A COMPREHENSIVE REVIEW OF DATA MINING APPLICATIONS IN QUALITY IMPROVEMENT AND A CASE STUDY

A THESIS SUBMITTED TO
THE GRADUATE SCHOOL OF NATURAL AND APPLIED SCIENCES OF
MIDDLE EAST TECHNICAL UNIVERSITY

BY

FATMA GÜNTÜRKÜN

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR
THE DEGREE OF MASTER OF SCIENCE
IN
STATISTICS

AUGUST 2007

Approval of the thesis:

A COMPREHENSIVE REVIEW OF DATA MINING APPLICATIONS IN QUALITY IMPROVEMENT AND A CASE STUDY

submitted by **FATMA GÜNTÜRKÜN** in partial fulfillment of the requirements for the degree of **Master of Science in Statistics Department, Middle East Technical University** by,

Prof. Dr. Canan Özgen ______________________
Dean, Graduate School of **Natural and Applied Sciences**

Prof. Dr. Ali Uzun ______________________
Head of Department, **Statistics**

Assoc. Prof. Dr. İnci Batmaz
Supervisor, **Statistics Dept., METU** ______________________

Prof. Dr. Gülser Köksal
Co-Supervisor, **Industrial Engineering Dept., METU** ______________________

Examining Committee Members:

Prof. Dr. Gülser Köksal ______________________
Industrial Engineering Dept., METU

Assoc. Prof. Dr. İnci Batmaz ______________________
Statistics Dept., METU

Assoc. Prof. Dr. Murat Caner Testik ______________________
Industrial Engineering Dept., HÜ

Dr. Özlem İlk ______________________
Statistics Dept., METU

Dr. Ceylan Yozgatlıgil ______________________
Statistics Dept., METU

Date: (23.08.2007)

I hereby declare that all information in this document has been obtained and presented in accordance with academic rules and ethical conduct. I also declare that, as required by these rules and conduct, I have fully cited and referenced all material and results that are not original to this work.

Name, Last name: Fatma Güntürkün

Signature :

ABSTRACT

A COMPREHENSIVE REVIEW OF DATA MINING APPLICATIONS IN QUALITY IMPROVEMENT AND A CASE STUDY

GÜNTÜRKÜN, Fatma

M.S., Department of Statistics

Supervisor: Assoc. Prof. Dr. İnci Batmaz

Co-Supervisor: Prof. Dr. Gülser Köksal

August 2007, 111 pages

In today's world, knowledge is the most powerful factor for the success of the organizations. One of the most important resources to reach this knowledge is the huge data stored in their databases. In the analysis of this data, DM techniques are essentially used. In this thesis, firstly, a comprehensive literature review on DM techniques for the quality improvement in manufacturing is presented. Then one of these techniques is applied on a case study. In the case study, the customer quality perception data for driver seat quality is analyzed. Decision tree approach is implemented to identify the most influential variables on the satisfaction of customers regarding the comfort of the driver seat. Results obtained are compared to those of logistic regression analysis implemented in another study.

Keywords: Data mining, quality improvement, manufacturing, logistic regression, decision trees

ÖZ

VERİ MADENCİLİĞİNİN KALİTE İYİLEŞTİRMEDEKİ UYGULAMALARININ GENİŞ BİR ÖZETİ VE ÖRNEK BİR ÇALIŞMA

GÜNTÜRKÜN, Fatma

Yüksek Lisans, İstatistik Bölümü

Tez Yöneticisi: Doç. Dr. İnci Batmaz

Ortak Tez Yöneticisi: Prof. Dr. Gülser Köksal

Ağustos 2007, 111 sayfa

Günümüzde işletmelerin başarılı olabilmesi için sahip oldukları en önemli unsur bilgidir. Veri tabanlarında saklanan veri kümeleri bu bilgiye ulaşırken kullanılan önemli kaynaklardan biridir. Bu veri kümelerinin çok büyük olması nedeni ile analizlerinde kaçınılmaz olarak veri madenciliği tekniklerinin kullanılması gerekmektedir. Bu tezde öncelikle üretimde kalite iyileştirme amaçlı veri madenciliği tekniklerini kullanan çalışmaları içeren geniş bir literatür özeti sunulmuştur. Daha sonra örnek bir çalışmada sürücü koltuğu kalitesi için müşteri memnuniyeti verisi analiz edildi. Müşterinin sürücü koltuğundan memnuniyetini etkileyen en önemli değişkenlerin belirlenmesi için karar ağaçları yaklaşımı uygulandı. Bu uygulamadan elde edilen sonuçlar diğer bir çalışmada aynı veri kümesine uygulanmış logistik regresyon analizinden elde edilen sonuçlarla karşılaştırıldı.

Anahtar Kelimeler: Veri madenciliği, kalite iyileştirme, üretim, logistik regresyon, karar ağaçları

ACKNOWLEDGMENTS

First of all, I would like to give special thanks to my thesis supervisor, Assoc. Prof. Dr. İnci Batmaz and co-supervisor Prof. Dr. Gülser Köksal for their professional support, guidance and encouragements which are invaluable for me during my research.

I am also grateful to my friends Berna Bakır, İlker İpekçi and Vuslat Çabuk for the amusing time we spent when we are studying together and also for their help during writing this thesis.

I am also indebted to Assoc. Prof. Dr. Murat Caner Testik, Prof. Dr. Nur Evin Özdemirel, and Prof. Dr. Sinan Kayalıgil for their useful advice and criticism.

