real goals of the study. The traditional data mining approach has been to consider several techniques that are usually applied in order to find the one that fits the best. As discussed above, certain guidelines can be used for selecting appropriate techniques in data mining projects. For example, decision trees are useful because of the following characteristics: they are easy to manage and understand, they can work with categorical and numeric data, they are not affected by extremes values, they can work with missing data, they are able to reveal complex interactions and a lack of linear relationships, they are good at handling noise in data, and they can processing large data sets. However, decision trees also present some disadvantages. For continuous variables or multiple regressions, the use of many rules is usually required, and small changes in the data may generate considerably different tree structures.

Therefore statistical models may be sometimes preferred, because they are the only tools that can give an estimation of their own accuracy, and they use mathematics to obtain the best methods under the specified restrictions of the problems. Statistical models can also be very fast; they

provide models in which new data is easy to apply, and they are usually better for predictions for values outside the range of the data analyzed. Unfortunately, some statistical models require a great deal of statistical knowledge, so they are not easy to use, explain, or apply correctly for the common user. In addition, the statistical significance of the results may not imply a practical use.

Neural networks also have good advantages; their models, for example, are usually considered easy to use, and because they are universal approximations, it is possible to apply them to model a wide range of relationships or patterns. Nevertheless, because of the complexity of the neural network patterns they create, their models frequently are difficult to understand, they may require a lot of time to process large data sets, and they cannot be implemented in different software packages without difficulty [9].

Genetic algorithms are good for handling noise in data because noise usually behaves as an occasional mutation element rather than as a dominant factor. Genetic algorithms can also process efficiently large data sets; moreover, they are easy to understand and to integrate [9]. Unfortunately, however, operating genetic algorithms can be difficult because they may require specialized knowledge and they are not available in all data mining software packages.

Currently, data mining software packages are widely available; these correspond to a large selection of suppliers existing in the market. For that reason, in order to select the best data mining tool for the given conditions of a specific project or study, important features should be evaluated to determine whether they correspond with the requirements of the project.

Although several authors have suggested a number of features that should be analyzed when selecting a data mining package, they also have concluded, based in their own experience that “there is no one best data-mining tool for all purposes” [14]. This fact implies that when a significant change factor interacts with the project, a new evaluation of task and techniques may be also required.

Some of the most important characteristics to keep in mind when selecting data mining tools are presented in Figure 8. Considering the task involved in the project is an essential factor in selecting the best tool. First, it is very important to determine whether the software will be used for a specific type of project or used in a variety of different studies with multiple characteristics and requirements. If a data mining tool will be applied to a specific set of conditions, the evaluation of the tool should concern those conditions; additional features may be desirable but not required. Purchasing a tool that has unused features will be a waste of resources.

But if data mining tools will be employed in a variety of studies, with different types of data sets, in different formats, and involving resources with different infrastructures, wide selection criteria are needed. The organization should, under these circumstances, purchase the best tool they can afford that meets all of their requirements.

Price is another important element, especially with the wide range of options available in the market. One of the issues of data mining software is that it can be very expensive, while the returns on projects may be difficult to quantify or require a considerable amount of time before they can be precisely measured.

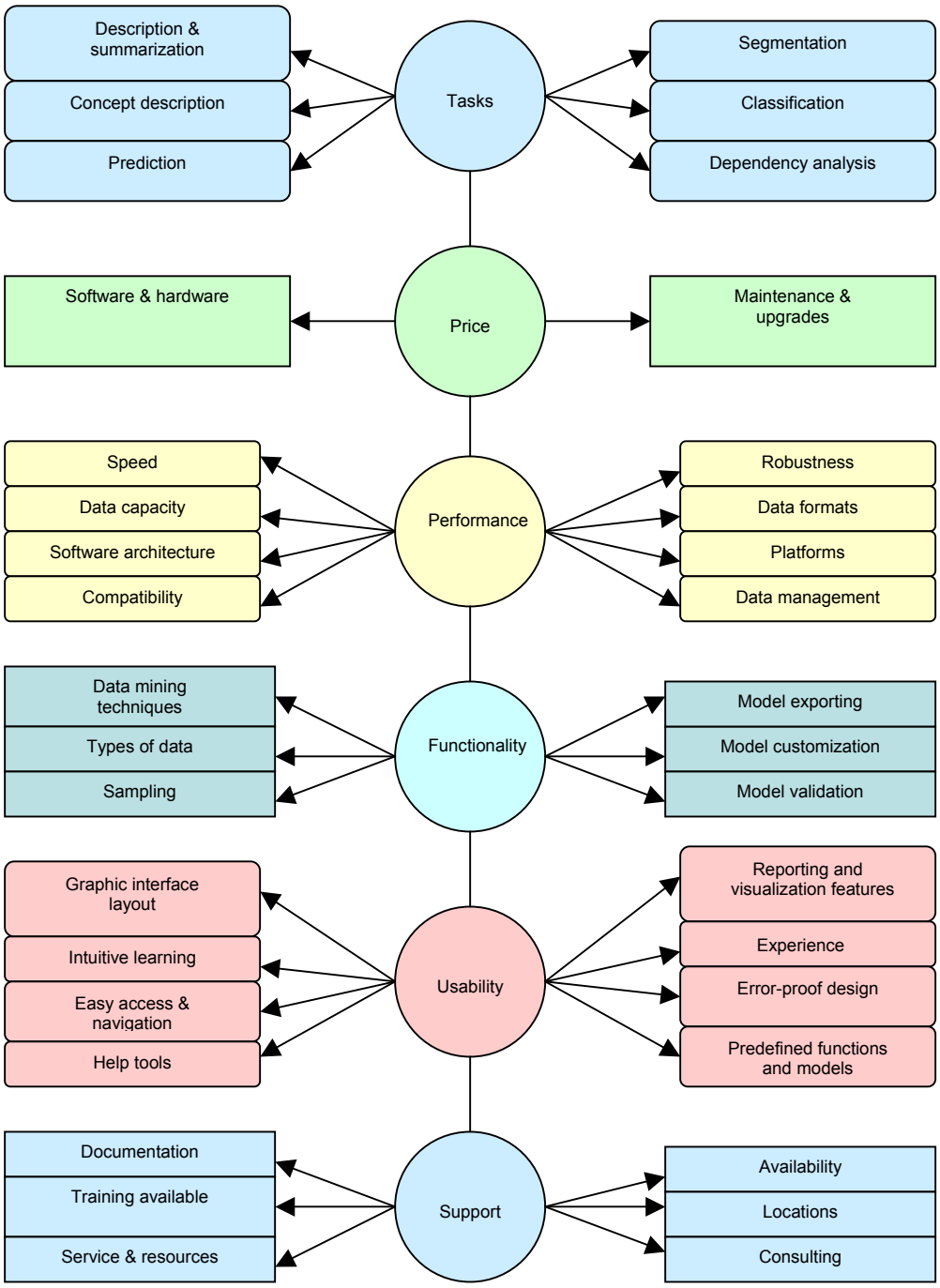


Figure 8. Proposed Factors for Selection of Data Mining Tools.

Yet the initial price of the software is not the only consideration. Because data mining tools are software-based, they greatly depend on technology and its development. Data mining tools are thus likely to be outdated very quickly, with new versions available in the market almost every year. Moreover, new versions may also require new hardware in order to be successfully applied, and they may not always be 100 % compatible with old technologies. For that reason, a comparison analysis between the different applications would identify which of the different solutions are more appropriate for a particular study. In order to perform this analysis, the total investment cost and all the consequent expenses of each of the different packages should be estimated. Regrettably, for most data mining projects, only the initial price is considered, while other significant expenses, which play a role in any project, are ignored.