I give my special thanks to Aslıhan Sertkaya for her perfect friendship.

Love and thanks to my lovely husband Mustafa Hakan Güntürkün.

Lastly thanks to my mother and father for their endless support and love they provided throughout my life.

LIST OF FIGURES

LIST OF TABLES

TABLE OF CONTENTS

CHAPTER 1

INTRODUCTION

As long as knowledge is another name of power, organizations give much importance to knowledge. When reaching this knowledge, they make use of the data in their databases. Data is important since it provides them to learn from the past and to predict future trends and behaviors. Today most of the organizations use the data collected in their databases when taking strategic decisions.

The process of using the data to reach this knowledge consists of two steps as collecting the data and analyzing the data. In the beginning, organizations face with difficulties when collecting the data. So, they have not enough data in order to make suitable analysis. In the long run, with the rapid computerization, they are able to store huge amount of data easily. But at this time they face with another problem when analyzing and interpreting of such large data sets. Traditional methods like statistical techniques or data management tools are not sufficient anymore. Then, in order to manage with this problem the technique called data mining (DM) has been discovered.

DM is a new useful and powerful technology that supports companies to derive strategic information in their databases. It has been defined as: 'The process of exploration and analysis, by automatic or semiautomatic means of large quantities of data in order to discover meaningful patterns and rules" [1]. The expression meant by meaningful patterns and rules is: easily understood by humans, valid on new data, potentially useful and novel. Validating a hypothesis that the user wants to prove can also be accepted as a meaningful patterns and rules. In sum, it is essential to derive patterns and rules that help us to reach strategic and unimagined information in DM.

1.1 SCOPE OF THE THESIS

This thesis consists of two parts. In the first part, a comprehensive literature review including the DM applications on manufacturing data is presented. First of all, the articles published in the literature between the years 1995-2007 are searched. Then we examine 80 articles which are in the scope of this study. To this end, a table was prepared including the information about the main aim of the articles and the DM process that they followed to reach that specified aim. In this table, the DM process was mentioned in detail. It includes information about the quality tasks, DM tasks and DM tools that are used in the articles. It also covers the information about which manufacturing data was used, how it was collected, what the size and structure of data is and how it was preprocessed. Later on, by using this table some summary tables and graphs were derived. By those tables and graphs we see the connections between the DM tools and DM tasks and also between DM tasks and quality tasks. Additionally, there were also some graphs relating the number of articles published and the number of each quality tasks studied in each year. Besides these, we searched the thesis and dissertations published in the literature between the years 2002-2007. And, we found three dissertations that are related to our subject. Then we summarized information about these three dissertations at the end of the literature review part of the study.

In the second part of this thesis, a case study on the analysis of customer quality perception data for driver seat quality improvement is presented. The aim of this study is to identify the most influential variables that affect the customer satisfaction, and by this way providing a decision support to the company for a new design of an optimal prototype of a driver seat. For this aim, the data which had been collected for another ongoing study [99] was used. In that ongoing study [99], a face to face questionnaire had been applied to one of the automotive company's customers in Turkey. Thus, the information about the customer satisfaction and the related factors had been gathered from 80 customers. The decision tree approach was used to determine the important factors regarding the deriver seat on customer satisfaction from the driver seat. At the end of this case study, the results obtained from the decision trees was compared and discussed with the results obtained from the logistic regression which was applied in the scope of another ongoing study [99].

1.1 OUTLINE

This thesis consists of six chapters. The first chapter is an introductory chapter in which the definition and the importance of the DM are mentioned. In the second chapter, some background information about the DM is presented. In addition, some common DM tasks and techniques are explained, and the main steps of the DM process are mentioned. At the end of the second chapter, some popular application areas of the DM are presented. In the third chapter a comprehensive literature review on DM techniques used in manufacturing data is introduced. Besides some summary tables and graphs of the articles published on quality improvement in manufacturing between the years 1995 and 2007 are presented. In the fourth chapter, materials and methods used in this thesis are introduced. The software used in the case study, three decision tree algorithms and the logistic regression technique are explained shortly. In the fifth chapter a case study on the customer quality perception data for driver seat quality improvement is presented. The DM process which starts with data gathering, preprocessing and continue with modeling and checking the performance of the developed models are mentioned in detail. In the final chapter, conclusion of this study and the possible works that can be done in the future are presented.

CHAPTER 2

DATA MINING PROCESS, TECHNIQUES AND APPLICATION

2.1 DM PROCESS

According to Fayyad DM refers to ‘a set of integrated analytical techniques divided into several phases with the aim of extrapolating previously unknown knowledge from massive sets of observed data that do not appear to have any obvious regularity or important relationships’[7]. Another definition is, ‘DM is the process of selection, extrapolation and modeling of large quantities of data to discover regularities or relations that are at first unknown with the aim of obtaining clear and useful results of the database’[7]. By using these definitions and harmonizing them, in this study DM is considered as a whole process consisting of different steps where in each step different DM techniques can be used. Then, several steps of the DM process and the possible DM techniques that might be used in each step are tried to classify as follows [3, 4, and 5].

1. Data gathering
    i. Feature determination
    ii. Database formation
2. Data preprocessing
    i. Data cleaning
        1. Missing data handling
           [For example: Fill in missing value manually, Use a global constant to fill in missing value, Use the attribute mean to fill in missing value, Use the most probable value to fill in missing value, Ignore the tupple or the attribute etc.]