When determining the cost of a data mining solution, it is essential to include the potential costs of elements such as software for all applications and licenses; all the equipment required; installation; personnel and staff training; maintenance and support; the future cost of upgrading software and hardware; and any other cost that may be incurred during the project. After estimating all the possible costs, the alternatives, which are in most cases mutually exclusive, can be evaluated using equivalent worth methods such as present worth (PW), annual worth, (AW) or future worth (FW) [33].

If the benefits are determined to be equal for all the alternatives, only costs should be compared. However, if benefits are not equal, the difference between the benefit and the costs should be compared. Furthermore, to find the present worth of cost and benefits, a nominal (market) interest rate should be chosen as the minimum attractive rate of return (MARR) for which the project is a good investment from the point of view of the organization.

The present [33] worth for the costs can be calculated by the following equation:

$$\text{PW (Cost)} = \sum_{k=0}^{n} Cost_k (1+i)^{-k}$$

where *i*=effective interest rate or MARR, per compounding period;
k = index for each compounding period;
Cost $_k$=amount estimated of expenses at the end of period K;
n = number of compounding periods in the planning horizon.

The best alternative in this case will be the one with a lower present worth value--in other words, the lower cost.

If benefits are not equal for all the alternatives, present worth values should be calculated for the difference between the benefits and costs of each alternative. If the benefits are greater than the cost, the alternative may be viable and the best alternative will be the one with a greater net value. If costs are greater than the benefits, a detailed economic feasibility analysis should be conducted in order to determine if the project's value is enough to justify its completion.

$$\text{PW (Net)} = \sum_{k=0}^{n} (Benefits_k - Cost_k)(1+i)^{-k}$$

where *i*=effective interest rate or MARR, per compounding period;
k = index for each compounding period;
Benefits $_k$ = amount estimated of benefits or incomes at the end of period K;

Cost $_k$=amount estimated of expenses at the end of period K;
n = number of compounding periods in the planning horizon.

Unfortunately, because the useful life of a given software package is difficult to estimate, care must be taken when selecting study periods. Moreover, in data mining projects, the repeatability assumption may not be an adequate method. Because software life cycles are becoming shorter [33], its value can be affected by new version releases, and most of the software assets do not have market value at the end of the useful life. Thus, the co-terminated assumption can be used in which, for all the investment alternatives whose useful lives are less than the study period, all the cash flows are reinvested at the MARR until the end of the study period [33].

The future worth method is then calculated using the equation [33]:

$$\text{FW (Net)} = \left(\sum_{k=0}^{d} \left(Benefits_k - Cost_k \right) (1+i)^{d-k} \right) (1+i)^{n-d}$$

where *i*= effective interest rate or MARR, per compounding period;
k = index for each compounding period;
Benefits $_k$ = amount estimated of benefits or incomes at the end of period K.
Cost $_k$=amount estimated of expenses at the end of period K.
d = useful life of the alternative.
n = number of compounding periods in the study period.

Thus, the alternative with the greater positive net future value will be the best from an economic point of view. However, sometimes a useful life is not easy to determine for all data mining projects. In those cases, a probabilistic approach can be used. A discrete probability distribution can be estimated for the useful life on a given alternative. Probabilities of various

useful life values can be projected; keeping in mind that the summation of the probability values for all the possible useful lives must be equal to 1. With these probabilities, expected values, variance, and standard deviation of the present worth values can be found by using the following equations [33]:

Expected (PW)

$$E(PW) = \sum_{j=0}^{l}\left(\left(\sum_{k=0}^{n}\left(Benefits_k - Cost_k\right)(1+i)^{-k}\right)_j \times p_j\right)$$

where $\sum_{j=0}^{l} p_j = 1$;

i =effective interest rate or MARR, per compounding period;

l =number of the possible useful lives for a given alternative;

j = index for each useful life;

p_j =probability associate with a useful life value;

k = index for each compounding period;

Benefits $_k$ = amount estimated of benefits or incomes at the end of period K;

Cost $_k$=amount estimated of expenses at the end of period K;

n = number of compounding periods in the study period.

Variance (PW)

$V(PW) = E[(PW)^2]-[E(PW)]^2$

$$= \sum_{j=0}^{l}\left(\left(\sum_{k=0}^{n}\left(Benefits_k - Cost_k\right)(1+i)^{-k}\right)^2{}_j \times p_j\right) - \left(\sum_{j=0}^{l}\left(\left(\sum_{k=0}^{n}\left(Benefits_k - Cost_k\right)(1+i)^{-k}\right)_j \times p_j\right)\right)^2$$

and the standard deviation:

$SD(PW) = [V(PW)]^{1/2}$

In order to select the best alternative in this case, the expected values, variance, and standard deviation of all the alternatives should be compared, the best candidate being the one with the greatest positive expected value and very low variance and standard deviation values.

Additionally, sensitivity analysis is another suitable method that can be employed to select the best alternative when considering different useful lives for a project. Sensitivity analysis is a good method to apply when considering risk and uncertainty in decision-making activities for projects. Both risk and uncertainty are caused by the lack of precise knowledge about the future, and the main idea behind sensitivity analysis is to determine the degree to which changes in a given factor or estimate would affect capital investment decision [33].

Sensitivity analysis, in fact, is not only used to analyze the effect in variations of useful lives in projects, but also to study the effects of selling prices, capacity utilization, and other combinations of factors. The most common techniques in sensitivity analysis are breakeven analysis, sensitivity graphs, and combinations of different factors.

Breakeven analysis tries to find the breakeven point among different alternatives. This point represents the value that designates that alternatives are equally attractive. Breakeven points are usually found by solving the following equations [33]:

$$EW_1 = f_1(y) = EW_2 = f_2(y) = \ldots = EW_n(y)$$

Where EW_f is the equivalent worth of the alternative f (usually annual worth) and y is the common factor in which a breakeven point is intended to be found. Here it is important to notice that the complexity of the solution increases according with the number of alternatives considered in the study

The sensitivity graph, on the other hand, is usually applied in cases where breakeven analysis cannot be used. This technique consists in plotting

the different results such as the PW value, obtained by using the estimates available for each of the factors analyzed [33].

Finally, the “combinations of factor method” analyzes the combined effect of several factors due to uncertainty. This method selects the most sensitive factors in a project using sensitivity graphs, then scenarios—such as optimistic, most likely, and pessimistic--are created using combinations of levels with these factors. The impact of the factors in each scenario is evaluated by comparing PW, AW or FW values obtained in each scenario and then the best option is selected [33].

A major difficulty in these methods is that for several alternatives, several alternatives, several scenarios, and several factors, the computational analysis required may comprehend a considerable amount of effort. Consequently, if useful life is considered as the main factor for data mining projects, a sensitivity analysis can be obtained by comparing the FW values of the alternatives. Thus, it is possible to compare the values obtained using different estimates of d, with the same value of n such as:

$$FW_p\ (\text{Net}) = \left(\sum_{k=0}^{d_p} (Benefits_k - Cost_k)(1+i)^{d_p - k} \right) (1+i)^{n-d_p}$$

Where $d_p \leq n$, d_p are the values estimated of useful lives for all the alternatives, n is the number of compounding periods, and there are p different values of useful lives in the study. FW values then can be compared within each alternative, and the best alternative with the greater FW values should be selected.

Besides price, another important element to analyze for the selection of data mining tools is performance. Performance tries to measure the ability of data mining tool to handle and perform efficiently all of its correspondent

tasks. It can be established by the amount of time required for a specific task given a selected set of conditions, but it can also include such elements as robustness, data-size capacity, data formats, software architecture, platforms of operation, and compatibility with other software packages.

One way to evaluate and compare the features corresponding to each of the available data mining tools is a decision matrix [14] in which all the most important and relevant characteristics are selected and then weighed and measured corresponding with the preferences for and requirements of the project, the stakeholders, and the organization. Performance benchmarking analysis among different software solutions can also be included in this approach.