2. Outlier and inlier handling
3. Inconsistent data handling

ii. Data integration

iii. Data transformation

1. Smoothing,
2. Aggregation,
3. Generalization,
4. Normalization of data

   [For example: min-max normalization, z-score normalization, normalization by decimal scaling etc.]
5. Attribute construction
6. Discretization and concept hierarchy generation

   [For example: S-based techniques (DA etc.), NN-based (SOM) etc.]

iv. Data reduction

1. Data cube aggregation
2. Dimension reduction(Feature selection)
   a. Feature wrapper
   b. Feature filter
3. Data compression

   [For example: Wavelet transform, S-based techniques (PCA, FA etc.) etc.]
4. Numerosity reduction

   [For example: S-based techniques(R, Histograms etc.), Clustering etc.]

v. Over sampling

3. Modeling

i. Predictive model

1. Classification

   [For example: S-based techniques(R, BC, LR etc.), DT-based(OC1, ID3, CHAID, ID5R, C4.5 AND C5, CART, QUEST, Scalable DT techniques, Statistical batch-based DT learning etc.), NN-based (w.r.t. learning algorithm:

backpropagation, Levenberg-Marquart; w.r.t. architecture: RBF, Perceptrons, ARTMap BNN), R-based (Generating Rules from a DT, Generating Rules from a ANN{ Rectangular basis function network }, Generating Rules without a DT and ANN {PRISM, RST, FST etc.}), Combining techniques (Integration of FST and RST, FAN, EN, CC etc.), SVM etc.]

2. Prediction

   [For example: S-based techniques (Parametric {MLR as RSM, GLM as ANOVA, MANOVA, TM, NRM as Generalized Additive Models, RR, BR, TSA as exponential smoothing etc.}, Nonparametric {ANOVA as Kruskal-Wallis, R, TSA as Moving average etc.}), DT-based (ID3, C4.5 and C5, CART, CHAID, Scalable DT techniques etc.), NN-based(w.r.t. learning algorithms: Feedforward Propagation, backpropagation; w.r.t. architecture: RBF{RBF network as Gaussian RBF NN}, Perceptrons, BNN), R-based (Generating Rules from a DT, Generating Rules from a ANN, Generating Rules without a DT and ANN{PRISM}), CBR, FEMS, SVM, Combining techniques (Modular ANN { FNN, Fuzzy ARTMAP NN, ANFIS etc.}) etc.]

ii. Descriptive model

1. Clustering

   [For example: Hierarchical methods (Agglomerative, Divisive), Partitional methods (Minimum spanning tree, Squared error, K-means, Nearest neighbor, PAM, Bond energy, GA, NN based {w.r.t. learning rule: Competitive as SOM, LVQ etc.}, Non competitive (Hebian)), Rule-based (Generating Rules from a ANN) etc.]

2. Summarization(Visualization and Statistics)

a. Visualization

[For example: S-based (Histograms, scatter plots, box plots, pie charts, 3_D plots etc.) etc.]

b. Statistics

[For example: Descriptive statistics (mean, median, frequency count etc.), Density estimation etc.]

c. Tables

3. Association

   a. Basic methods

      [For example: Apriori, Sampling, Partitioning etc.]

   b. Advanced association rules method

      [For example: Generalized association rules, Multiple-level association rules, Quantitative association rules, Using multiple minimum supports, Correlation rules etc.]

4. Optimization

   [For example: S-based (TM, RSM), NN-based, GA, SA, SQP, Levenberg-Marquardt method etc.]

S-based: Statistical based, DT-based: Decision tree based, NN-based: Neural network based, R-based: Rule based, R:Regression, PCA:Principle component analysis, RBF:Radial basis function, SVM:Support vector machines, CBR: Case based reasoning, GA:Genetic algorithms, SE:Subjective and empirical approach, BNN:Bayesian networks, FAN:Fuzzy adaptive network, EN:Entropy network, CC:Composite classifiers, TSA:Time series analysis, FEM:Finite element modeling, GSA:Grey superior analysis, SA: Simulated annealing, SQP: Sequantial quadratic programming method, DA: Discriminant analysis, CA:Correlation analysis, BC:Bayesian classification, GLM:General Linear Models, NRM:Nonlinear regression models, RR:Robust regression BR:Bayesian Regression, FA: Factor analysis, LR:Logistic regression, MLP:Multi linear perceptron, LVQ: Learning vector quantization

2.2 MAIN DATA MINING TASKS

There are different ways of distinguishing interesting patterns or trends in a huge amount of data set which are called DM operations or tasks. In literature, there exist different DM task categorizations. For instance, one is the categorization involving Prediction, Classification, Clustering, Affinity Grouping or Association Rules and

Visualization and Statistics [1]; and another involving Classification, Regression, Clustering, Summarization, Dependency Modeling, Change and Deviation Detection [12]. There are also some other categorizations involving various classes as Outlier Analysis and Text Mining [13]. We are mainly interested in the following DM tasks. Optimization in the categorization does not exist in the literature. So that, we newly defined this category. We defined it because although the papers commonly used DM tools for optimization purposes, there were not any DM tasks suitable to this case. The few rest that out of the scope of this study as Text mining, Web mining, Spatial mining and Temporal mining or in the scope of this study but not studied in the papers placed in the table as Affinity Grouping or Association Rules, Visualization and Statistics are listed in the others part. The analyst can apply one or several of them during the analysis on the dataset.