However, it is important to keep in mind that many of the data sets available for software demonstrations are prepared and specifically designed for a given set of products and conditions controlled by the software vendors, and are rarely compared with real data. If this is the case, it is a better option to prepare a specific data sample containing actual (or very similar) data from the company's current applications and compare all the performance of all solutions with it, instead of relying only on product descriptions provided by the software vendors.

Functionality measures the ability of software to work under different sets of circumstances. It involves elements such as the number of different data mining techniques included, the degree of customization of models and algorithms that the package supports, and the different types of data that can be utilized. Reporting capabilities, sampling techniques, model validation and model exporting are also other features that can be analyzed under the category of functionality [14]. These may be helpful in measuring the degree of flexibility that a given data mining solution really provides, particularly with

cases in which the data mining tools must be employed in a wide spectrum of studies.

The usability of the software package is another important element that must be analyzed when selecting a data mining tool. All data mining tools must be easy to learn, understand, and use, so that they may be applied effectively. Selecting an otherwise excellent package that is very difficult to use or figure out may risk acceptance of the results, may create more resistance from the point of view of the users, and may cause the project to fail. The users should be confident and understand what they are doing; otherwise the probability of errors dramatically increases, and nobody in the organization will believe in the results. Usability depends on several factors such as the graphic interfaces, access and navigation features, learning curves, experience required, help tools, reporting and visualization features, and predefined functions and models. All of these fundamentals must to be consistent with the main purpose of the project and should effectively contribute solving any difficulties that arise during its execution and implementation phases.

Finally, but not least, support is an essential issue of data mining applications, one that must be studied in detail. A good data mining application without convenient service support may cause considerable amounts of time and resources to be diverted to solve unexpected problems, conflicts or misunderstandings. This condition may influence not only the execution and completion of the project's schedule, but also may involve a far greater additional investment of assets.

Support can be measured by several different factors, such as the documentation provided by the vendors, the time available for inquiries and conflict solutions, the vendors' services and resources available for customer

support, locations, the training available and offered to the users, and consulting services for future projects or expansions. Although these elements are among the most important to consider when selecting data mining tools and applications, these are not the only ones, and in many cases not all of them are required.

The importance and weight of each factor should be determined and measured according to the specific conditions of each project; moreover, they must periodically be revised as a response to the dynamic conditions that influence the software market (see Figure 9).

Identify Resources Required: Hardware, Software, Data, and Personnel.

Once the goals and objectives for the project have been defined and the task, techniques and tools have been selected, it is essential to identify all

Criteria	Weight	Score	Total
Tasks			
Price			
Performance			
Usability			
Support			
Total	1		

1-10

Total Score for Alternative

Figure 9. Decision Matrix for the selection of Tools

the resources required for the project. Data mining projects may involve many resources, which may be classified into four types: software, hardware, data and personnel.

Many of the companies and institutions willing to implement a data mining project may already have an existing infrastructure of resources available. Organizations may already own, for example, networks, servers, database management systems, data repositories, or data warehouses that can be employed in the project. Data analysts, server administrators, and technicians working in the company can also be very useful, even if they are not directly considered stakeholders themselves; they can provide an invaluable source of information thanks to their knowledge and experience with current systems.

Identifying all these resources is vitally important to determine their accessibility, functions, and involvement in new data mining projects. Unfortunately, although institutions may have already acquired these types of resources, they may be currently assigned to other different projects in execution or they can be unavailable during the implementation of the project. A careful evaluation of resource capacity and availability should be conducted in order to determine possible involvement in the project.

Identify Additional Resources Requirements

Once it has been established what resources are available, an estimation of the remaining resources required for the project should be conducted. This study must include all the software, hardware, personnel and data that are required and are not already available in the organization. Here,

special care should be given to determine the amount, type, and format of information that are needed. If the information necessary for conducting a data mining project is not completely available, the results of the study will be misleading.

The quality of model built with a data mining study depends on the quality of the information on which it has been based. If this information contains large numbers of errors and inconsistencies, the inconsistencies and the errors may be also be shown in the results predicted by the model. Information also may become outdated because of changes in processes, workstations, operations and products, so models may produce predictions that are no longer valid.

For that reason, although data may be already available for the project, this data may not be consistent enough to create adequate rules, patterns and relationships and new data should be collected. In those cases a data recollection process should be conducted before data preparation can be initiated.

Determine Feasibility of Project

Once the resources necessary for the project have been identified, a feasibility analysis is essential to determine whether the project is viable. Feasibility analyses can be divided into four major sections: operational, technical, schedule, and economic.

The operational feasibility analysis determines whether the project can work, as well as whether it would be accepted in the organization. It can be performed by surveying end users to establish how they feel about the project, identify their concerns, and ascertain how they would react once the models were implemented. Technical feasibility, in contrast, concerns the availability of the technology required to implement the project. Although data

mining algorithms are continually evolving, it may be that for a specific set of conditions, the existing applications or solutions are not enough and new applications must be developed first.

The schedule feasibility analysis determines whether the project can be successfully completed within a desirable or required timeframe. For certain projects, for example, the amount of data available would not be sufficient, and more time would be required in order to collect more information before the models can be successfully developed.

In other cases, by the time the project has being successfully implemented; changes in processes, products, materials or workstations can cause the model to be no longer necessary or valid. For that reason, if unexpected changes occur during the development phases of the project, schedule feasibility should be updated in order to guarantee that the study will generate the expected benefits.

An economic feasibility study [37] involves determining whether the benefits generated by the project are economically attractive enough to make it worth implementing the project. One approach to evaluate the economic feasibility of data mining projects is to use a benefit/cost ratio method. While costs in data mining projects are relatively easy to estimate, benefits may not be as evident or easy to calculate. This is because favorable outcomes may have unexpected impacts that are difficult to quantify and mostly depend on the type of organization, department, or section in which they are implemented. The benefit-cost ratio [33] is determined by calculating the ratio between the benefits and the cost of the project.

Benefit/ Cost = $\frac{Benefits}{Cost} = \frac{PW(B)}{PW(Total\_Cost)}$

$= \frac{PW(B)}{Investment + PW(O\&M)}$

where PW = present worth

B = benefit

O&M = Operation and maintenance costs.

A variation of the benefit-cost analysis is the modified benefit-cost ratio [32]:

Modified Benefit/Cost= $\frac{PW(B) - PW(O\&M)}{Investment}$

In order to perform an economic feasibility study, both the benefits and the cost should be estimated as well as possible. If the benefit cost ratio is equal to or greater than 1, the project is economically attractive [33].

As mentioned above, to find the present worth of cost and benefits, a nominal (market) interest rate should be chosen as the minimum attractive rate of return (MARR) for which the project is a good investment from the point of view of the organization.

The present worth for the benefits can be calculated by the equation [33]:

PW (B) = $\sum_{k=0}^{n} Benefits_k (1+i)^{-k}$

where *i*=effective interest rate or MARR, per compounding period.

k = index for each compounding period.

Benefit $_k$=amount estimated of benefit at the end of period K.

n = number of compounding periods in the planning horizon.

In a similar way, the present worth for the costs can be calculated by the following equation:

$$\text{PW (O\& M)} = \sum_{k=0}^{n} O\,\&\,M_k\,(1+i)^{-k}$$

where *i*=effective interest rate or MARR, per compounding period.

k = index for each compounding period.

$O\&M_k$=amount estimated of operation and maintenance costs at the end of period K.

n = number of compounding periods in the planning horizon.

It is also important to remember that data mining costs can include several different categories. These include the following:

- Software
- Hardware
- Installation
- Training
- Maintenance and support costs
- Consulting and outsourcing

Benefits from a data mining project vary according to the goal, the strategies, and the type of study. Some of the benefits that a data mining project may represent are:

- Increase in productivity.
- Reduction of cost.
- Increase in product quality.
- Increase in personnel safety.
- Process improvement.