2.2.1 Data gathering and preprocessing

The first step of DM applications is data gathering. In this part, it is aimed to obtain the right data. For this purpose, all the available data sources are examined then the right data for the recent analysis is selected. It includes two steps: Feature determination and database formation. In feature determination it is determined the name of the variables whose data is collected. Whereas in database formation, collected data is returned to database format.

The second step is data preprocessing. The goal of this step is to investigate the quality of the selected data then transform in order to make it suitable for further analysis. This part is important since real life data is incomplete, noisy and inconsistent. Data preprocessing consists of data cleaning, data integration, data transformation and data reduction.

Data cleaning deals with filling the missing values, detecting the outliers then smoothing the noisy data and correcting the inconsistencies in the data. Methods for missing values are listed in DM Process part. All of them have some advantages and disadvantages respectively. For example, if tupple does not contain many missing values ignoring the tupple is not an effective method. Similarly, filling in missing value manually is time consuming. Although using a global constant to fill in the missing values is a simple method, it is not recommended. Filling in missing values with the most probable value is the most commonly used technique. Some methods like regression and decision tree induction are also used in this technique [4].

Noisy data is another important problem if real life data is used. Noise is a random error or variance in a measured variable. Clustering technique, scatter plots, box plots are helpful for detecting the outliers. And, some smoothing techniques like binning and regression are used to get rid of noise. Lastly, there may be inconsistencies in data. It is due to error made at data entry or data integration. It may be corrected by performing a paper trace [5].

Data integration is combining necessary data from multiple data sources like multiple databases, data cubes, or flat files. Some problems may occur during the data integration. To illustrate, if an attribute can be derived from another table it indicates to redundancy problem. Another problem is detection and resolution of data value conflicts. Since different representation, scaling or encoding can be used, for the same entity, attribute values can be different in different data sources. As a conclusion, we should be more careful in data integration in order not to face with such problems.

Data transformation is changing the data into convenient form for DM analysis. It includes smoothing, aggregation, generalization, normalization of data and attribute construction. Aggregation is summarization of data and it is used when building a data cube. An example of generalization is, changing the numeric attribute age into young, middle-aged, and senior. Normalization is changing the scaling of the value in order to be fall it within a desired range. Many methods are used for normalization. Some of them are min-max normalization, z-score normalization and normalization by decimal scaling. Attribute construction is building new useful attributes by combining other attributes inside the data. For instance, ratio of weight to height squared (obesity index) is constructed as a new variable so that it may be more logical and beneficial to use it in analysis.

Data reduction is changing the representation of data so that its volume becomes smaller while the information it includes is almost equal to the original data. It is important since the datasets are huge and doing analysis on this data is both time consuming and impractical. Methods for data reduction are listed in DM Process part.

Data gathering and data preprocessing are parts of data preparation. It includes choosing the right data then convert it into suitable form for the analysis. Data preparation is the most time spent part of the DM applications. In fact, about half of the time is spent in this part in DM projects. Much importance should be given to this part if we do not want to come up with any problem during the process.

2.2.2 Classification

Classification is an operation that examines the feature of the objects then assigns them to the predefined classes by the analyst. For this reason it is called as "supervised learning". The aim of it is to develop a classification or predictive model that increases the explanation capability of the system. In order to achieve this, it searches patterns that discriminate one class from the others. To illustrate, a simple example of this analysis is to predict the customers or non-customers who had either visited the website or not. The most commonly used techniques for classification are DT and ANN. And, it is frequently used in the evaluation of credit demands, fraud detection and insurance risk analysis [1].

2.2.3 Prediction

Prediction is a construction of a model to estimate a value of a feature. In DM, the term “classification” is used for predicting the class labels and discrete or nominal values whereas the term “prediction” is mainly used for estimating continues values. In fact, some books use the name ‘value prediction’ instead of ‘prediction’. Two traditional techniques namely, linear and nonlinear regression (R/NLR) and ANN are commonly used in this operation. Moreover, RBF is a newly used technique for value prediction which is more robust than traditional regression techniques [1].

2.2.4 Clustering

Clustering is an operation that divides datasets into similar small groups or segments according to some criteria or metric. Different from classification, there are no predefined classes in this operation. So it is called as “unsupervised learning”. It is just an unbiased look at the potential groupings within a dataset. It is used when there are suspected groupings in dataset without any judgments about what that similarity may involve. It might be the first step in DM analysis. Because, it is difficult to derive any single pattern or develop any meaningful single model by using the entire dataset. Clearly, constructing the clusters reduce the complexity in dataset so that the performance of other DM techniques are more likely to be successful. To illustrate, instead of doing a new sales company to all customers, it is meaningful firstly creating customer segments than doing the convenient sales companies to the suitable customer segments. Clustering often uses the methods like K-means algorithm or a special form of NN called a Kohonen feature map network (SOM).

2.2.5 Others

Affinity grouping is to find out what things go together whereas association rule is to find interesting associations or correlation relations within the dataset. It is also known as ‘market basket analyses’ since its most common application. It examines the relationships between items in a single transaction. An example of this operation is to find out the products sold at the same time.

Visualization and statistics give us information about where to start looking for and explanation. They just simply describe what is going on in a complicated database. They also provide us to distinguish the objects which are different from the general behavior of the data. Histograms, scatter plots, box plots, pie charts and 3_D plots are used for visualization. Some statistics like mean, median, frequency count etc. are also used in this analysis. Lastly, these analyses have successful applications in fraud detection [1].