- Waste reduction.
- Reduction in production times.
- Increase in sales.
- Improvement in design of new products.

Develop Data Model

The creation and development of a data mining model is another important step in a data mining project. Data mining models can be automatically produced by data mining tools or programmed using the rules, patterns, or relationships that the tool discovers. Not all data mining projects require the creation of a model. In some cases, the information provided by a data mining tool is good enough to be used alone, to implement changes in a manufacturing process for example, or to select a specific combination of variables and materials. The following section describes the major phases that must be performed for the development of a data mining model; the physical development of the model in many cases is optional and depends on the requirements of the organization.

Data Gathering

As discussed above, many data mining studies assume that required information is already available; unfortunately, that is not always the case. Information is a dynamic asset which changes in time. Products, processes, operators, regulations, services, costumers, suppliers and materials are dynamic factors that frequently change. And so does the information concerning them.

When identifying sources of information for data mining projects, it is important to consider essential aspects of information such as [34]:

- Owners
- Persons responsible
- Formats available
- Cost of retrieval
- Size
- Security requirements
- Privacy

Privacy and security are special issues that must be managed with extreme care. Regulations can restrict the use of certain data or require special authorization in order to use it.

As a good data documentation strategy, elements such the number of fields and columns, data types, data descriptions, units of measurement, ranges of values, and primary keys must be gathered. They can be implemented as data description reports [34] and would prove very helpful in understanding the relationships and patterns uncovered by data mining tools and the models generated with them.

Additionally, in order to perform a good data mining analysis, the information available must be as good as possible. In other words, errors, missing values, and other noise factors should be avoided. And what better place to do it than directly from the sources? Data sets can be cleaned in data warehouses after it has been collected, but this process also implies the loss of information, which can help create better rules and models.

The same happens when data is summarized. Most of the information gathered when data is summarized is only good for answering the initial questions for which it was initially requested, while more detailed information can help to answer new or unexpected queries. In order to enhance

performance, summarization can be done, as long as data is kept at a good level of detail [6].

Once the necessary data and information have been identified, suitable and uncorrupted sources are then selected and the recollection methods are defined. Data must be stored in a consistent and adequate format that facilitates the analysis, interpretation, integration and retrieval operations. Sources of information can be forms, surveys, reports, or information available in servers of operational databases, external data marts, or repositories, which can be generally accessed by SQL code executed at the servers and generated by application program interfaces or gateways.

Examples of getaways include ODBC (Open Database Connection), OLE-DB (Open Linking and Embedding for Databases), and JDBC (Java Database Connection) [18].The main features of a gateway such as ODBC, for example, are to set up a connection to a database server, send SQL statements, receive results of transactions, and handle errors [39].

Procedures for updating may also be required according with the frequency in which the data is modified. Documentation should also be provided to successfully understand the overall design and data structures, so future modifications can be easily implemented. Furthermore, data elements such as metadata or “data about data,” are currently being included to define data warehouse objects such as names, definitions, and data management operations [18].

Unfortunately, if organizations are already working with data warehouses, performing data mining directly in them is not always the best option. Centralized data warehouses are suitable for performing cross functional analysis on complete data [6]. But when executing data mining

tasks, the overall normal performance of data warehouses can be severely affected by the complexity of the data analysis. For large amounts of data, normal queries can take several hours or even longer; additionally, operational databases, and other applications are usually designed to efficiently perform specific series of tasks, with a specific workload level. Increasing the workload levels will then substantially affect the performance of the whole system.

At the beginning of the data mining process, data mining techniques also require that the data and information to analyze is stable and does not constantly change, so that rules and models can be built more accurately. After the models are completed, then real-time data can be studied and classified to discover patterns, predict results, or determine courses of actions. For that reason, in some cases, it may be worthwhile to consider maintaining separate data warehouses, data marts, or data repositories to promote high-performance operations in all systems and to guarantee the data stability required.

A data warehouse contains information that usually concerns the whole organization, but in some projects all this information is not required. Data marts usually include subsets of data related to a specific department or specific subjects. Data warehouses are frequently built using the fact constellation schema because it can model “multiple interrelated subjects”; data marts, in contrast, use either the star or the snowflake schema, since both models are oriented toward single-structure objects; but the star schema is the one that is most frequently used because it is considered to be more efficient [18].

Another important consideration when constructing data marts or repositories is to use “a solid core of detail data” [6]; in cases when the data

must be gathered from legacy systems, this approach will greatly enhance the performance and stability of the system. Data repositories and data marts should also be built as close as possible to the original data, since summarization also implies, in most cases, loss of information. Finally, the use of prototypes can also be considered; prototypes are valuable tools of the systems development and design, but they can also be employed to understand and optimize data structures for storage.

Data Preparation

In many cases the transformation of data is also required. Data must be cleaned and integrated in order to correct possible inaccuracies, remove irregularities, eliminate duplicated data, detect and correct missing values, and check for any possible inconsistencies, before the analysis can began. Data mining tools can effectively create valid and insightful models only when the information provided is free of nuisance and noise factors.

The best way to guarantee the quality of information is to address it directly at the source. Good validation and consistency checks are ideally formed when data is initially gathered. Unfortunately, this is not always possible, since the only information available may already be polluted. Therefore, several techniques can be performed on data in order to clean and prepare it for data mining analysis.

In the case of missing values, the approaches that are usually considered are the following [18]:

- Ignore records
- Fill in values manually
- Create a special value or category

- Use the mean value of the distribution
- Use the mean value of the same class
- Use the most probable value.

Unfortunately, using these approaches may bias or alter the data, increasing the degree of contamination already existing. Filling in the values manually for thousands of records is not a viable solution, and the information may no longer be available. Creating special categories of unknown values is also risky; using them may generate data for which data mining tools mistakenly assume nonexistent patterns or relationships. Incorrect models may be built as a result.

Some authors have also observed that eliminating all records with missing values may reduce the amount of data available to a very small sample. Yet missing values can be an interesting characteristic of data that may itself be worthy enough for analysis [34]. Additionally, the use of the values of the mean may work well in some cases, but in others will conceal the real patterns or relationships that exist.

The most probable value is currently one of the most-used strategies; however, it is important to recognize that using this technique assumes existing relationships within data in order to discover other relationships. The only true solution for the problem of missing values resides, as mentioned above, in the initial process of data-gathering. Missing values can be avoided with strict controls in data entry processes, and with efficient and effectively designed methods that verify and validate the information directly at the source.

In order to eliminate noise in data, several other methods can be employed [18]:

- Binning
- Clustering
- Combine inspection
- Regression

Binning analysis sorts data values grouped in buckets or bins, and “smoothes” them by replacing the data in every bin with the correspondent mean, median, or closest boundary of each group, respectively. Clustering can also be used to eliminate noise by identifying extreme values out of the range of groups where similar values are identified. The combined inspection method requires the interaction between computer and human inspection. In this method, computers predict values that are compared against a certain threshold; when values exceed the threshold, they are presented to a human user that finally decides whether the value can be accepted or must be ignored.

The regression method, however, fits data to a certain function. Regression can be linear, multiple, or multidimensional depending on the number of variables used in the analysis [18]. Although noise detection is an important element for accurate data mining analysis, it must be implemented with caution. Noise reduction can also eliminate important patterns and relationships within data elements that must be identified, especially in applications such as process monitoring and quality control.

Other data transformations that can be useful for data mining analysis are aggregation, generalization, normalization and attribute constructions. With aggregation and generalization, data can be analyzed at different summarized and hierarchical levels [18].

Normalization transforms data into a specific range of values, while attribute construction allows data mining tools to analyze new variables generated using values already available. Additional available methods are the PCA, wavelet analysis, and event representation, which have been discussed in previous sections.