There are also some advanced topics which are out of the scope of this thesis as web mining, spatial mining and temporal mining. Web mining is mining of the World Wide Web data which has a very huge size. There are several web mining tasks. One classification of Web mining activities is Web content mining, Web structure mining and Web usage mining [3]. Web content mining examines both the content of web pages and the results of web searching. Basic content mining is a type of text mining. It contains keyword searching, similarity measures, clustering and classification. Web structure mining is used to classify Web pages or to create similarity measures between documents. Web usage mining is the mining of Web usage data, or Web logs which is a listing of page reference data. It is used for several aims. As an example it analyzes the sequence of pages that a user accesses then makes a profile about that user. It can also be used to identify the quality and effectiveness of the pages of the site [3].

Spatial mining is the mining of the spatial data which have a spatial or location component. Some examples of the spatial mining applications are in geographic information systems, geology, resource management, medicine and robotics. It is used it to define the relationships with respect to direction. Several techniques that appear in DM Process part can directly be applied to spatial data. But, there also are some specific techniques and algorithms for spatial data mining [3].

Temporal mining is the mining of temporal data which includes multiple time points rather than one time point. There are several examples of temporal data. The information collected by satellites as images and sensory data, printouts of heartbeats and several different brain waves are some of the examples of it. Mining of the temporal data is a bit complicated. For instance, time series data can be clustered and similarities are found. But finding the similarities between the two time series data is a bit difficult. Association rules derived from temporal data include temporal aspects and relationships. Temporal mining can also be used in the combination of the Web usage and spatial mining.

2.3 MAIN DATA MINING TECHNIQUES

In DM operations, well-known mathematical and statistical techniques are used. Some of these techniques are collected in some heads like S-based, DT-based, NN-based and Distance-based. And the rest of them which are not covered with these four heads are listed in the others part. Here we only mentioned the commonly used or known techniques within these heads.

2.3.1 Statistical – based techniques

One of the commonly used S-based techniques is R. "Regression analysis is a statistical technique for investigating and modeling the relationship between variables"[101]

The general form of a simple linear regression is

$$y_i = \alpha + \beta x_i + \varepsilon_i$$

In this equation α is the intercept, β is the slope and ε is the error term, which is the unpredictable part of the response variable y_i. α and β are the unknown parameters to be estimated. The estimated values of α and β can be derived by the method of ordinary least squares as follows:

$$\hat{\beta} = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2}$$

and

$$\hat{\alpha} = \bar{y} - \hat{\beta}\,\bar{x}$$

Regression analysis must satisfy some certain assumptions. There assumptions are predictors must be linearly independent, error terms must be normally distributed and independent and variance of the error terms must be constant. If the distribution of error term is different than normal distribution then the GLM which is a useful generalization of ordinary least squares regression is used. The form of the right hand side can be determined from the data which is called nonparametric regression. This form of regression analysis requires a large number of observations since the data are used both to build the model structure and to estimate the model parameters. Robust regression is a form of regression analysis which circumvents some limitations of traditional parametric and non-parametric methods. It is highly robust to outliers. If the response variable is non-continuous then the logistic regression approach which is mentioned in Chapter 4 is used [102].

ANOVA which stands for analysis of variance, is another well-known S-based technique. “It is a statistical procedure for assessing the influence of a categorical variable (or variables) on the variance of a dependent variable” [103]. It compares the difference of each subgroup mean from the overall mean with the difference of each observation from the subgroup mean. If there is more variation between-groups differences, then the categorical variable or factor is influential on the dependent variable. One-way ANOVA measure the effects of one factor only, whereas two-way ANOVA measure both the effects of two factors and the interactions between them simultaneously. The F-test is used to measure the effects of the factors [104]. It must satisfy some certain assumptions as independence of cases, the distributions in each of the groups are normal and the variance of data in groups should be the same. When the normality assumption fails, the Kruskal-Wallis test which is a nonparametric alternative can be used [105].

2.3.2 Decision tree – based techniques

Decision trees are the tree shaped structures that are the most commonly used DM techniques. Construction of these trees is simple. The results can easily be understood by the users. In addition, they can practically solve most of the classification problems. In a DT model, there are internal nodes which devise a test on an attribute and branches show the outcomes of the test. At the end of the tree, leaf nodes, which represent classes, take place. During the construction of these trees, the data is split into smaller subsets iteratively. At each iteration, choosing the most suitable independent variable is an important issue. Here, the split which creates the most homogenous subsets with respect to the dependent variable should be chosen. While choosing the independent variable, some attribute selection measures like information gain, gini index etc. are used. Then, these splitting processes according to the measures continue until no more useful splits are found. In brief, DT technique is useful for classification problems and the most common types of decision tree algorithms are CHAID, CART and C5.0. These algorithms are mentioned in Chapter 4.

2.3.3 Neutral network – based techniques

NN supports us to develop a model by using historical data that are able to learn just as people. They are quite talented for deriving meaning from the complicated dataset that are difficult to be realized by humans or other techniques. To exemplify,

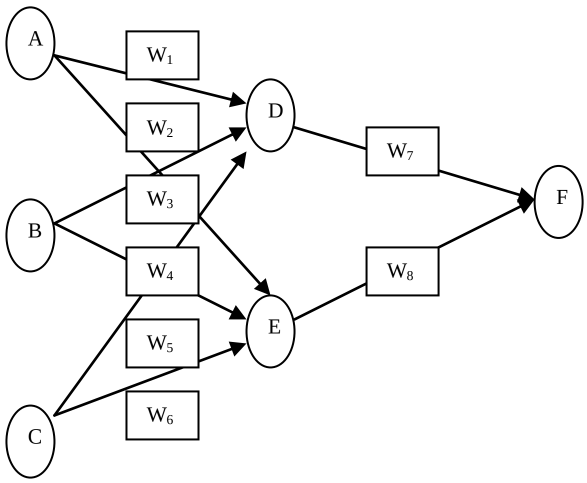


Figure 2.1 Example of a Neural Network Architecture

It simply consists of combining the inputs (independent variables) with some weights to predict the outputs (dependent variables) based on prior experience. In Figure 2.1, A, B and C are input nodes and they constitute the input layer. In addition, F is the output node and constitutes the output layer. Moreover, in most of the NNs, there are one or more additional layer between the input and output layer which are called “hidden layers”. In the Figure 2.1, D and E are the hidden nodes and constitute a hidden layer. The weights are also shown on the arrows between the nodes in the same figure. Additionally, if we look at the strengths and weaknesses of this technique firstly, it is more robust than DT in noisy environments. Then, it can improve its performance by learning. However, the model developed is difficult to understand. Moreover, learning phase may fail to converge. Input attribute value must also be numeric. As a result, NNs are useful for most prediction and classification operations when just the result of the model is important rather than how the model finds it.

Backpropagation is the most commonly used learning technique. It is easily understood and applicable. “It adjusts the weights in the NN by propagating weight changes backward from the sink to the source nodes” [3].

Perceptron is the simplest NN. In this architecture, there is a single neuron with multiple inputs and one output. A network of perceptrons is called a multilayer perceptron (MLP). MLP is the simple feedforward NN and it has multiple layers [3].

Radial basis function network is a NN which has three layers. In hidden layers Gaussian activation function is used whereas in output layer a linear activation function is used. Gaussian activation function is a RBF with a central point of zero. “RBF is a class of functions whose value decreases (or increases) with the distance from a central point” [3].

2.3.4 Hierarchical and Partitional techniques

Cluster analysis identifies the distinguished characteristics of the dataset, and then divides it into partitions so that the records in the same group are similar and between the groups are different as much as possible. The basic operation is the same in all clustering algorithms. Each record is compared with the existing clusters then it is assigned to the cluster whose centroid is the nearest. Later, centroids of the new clusters are calculated and once again each record assigned to the new cluster with the closest centroid. At each iteration the class boundaries, which are the lines equidistant between each pair of centroids, are computed. This process continues until the cluster boundaries stop changing. As a distance measure, most of the clustering algorithms use the Euclidean distance formula. Certainly, nonnumeric variables must be transformed in order to be used by this formula.

Hierarchical clustering techniques can generate sets of clusters whereas partitional techniques can generate only one set of clusters. So in partitional techniques user has to specify the number of clusters. In an agglomerative algorithm, which is one of the hierarchical clustering techniques, each observation is accepted as one cluster. Then it continues to combine these clusters iteratively until obtaining one cluster. On the other hand, in K-means clustering, which is one of the partitional clustering techniques; observations are moved among sets of clusters until the desired set is obtained [3].

2.3.5 Others

Genetic algorithm is an optimization type algorithm. It can be used for classification, clustering and generating association rules. “It has five steps:

1. Starting set of individuals, P.
2. Crossover technique.
3. Mutation algorithm.
4. Fitness function
5. Algorithm that applies the crossover and mutation techniques to P iteratively using the fitness function to determine the best individuals in P to keep. The algorithm replaces a predefined number of individuals from the population with each iteration and terminates when some threshold is met” [3].

This algorithm begins with a starting model which is assumed. Then using crossover algorithms, it combines the models to generate new models iteratively. And, a fitness function selects the best models from these. At the end, it finds the “fittest” models from a set of models to represent the data.

2.4 APPLICATION AREAS

Today main application areas of DM are as follows:

Marketing:

- Customer segmentation
  - Find clusters of ‘model’ customers who share the same characteristics: interest, income level, spending habits, etc.
- Determining the correlations between the demographic properties of the customers
- Various marketing campaign
- Constructing marketing strategies for not losing present customers
- Market basket analysis
- Cross-market analysis
  - Associations/correlations between product sales
  - Prediction based on the association information
- Customer evaluation
- Different customer analysis
  - Determine customer purchasing patterns over time
  - What types of customers buy what products
  - Identifying the best products for different customers
- CRM
- Sale estimation

Banking:

- Finding the hidden correlations among the different financial indicators
- Fraud detection

- Customer segmentation
- Evaluation of the credit demands
- Risk analysis
- Risk management

Insurance:

- Estimating the customers who demand new insurance policy
- Fraud detection
- Determining the properties of risky customers

Retailing:

- Point of sale data analysis
- Buying and selling basket analysis
- Supply and store layout optimization

Bourse:

- Growth stock price prediction
- General market analysis
- Purchase and sale strategies optimization

Telecommunication:

- Quality and improving analysis
- Allotment fixing
- Line busyness prediction

Health and Medicine:

- Test results prediction

- Product development
- Medical diagnosis
- Cure process determination

Industry:

- Quality control and improvement
    - Product design
        - Concept design
        - Parameter design (design optimization)
        - Tolerance design
    - Manufacturing process design
        - Concept design
        - Parameter design (design optimization)
        - Tolerance design
    - Manufacturing
        - Quality monitoring
        - Process control
        - Inspection / Screening
        - Quality analysis
    - Customer usage
        - Warranty and repair / replacement
- Logistic
- Production process optimization

Science and Engineering:

- Analysis of scientific and technical problems by constructing models using empirical data

[11, 106].