Model Development

The model development phase is an iterative process in which data mining tools analyze data and generate rules or identify patterns and relationships. Unfortunately, though, the rules, patterns or relationships identified by the different algorithms do not always have a significant meaning or use. Human experts are then required to identify, choose, and decide which are the most important rules and significant models.

For this aspect, training plays a very important role. Users must understand not only how to manage the software packages, but also what the data really represents. Additionally, for specific tasks such as process monitoring, quality control or product design, users must also have an authentic understanding about the process, tasks, materials and conditions involved. Only in this way is truly insightful information discovered.

In model development, data is explored in order to identify the most relevant fields. The data available is then divided randomly in subsets, at least one set for training and at least one set for validation. When the fields are selected and the data has been prepared, the best predictors are found, using the training data sets. With these predictors, several models are iteratively explored in order to find the most suitable for the project. The models are created using the rules and relationships discovered according to the data mining task and techniques selected for the project. Finally, the predictions of the models must be tested against the validation sets. If the

predictions are sufficiently good, the models can be implemented; otherwise, more models are explored in an iterative process.

Model Validation

The purpose of the model validation phase is to determine whether or not the models created by the data mining tools can correctly predict the behavior of the variables represented by the data. As mentioned above, a validation data set can be used to verify whether the predicted values of the model are close enough to the behavior expressed by the data in the validation data set. In order to perform this task, thresholds can be assigned according to the specific needs and conditions of each project. Cross-validation and bootstrapping [34] are two validation techniques that can be used to estimate the errors of the models. The error values can then be compared with thresholds to verify that the models are valid. However, even when a model successfully predicts the values in both the training and the validation sets, it is not guaranteed that the same model will always successfully predict the values of the variables represented by similar or new data.

For example, if a given system is affected by different external factors which were not present before, an old model would no longer be valid and new models would be needed. For this reason, threshold values should be periodically revised according to the current requirements of the organization. Furthermore, additional testing with new data is required if the intention of the project is to predict the behavior of a real system.

Finally, It is also important to remember that if noise was removed from the original data before the models were originally created, new data and information proceeding from the original sources may also contain noise and should be filtered before it can be analyzed by the models.

Implement Model

Once a model is validated, it can be implemented according to the goals and objectives initially established for the project. Implementation is an important phase and also requires analyzing and interpreting the results generated by the models. Not all data mining projects require the implementation of a specific model. However, the information gathered during the process, and the rules, relationships, or patterns discovered, can be used to solve specific problems, give recommendations, make decisions, or identify the necessity of further studies.

If the models are implemented with other applications, implementation itself can be considered as a part of a system analysis and design process, and it would require additional testing. Models can be used to classified specific records, assign probabilities, or generate special orders or reports. For that reason, additional Interface programs and software packages may be needed.

DFDs (data flow diagrams) and ERDs (entity relationship diagrams) can be used to analyze and design new applications. After programming has been completed, alpha and beta testing can then be executed. These tests are intended to guarantee that the system works intended in its design. For the test implementation, a top-down approach is recommended, in order to reduce cost and errors, as well as to facilitate system integration between the different modules. Additionally, the test and design operations must directly involve the final users. The inclusion of final users is essential, because the real acceptance and the success of the project depends on them.

The changeover phase of the final applications can be executed using the parallel or the phased method [37], according to the characteristics of the

project and the requirements and expectations of the stakeholders. If the system involves critical activities, they should be performed parallel to the older system, in order to increase the operation's reliability and security. Performance of new options, operations, and features should then be analyzed and compared with those of the old methods in order to detect possible flaws.

Using a phased changeover method will allow analysis of changes and customization of the system to specific requirements by case [37]. This option will also ensure that the new system works as intended for each element, before proceeding with the others. Moreover, evaluation of the project can also be measured using the decision matrix in Figure 10.

Weight

Score: (0-5) 0 -None 5 -Excellent $\sum p = 1$

	Score	Weight	Total
Did the Project meet Organization's Expectations?	A1	P1	A1xP1
Did the project meet stakeholders requirements?	A2	P2	A2xP2
Did the project achieved the expected benefits ?	A3	P3	A3xP3
Was the technique successfully selected ?	A4	P4	A4xP4
Was the performance of tools, software, and hardware satisfactory?	A5	P5	A5xP5
Was the model implementation (if required) successful?	A6	P6	A6xP6
Total			$\sum Ai \times Pi$

Figure 10. Decision Matrix for Project Evaluation

Establish On-going Support

Finally, data mining projects, in many cases, may also require the inclusion of a support phase. Maintenance operations must be periodically conducted for the equipment; moreover, the data and information residing in data marts, data repositories, and data warehouses must be protected by performing periodic back-ups. Back-ups can be full, differential, or incremental, according to the requirements of any given case.

Additionally, new types or sources of information, new versions of software packets, new operational systems, or new equipment may be available. In some other cases, the original models would need to be periodically updated, refined, or completely built again. The support phase must ensure that both the model and the corresponded applications are working appropriately and correspond with the specifications of the project.

Conclusion

By using a systems analysis approach, this chapter presented a proposed methodology for using data mining in solving problems related to industrial engineering. The proposed methodology encompasses five major phases: analyze the organization, structure the work; develop data model, implement model, and on-going support. Each of these phases has been described in detail and covers the major steps that any data mining project in industrial engineering must sustain from the origin of the project to its final implementation and support phases. The proposed methodology presents a solid framework capable of enabling industrial engineers to apply data mining in a consistent and repeatable way, which would enable them to evaluate data mining projects, duplicate results, or determine where the errors have occurred in their data mining projects.

CHAPTER 5

CONCLUSIONS AND RECOMMENDATIONS

Using the relationships, patterns, and rules found by data mining tools, industrial engineers may discover unexpected and useful information that can lead to a better understanding of systems and processes. This information can then be used to design new processes and new products, or to create modules and expert systems capable of controlling and optimizing systems. Industrial engineers can also use these modules to obtain better performance and resource utilization. The methodology developed in this research can help in these efforts.

Application in Industrial Engineering

An important consideration for successful application of data mining in industrial engineering is the perspective of the data mining methodology. The traditional focus has been centered on a statistical point of view. This approach is not systems-oriented, and it lacks some fundamental components needed in an information system project. Specifically, it does not incorporate the analysis, design, and implementation phases of an information systems project. Since the objective of data mining is to produce complete information for decision-making, its application should parallel the efforts of information system development. Additionally, the traditional approach generally does not consider the roles of the organization and the stakeholders during the project. It does not see data mining as an integral element of the organizational system. Instead most data mining methodologies focus on compatibility with specific software packages. What the industrial engineer needs is a software-independent methodology that

focuses on the role of information within the organization and its interactions with the other entities of the organizational system.

Another difficulty for industrial engineers who want to apply data mining is that traditional data mining approaches combine into a single step the selection of tools and techniques. Considering tools and techniques together, and early in the process, creates the risk of overlooking the organization's goals and requirements during the decision-making process. As a result, the selected tools and techniques may not be appropriate for the organization. This would then render the data mining effort useless. The models generated may not truly represent the behavior of the organizational entities for which it was initially intended.

Data mining techniques must be selected according to the organizations' goals and for their data requirements. In case the data available is not sufficient to perform a data mining project, more data and information should be collected. If data mining tasks and techniques are selected only for the data available, the patterns and resulting models may not apply to the organization's requirements.

The traditional approach of data mining has been focused on the possibility of analyzing available information that has already been stored in databases, data warehouses, or data marts. But in this study, the approach is to implement systems analyses that provide a basis for the design and development of data marts and data warehouses that will allow the creation of f data mining projects in industrial engineering areas. This approach calls for identifying informational needs, analyzing existing data sources, and either augmenting existing data stores or creating new ones to make needed data accessible.