CHAPTER 3

DATA MINING APPLICATIONS ON MANUFACTURING DATA

The environment in which the tools and labor are used to make things for use or sale is called "manufacturing/process environment". There are various types of these environments. Two of them will be described in the following, namely dynamic environments and discrete environments. An example of the dynamic environment is the chemical process industry. It is dynamic because we can not divide the products under process into different parts. Whereas the most common example of the discrete environment is the parts industry in which we can divide the products in process to different parts. There exist many common and different properties of them.

Our aim is to control the quality characteristics of the processes described above and to keep the quality in the specified target levels. There may be some variations around these target levels, and the most significant question here is to determine the main reasons that cause these variations. We will call these reasons as the "causes of variations".

The first kind of these causes of variation is called the common or chance causes, which are seen as normal under the usual manufacturing process. Actually these may occur because of the natural reasons as the change in the temperature or humidity, or they may be sourced from the changes out of the control of the factory as technological or economical changes. The most remarkable property of these is that they are not removable and seemed as an inherent part of the process.

The second kind of these variations is the assignable or special causes of variation. These are originated from the faults of the process inside such as the faults of labors and machines. More specifically the reasons may be the machine malfunctions, the misuse of the raw materials, operator mistakes, etc. The most significant property of these is that these are controllable. Hence, the aim of the quality control techniques is to determine and remove these types of causes.

Most of the industries in dynamic environments use automatic process control systems. However in these environments the observation series which represent the deviations from target can be autocorrelated. Since there are such difficulties, we will only deal with discrete environments rather than dynamic processes.

In the discrete environments (and also in dynamic environments) there are lots of quality control and improvement activities in usage. These are used in all stages of product development, namely in product design, in manufacturing process design, in manufacturing and in customer usage. At all of these cases different activities are performed. In this study quality control and improvement activities that occur during manufacturing stage is studied. These include quality monitoring, process control, inspection/screening and quality analysis.

3.1 QUALITY CONTROL AND IMPROVEMENT ACTIVITIES

In both discrete and dynamic environments, various quality control and improvement activities can be performed. These are explained in Table 3.1[100].

Table 3. 1 Quality control and improvement activities and methods

Product developm ent stage	**Quality control and improvement activity**	**Description of the activity**	**Methods**
Product design	Concept design	Designer examines a variety of architectures and technologies for achieving the desired function of the product and selects the most suitable ones for the product. Involves innovation to reduce sensitivity to all noise factors.	QFD, Pugh's concept selection, TRIZ, technological forecasting, etc.
	Parameter design	The best settings are determined for the control factors, which yield desired performance of the product no matter how common sources of variation behave. During parameter design we assume wide tolerances on the noise factors and assume that low-grade components and materials would be used.	Design of experiments, response surface modeling and analysis, ANOVA, regression, optimization
	Tolerance design	If at the end of parameter design we cannot reduce sensitivity of the design to noise factors sufficiently, then we make a trade-off between reduction in the variability and increase in the manufacturing cost. That is we selectively reduce tolerances and selectively specify higher grade material in the order of their cost effectiveness. Inclusion of a suitable compensation	Statistical tolerancing, cost analysis, etc.

Table 3. 1 (cont'd.)

		system such as a feedback control system can be considered as a tolerance factor to be optimized along with the component tolerances.	
Manufacturing process design	Concept design	Similar to the concept design of products. Involves innovation to reduce unit-to-unit variation.	QFD, Pugh's concept selection, TRIZ, technological forecasting, etc.
	Parameter design	Similar to the parameter design of products.	Design of experiments, response surface modeling and analysis, ANOVA, regression, optimization
	Tolerance design	Similar to the tolerance design of products.	Statistical tolerancing, cost analysis, etc.
Manufacturing	Quality monitoring	The goal is to recognize promptly the existence of an assignable cause, finding the root causes and correcting for it.	Statistical process control (control charts), Principal Component Analysis, etc.
	Process control	The goal is to send information about errors or problems discovered in one step to the next step in the process so that variation can be reduced.	Methods of compensating for known problems such as feedback control, feed forward control,

Table 3. 1 (cont'd.)

			manual adjustments
	Inspection / Screening	All units produced are inspected (measured) and the defective ones are discarded or repaired.	Pattern recognition, automated inspection
	Quality analysis	Finding factors that significantly affect quality, modeling relations between input and output characteristics of quality, and predicting what quality will be for a given set of input parameters for providing feedback to product/process design (re-design) and other corrective actions for quality improvement.	ANOVA, Regression, Classification, Clustering, Rule Induction, etc.
Customer usage	Warranty and repair / replacement	Compensate the customer for the loss caused by a defective product through warranty or repair or replacement	Replacement analysis

In this thesis, it is focused on process parameter design (design optimization) and quality analysis activities that occur during manufacturing for the purpose of quality improvement. And, quality analysis activities are classified as follows:

Process and product quality description:

- Reducing attributes/variables, which do not affect the quality significantly
- Simplifying the DM task by reducing the number of attributes/variables
- Ranking the attributes/variables to based on their significance
- Identifying significant attributes/variables for quality
- Identifying how low, medium and high yielding products are naturally grouped in data
- Finding the most probable causative factor(s) that discriminate between low yield and high yielding products

Predicting quality:

- Predicting one of the determinant of the quality for real outputs

Classification of quality:

- Classifying a quality characteristic of interest for nominal or binary outputs
- Classification of faults

Parameter optimization:

- Optimization of process/product parameters based on the learned characteristics of the cases yielding highest quality.