In this approach, analyzing the organization's and the stakeholders' needs, requirements, goals, and strategies is a vital step. The stakeholders' involvement is also a key factor for success in any data mining project. The adequate fulfillment of all their needs, requirements, and expectations strongly determines the implementation and success of any data mining project. Additionally, for any project conducted in areas of industrial engineering, knowledge of the factors and processes involved have a positive effect and facilitate its successful implementation.

Documentation is an essential part of data mining projects because the models created with rules extracted using data mining tools can be very complex. As a result, a good documentation effort is required during the entire project; only in this way will analysts and programmers be able to keep track of all changes performed in any data source and to make all the modifications needed for the corresponding models. This documentation serves as a description of the data stores and the data mining efforts and their results, but it also provides a baseline upon which future enhancements can be made.

Application Concerns

Data mining is not a magical tool. The projects using data mining tools and techniques are not always easy to implement and may require considerable amounts of time and resources. Industrial engineering projects require especially careful planning, preparation, and study in order be successfully implemented and to obtain significant results. The more planning and preparation on the front-end, the better the chance of success on the back-end.

Not all data mining projects produce useful results. Data mining projects are based on the assumption that useful information, relations, and patterns can be extracted from data. However, if such elements do not in fact

exist, they cannot be found. Moreover, the patterns, rules, and relationships present in data may be caused by chance alone and may not represent the exact behavior of a given system.

Data mining projects in industrial engineering, however, are capable of generating a wide variety of benefits for organizations. They can, for example, improve the stability of processes, reduce delays, increase efficiency in material flows, improve quality of products, reduce downtimes and repairs, decrease maintenance costs, improve scheduling of tasks and operations, reduce energy consumption, reduce waste in operations, and improve products design and operations safety.

With each new day, more and more data mining applications are discovered and implemented. Data mining is already helping companies and organizations to manage and allocate their resources in more effective and efficient ways, thus reducing cost and improving the quality of products and services they offer.

But data mining is not simply a tool for producing useful data. It is also a tool that can help industrial engineers to better understand processes, system behaviors, and interactions with their environment. As with any tool, the results may vary depending on how data mining is applied.

Application Issues

As previously discussed, the applications of data mining in manufacturing and other areas of industrial engineering require good knowledge and understanding of the processes and systems involved. As a result, data mining analysts must not only have good awareness of and familiarity with all the processes and variables of the study, but also, they

must be certain that all of the data sources available are valid, correspond to the specifications and requirements, and are suitable for analysis.

Consequently, the training and preparation required for these types of studies may be extensive--not only for the selection of the corresponding tasks and techniques, but also for the application of the necessary tools. Many of the data mining tools and solutions available are very complex and not intuitive for the users; they include a wide spectrum of choices available for algorithms, tasks, and models. Therefore the training required for data mining projects must also include training in data mining software packages, which may take 80 hours or longer.

Additionally, not all companies and organizations are prepared to conduct data mining studies. Companies may not have considered the possibility of using data mining in their records and historical information. So, when the opportunity of introducing data mining techniques appears, it may be that not all the required information is available. Or it may be that, the data is distributed in many different types of databases, locations, and formats.

These conditions significantly degrade data mining performance, require the application of special data cleaning and data preparation techniques, and represent a considerable new investment of time and other resources. The best solution to this problem is to design databases, data marts, data repositories, and data warehouses, so that data mining can be successfully applied in order to fulfill all stakeholders' requirements and expectations.

Another consideration is that not all available data is useful, and historical data may not generate good rules or prediction models. Some processes in manufacturing are so dynamic and flexible that the information available may not correspond with existing conditions or current products and

processes. Additionally, in some cases information may no longer be relevant, and analysts may be unaware of the data's currency. Wrong assumptions may be made, thereby, generating erroneous or inaccurate models that lead to inappropriate recommendations.

For that reason, caution is strongly recommended in selecting data sources for the studies. Data mining analysts in industrial engineering projects are required to have a good understanding of the process and the selected tools and techniques. They also need to know the origin of the data before the data mining process begins.

Only when these requirements are met will the rules, models, interactions, and relationships found in the studies have a greater probability of representing the real systems. Analysts then can confidently appreciate what the information implies.

Finally, data mining projects can also be very expensive. Costs do not include only the value of the software packages (which ranges at present from $100 to $250,000 and can be segregated in many different modules), but also other related expenses, such as additional annual licensing fees, software, hardware, installation, training, technical support, maintenance, consulting, and outsourcing. These costs include up-front, one-time costs as well as on-going costs. Thus organizations must consider the economic feasibility of the data mining project throughout its useful life.

Future Work

There are many possible areas for the application of data mining in industrial engineering. Although many of them have been listed in this research, many others can also be found. The methodology proposed in this

research provides a structural basis from which additional studies can be developed.

Currently, there are public databases that contain information relevant to industrial engineering applications. Some of these databases can be accessed through the internet. There are several that have useful information for ergonomic applications. They include the following: (1) OSHA (Occupational Safety and Health Administration) Accident Investigation, which enables users to search texts of accident investigation summaries (OSHA-170 forms) and is located at: http://155.103.6.10/cgi-bin/inv/inv1; (2) CrashDatabase.com, which contains information on airplane crashes and is located at: http://www.crashdatabase.com/; and (3) ARIP (Accidental Release Information Program), which contains information about chemical accidents and is located at: http://d1.rtknet.org/ari/.

Although, many of these databases are not designed for data mining analysis and much of the information contained in them is in text format, they still can be used as a source of data for further analysis. Text data mining, which is presently under development by many different software companies, would be a suitable application for these cases. Text data mining is currently being used for email routing, document indexing, and document filtering; but in the future, it will be able to extract more detailed and comprehensive information from a wide variety of sources.

This research presented a conceptual model to be applied in industrial engineering applications of data mining. This methodology, however, should be applicable to a variety of data mining projects. The next step for this research is to test and improve this conceptual model. Data mining is a constantly evolving tool, so this research will endeavor to involve it dynamically in the industrial engineering toolbox.

LIST OF REFERENCES

1. Anonymous, “Mining for a competitive Advantage in your Data warehouse”, Techguide.com retrieved from the World Wide Web on January13, 2002. http://techguide.znet.com/html/datamine/.

2. Anonymous “Uncover gems of information”, SAS Institute. Retrieved from the World Wide Web on January 28, 2002. http://www.sas.com/products/miner/index.html

3. Anonymous, “DaimlerChrysler Drives Information Discovery’s Pattern Warehouse”, Information Discovery, retrieved from the World Wide Web on March 10, 2002 http://www.dataminig.com/casestudies.htm

4. Anonymous, “New solution helps better ensure the safety of workers who face the risk of hazardous gas exposure”. IBM Corporation. Retrieved from the World Wide Web on March 10, 2002. http://www2.software.ibm.com/casestudies/swcs.nsf/customername/8FD22AACF8B3001787256B59002C5896

5. Anonymous, “ UKH mines accident data with IBM in quest for safer public work places ”. IBM Corporation. Retrieved from the World Wide Web on March 10, 2002. http://www2.software.ibm.com/casestudies/swcs.nsf/customername/D6A0304882C71DF60025689700062B7E.

6. Armstrong, Rob., Coffin, Tom ., and Rolf Hanusa. “Secrets of the Best Data Warehouses in the world”. Coffin Data Warehousing .2000. USA.

7. Benning, Stacey. Denning, Michelle. Jaquint, Cooch., and Rusell,Paul.“Data Mining” . University of Iowa, retrieved from the World Wide

Web on March 10, 2002. http://www.biz.uiowa.edu/class/6k180_park/Student-Reports/sbenning/

8. Bertino, Elisa, Catania, Barbara and Caglio, Eleonora. “Applying Data mining Techniques to Wafer Manufacturing” ”, retrieved from the World Wide Web on April 24, 2001.http://citeseer.nj.nec.com/bertino99applying.html

9. Bose, Idranil., and Mahapatra, Radha., “Business data mining- a machine learning perspective”. Information & Management. V39 .2001.

10. Braha, Dan. “Data Mining for Design and Manufacturing: Methods and Applications”. Kluwer Academic Publishers. 2001.

11. Cawley, Jeffery L. “Mine your P’s & Q’s”. Industrial Computing, June 1999.

12. Chapman, Pete. Clinton, Julian. Kerber, Randy. Khabaza, Thomas. Reinartz, Thomas. Shearer, Colin. And Wirth, Rüdiger. “CRISP-DM 1.0, Step by Step data mining guide” USA .SPSS Inc. 2000, retrieved from the World Wide Web on January 10, 2002. http://www.spss.com/CRISPDM/.

13. Chen, Nianyi., Zhu, Dongping Daniel, and Wang, Wenhua, “Intelligent material processing by hyperspace data mining”. Engineering Applications of Artificial Intelligence, V13. 2000.

14. Collier, Ken. Sautter, Donald. Medidi, Maralidhar. et at. “ Methodology for evaluating and selecting Data mining software”. Center for Data Insight (CDI). Northern Arizona University. Retrieved from the World Wide Web on January 15, 2002. http://insight.cse.nau.edu.

15. Famili,A., Shen, Wei-Min., Weber.,and Richard., Simoudis, Evangelos. “Data Preprocessing and Intelligent Data Analysis”. Intelligent Data Analysis. V1. 1997.

16. Ford, Denise Darcell. “Design of a Conceptual Information System for Industrial Engineering Applications in a Manufacturing Organization”. A Thesis presented for the Master of Science Degree, The University of Tennessee, Knoxville. 1983.

17. Grossman, Robert L., Kamath, Chandrika., Kegelmeyer, Philip., Kumar, Vipin., and Namburu, Raju R. “Data Mining For Scientific and Engineering Applications”. Kluwer Academic Publishers. 2001. Netherlands.

18. Han, Jiawei. And Kamber Micheline. “Data Mining, Concepts and Techniques”. Morgan Kaufmann Publishers. 2001.

19. Inmon, H. Zachman, John, and Geiger, Jonathan. “Data Stores Data Warehousing and the Zachman Framework, Managing Enterprise Knowledge”. McGraw-Hill.1997.USA.

20. Jackson, Denise. “A Methodology for the Quantification of Knowledge Work”. Dissertation presented for the Doctor of Philosophy Degree, The University of Tennessee. 1989.

21. Koonce, D.A. Tsai , S.-C. and Fang, Cheng-Hung. “A Data Mining Tool For Learning From Manufacturing Systems” Computer and Industrial Engineering V33. 1997.

22. Koonce, D.A., and Tsai , S.-C. “Using data mining to find patterns in genetic algorithms solutions to a job shop schedule”. Computer and Industrial Engineering .V38. 2000.

23. Landis, Raymond B. ”Studying Engineering, A Road Map to a Rewarding Career”. Discovery Press, Burbank California, 1995. USA .

24. McDonald, Chris J. “New tools for yield improvement in the integrated circuit manufacturing: can they be applied to reliability?”. Microelectronics Reliability. V39. 1999.

25. Milne, Robert. Drummond, Mike and Renoux, Patrick. “Predicting making defects online using data mining”. Knowledge-Base Systems. V11,1998.

26. Muth, John F., and Thompson, Gerald L. “Industrial Scheduling”. Prentice-Hall, Inc. 1963. USA.

27. Parsaye K., “Data Mines for Data Warehouses”, Information Discovery, retrieved from the World Wide Web on March 10, 2002 http://www.dataminig.com/dm4dw.htm

28. Paschal, G. Gray. “Use of Principal Component Analysis for Data Reduction for Training Neural Networks”, Thesis presented for the Master of Science degree, University of Tennessee, Knoxville. December 1996.

29. Shelly, Gary B. Cashman, Thomas J. Vermaat, Mysty E. and Walker, Tim J. “Discovering Computers 2001”. Shelly Cashman Series. 2000. USA.

30. Sohn, S Y., and Shin, H. “Pattern recognition for road traffic accident severity in Korea”. Ergonomics. V44. Issue 1. January 15. 2001.

31. Subramanian, Nirmala. “Data mining approach to improvement and standardization of operations in test facility”. Thesis presented for the Master of Science degree, Ohio State University. 1999.

32. Sule, Dileep R. “Industrial Scheduling”, PWS Publishing Company. 1997. USA.

33. Sullivan, William G., Bontadelli, James A., and Wicks Elin M. “Engineering Economy” Prentice- Hall Inc. 2000. USA.

34. Two crows corporation, “Introduction to Data Mining And Knowledge Discovery” Third Edition., 1999.

35. Wang, Xue Z. “Data Mining and Knowledge Discovery for Process Monitoring and Control”. Springer – Verlang London Limited. 1999. Great Britain.

36. Westphal, Christopher. Blaxton Tersa. “Data Mining Solutions, Methods for Solving Real-World Problems”., John Wiley & Sons , Inc. 1998 .USA.

37. Whitten, Jeffrey and, Bentley, Lonnie. “Systems Analysis and Design methods”. Irwin McGraw-Hill, Forth Edition.1998.

38. Witten, IAN, and Frank, Eibe. “Data Mining, Practical Machine Learning Tools and Techniques with Java Implementations”. Morgan Kaufmann Publishers.2000. USA.

39. Zandin, Kjell B. “Maynard’s Industrial Engineering Handbook”. MacGraw-Hill. 2001. USA.

APPENDICES

Key Terms

Attrition: Retention of Customers

Backpropagation: A training method use to calculate the weights in neural networks from data.

CRM (Customer Relationship Management): Process of studying and interacting with customers in order to maximize profits. It includes ensuring customer satisfaction as well as cutting service to unprofitable customers.

Data Mart: Small data warehouse focused on a single area such as a research project or department.

Data Repositories: Collection of resources that can be used to retrieve information, repositories usually consist of several databases tied together by a common search engine.

Data Visualization: Techniques for turning data into information by using visual representation and the capacity of the human brain to recognize patterns and trends.

Data Warehouse: A static copy of a database generally optimized for analysis or renormalized.

Decision Trees: Once of the most popular data mining techniques, that search for ways to divide data into subgroups that are as much similar as possible with regard to a target variable. Two of the most popular tree models are the CHAID and the CART.

Decision Support Software: Software that uses analysis to improve a decision making process.

Genetic Algorithms: A computer-based method of generation and testing combination of possible input parameters to find an optimal output. This process is based on natural evolution concepts such as combination, mutation, and natural selection.

Induction: Technique that infers generalizations from the information in data.

Neural Network: Model that mimic the brain through systems of equations, they learn by being trained with a data set.

Rule: A conditional statement that tells a model or a system how to react to a particular situation.

Schema: A technical blueprint of the database

Tuples: Records in a relational data base system.

Providers and Vendors

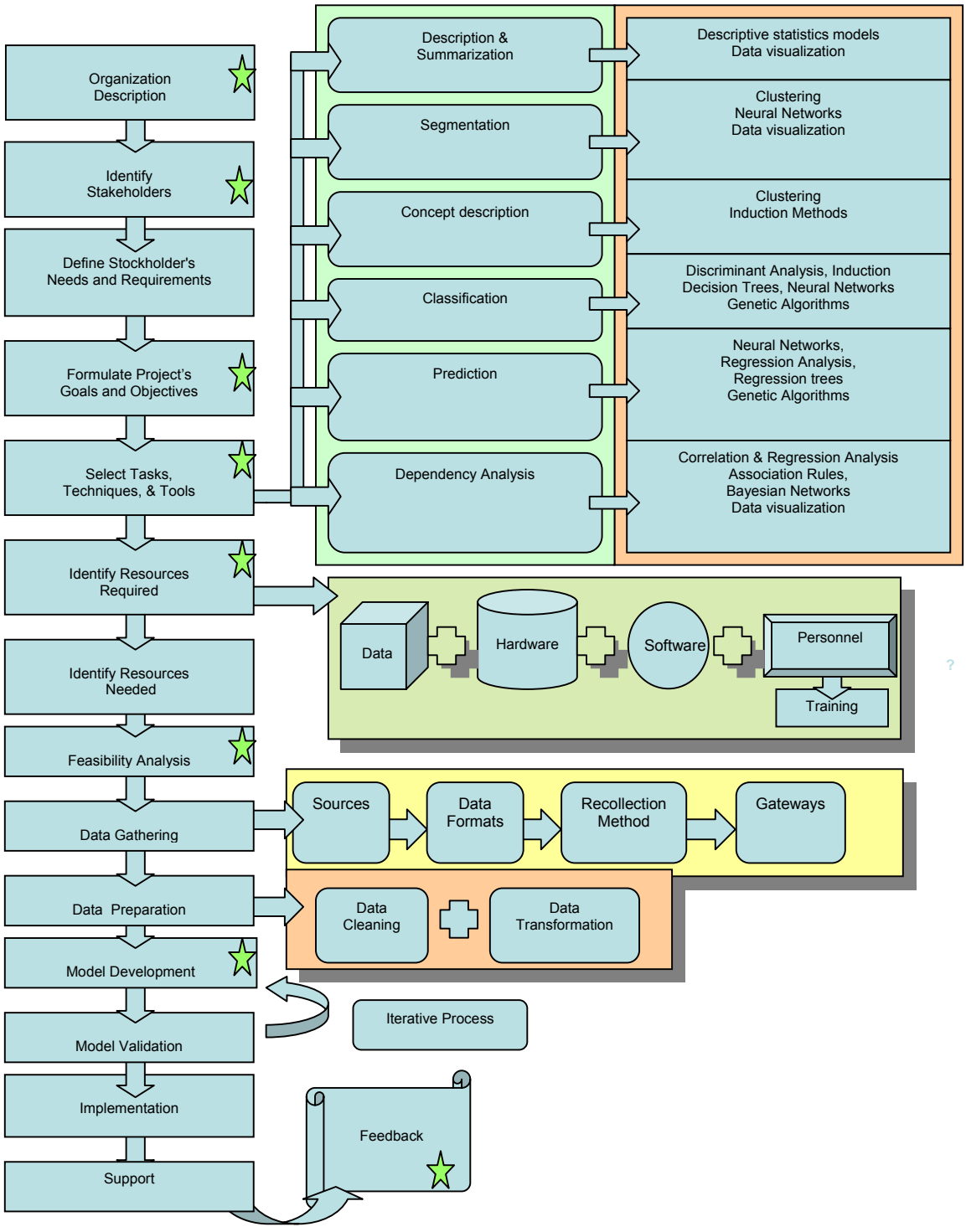
There are many companies and providers of data mining solutions; each of them offers different tools, algorithms and tools in their software packets. Some of the more important providers for data mining solutions are presented as follows:

(source: **http://www.kdnuggets.com/companies/products.html**).

- Abtech Corporation.
  www.abtech.com
- Advanced Software Applications.
  www.asacorp.com
- ANGOSS Software.
  www.angoss.com/
- Apower Solutions.
  www.apower.com
- ASA Corp.
  www.asacorp.com/
- Attar Software.
  www.attar.com/
- BrainMine.
  www.brainmine.nl/
- ClearForest Corp.
  www.clearforest.com/
- Cygron Pte Ltd.
  www.cygron.com/
- Data Description, Inc.
  www.datadesk.com/
- Data Mining Technologies.
  www.data-mine.com
- Exclusive Ore Inc.
  www.exclusiveore.com/
- IBM Global Business Intelligence Solutions.
  http://www-4.ibm.com/software/data/iminer/fordata/
- Information Discovery Inc.
  www.datamining.com
- Insightful Corporation.
  www.insightful.com/default_class5.asp
- Isoft.
  www.alice-soft.com/

- Logic Programming Associates Ltd.
  www.lpa.co.uk/ind_top.htm
- Manning and Napier Information Services (mnis).
  www.mnis.net
- MarketMiner.
  www.marketminer.com/
- Neural Technologies.
  www.neuralt.com/
- NeuroDimension, Inc.
  www.nd.com/
- ProGAMMA.
  www.gamma.rug.nl/
- Prudential System Software.
  www.prudsys.com/
- Quadstone.
  www.quadstone.com
- Rulequest
  www.rulequest.com/
- Salford Systems.
  www.salford-systems.com
- SAS Institute. Inc.
  www.sas.com
- Sentient Machine Research.
  www.smr.nl/
- SPSS.
  www.spss.com/
- Statsoft Inc.
  www.spss.com/
- Temis-Group
  www.temis-group.com/
- Urban Science.
  www.urbanscience.com/
- WhiteCross.
  www.whitecross.com/
- WizSoft, Inc.
  www.wizsoft.com/
- Webminer.
  www.webminer.com

Methodology Proposed



Suggested Applications of Data Mining to Industrial Engineering.

Area	Possible Application
Maintenance and Reliability	• Fault Diagnosis • Preventive and Predictive Maintenance Analysis • Total Productive Maintenance Models
Inventory Control	• Inventory Reduction Studies • Warehouse Management
Product and Process Development and Design	• Concurrent Engineering Analysis • Waste–free Process Design • Material Selection. • Material Handling Studies • Workstation Design • Tool Selection • Work Method Design • Product Safety Evaluation • Rapid Product Development • Process Integration • Plant Layout Design • Virtual Manufacturing • Quality Function Deployment. • High Adaptability System Development • Usability Analysis. • Simultaneous Engineering Studies. • Error Proof Design. • Flexible Manufacturing. • Quick response Manufacturing • Product Liability Studies
Work Measurement	• Motion Studies • Work Load Analysis • Standard Development • Allowance and Fatigue Studies • Learning Curves Analysis
Process Optimization and Improvement	• Productivity Improvement Analysis. • Analysis of Process Variations within Manufacturing and Assembly Operations. • Waste Analysis • Shop Floor Control • Downtimes Elimination • Setup Reduction • Assembly Lines Improvement • Product Pace Studies • Work in Process Reduction • Lot Size Studies • Queuing Analysis • Simulation Results Analysis • Value Management Studies

Suggested Applications of Data Mining to Industrial Engineering (continued)

Area	Possible Application
Labor	• Motivational Job Analysis • Training Improvement Studies • Job Evaluation Analysis • Incentive and Job Retention Models • Occupational workforce Composition Studies
Engineering Economy	• Payback Analysis • Cost Reduction • Activity Base Costing • Risk analysis
Quality Control	• Product quality Improvement analysis
Occupational Safety	• Accident Causation Models • Hazard Exposure Analysis • Occupational Risk Analysis • Accident Rates Studies.
Facility Layout and Design	• Operation Relocation Studies • Space Utilization Evaluation • Work Cells Analysis • Equipment Disposition Evaluation • Material Flow Studies
Logistics	• Facility Location Analysis • Product Location Studies • Relocation of Manufacturing Sites. • Supply Chain Oscillation Analysis
Ergonomics	• Design of Human-machine interfaces • Ergonomic Risk factor Analysis • Musculoskeletal Stress and Injury studies • Biomechanical Profile Analysis.
Scheduling	• Production and Labor scheduling • Finite Capacity Scheduling and Capacity Planning

VITA

Jose Solarte was born in Bogotá, Colombia on April 19, 1974. He graduated from the Calasanz High School in December 1991. In 1997, he received a BS degree in Industrial Engineering from the Pontifical Javeriana University in Bogotá, Colombia. Jose is currently pursuing his doctorate in Industrial Engineering at the University of Tennessee, Knoxville.