3.2 LITERATURE REVIEW

In this part of the thesis, a review of literature on the use of DM techniques for quality improvement is presented. The aim of this study is to overview the DM researches for quality improvement for the use of the academicians and practitioners. We want to achieve this by presenting; which DM techniques were used for which quality task, in which industries that these techniques were implemented, what type of data was used in the DM applications for quality improvement. The scope of this work is limited to manufacturing quality problems; predicting and classifying the quality, process and product quality description and parameter optimization. Related scholarly articles in the literature published between the years January 1995-February 2007 are covered. The journals and databases given below were scanned with the keywords "DM, quality, quality improvement, manufacturing, decreasing failure-faults, Principle component analysis (PCA), regression (R), DT, ANN, rough set theory(RST), Bayesian belief networks (BNN), genetic algorithms (GA), clustering, support vector machines (SVM), multidimensional scaling keyword" and some combinations of these words.

Journals:

- International Journal of Production Economics
- IIE Transactions
- Quality and Reliability Engineering International
- Quality Engineering
- Journal of the Operations Research Society
- Operations Research
- Management Science
- European Journal of Operations Research
- IEEE
- International Journal of Production Research
- Technometric
- Journal of Quality Technology

Databases:

- EbscoHost - Academic Search Premier
- Emerald Management Xtra
- ENGnetBASE (E-Handbook, E-Dictionary)
- MathSciNet (Mathematical Reviews and Current Mathematical Publications)
- Oxford Online Journals
- SCOPUS
- Springer Lecture Notes
- Taylor & Francis Online Journals
- Web of Science - Science Citation Index Expanded 1945

Following the DM Process presented in section 2.1, studies in literature on DM application are classified and presented in Appendix A. There exist 12 columns in this table. In the first column, name/names of the researchers and the published date of the articles are placed. In the second column, the main aim of the articles is mentioned briefly. Then the name of the manufacturing or process where the application is done is written. In the next column, data source and the numbers of records which consist of train, test and verification data sets are appeared. In the fifth column there exist the types and the number of the inputs whereas in the sixth column there exist the types and the number of outputs. In the next column, the software used in the implementations is presented. Later, quality tasks of the articles are written. In the following columns, data collection methods, DM tasks and tools are presented. And, in the last column results of the implementations whether they are successful or not are shown.

Examining the table in Appendix A, one can conclude that the DM tools were most commonly used in semiconductor manufacturing for quality improvement purposes [19, 26, 36, 46, 67, 79 etc.]. There also exist applications in integrated circuit manufacturing [58, 59 etc.], steel production [29, 31 etc.], plastic manufacturing [57, 69, 73 etc.] etc. [107]. Secondly, observational data was often used in the implementation (Figure 3.1). Experimental data was collected by using different designs of experiment. Moreover, although output data is usually continuous or binary types, there are different types of input data.

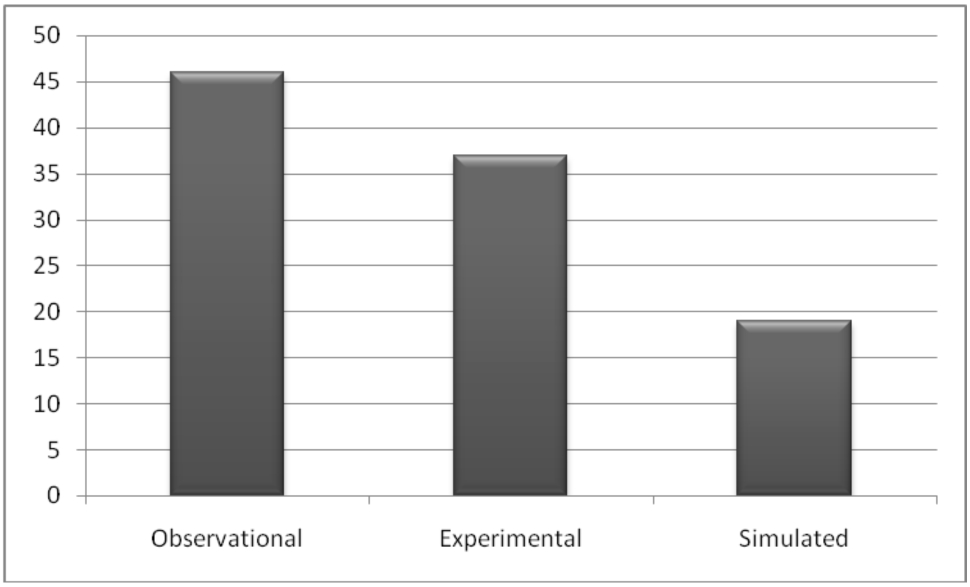


Figure 3.1 Source of the data used in articles

In Figure 3.2, it is seen that the number of articles published between January 1995 and February 2007 increase during the years. Especially, there are a lot of articles published in years 2002 and 2006. Furthermore, there does not exist any article which is in the scope of this study in 1995 and 1996. Therefore, we can reach a conclusion that DM is a newly used method on quality improvement in manufacturing.

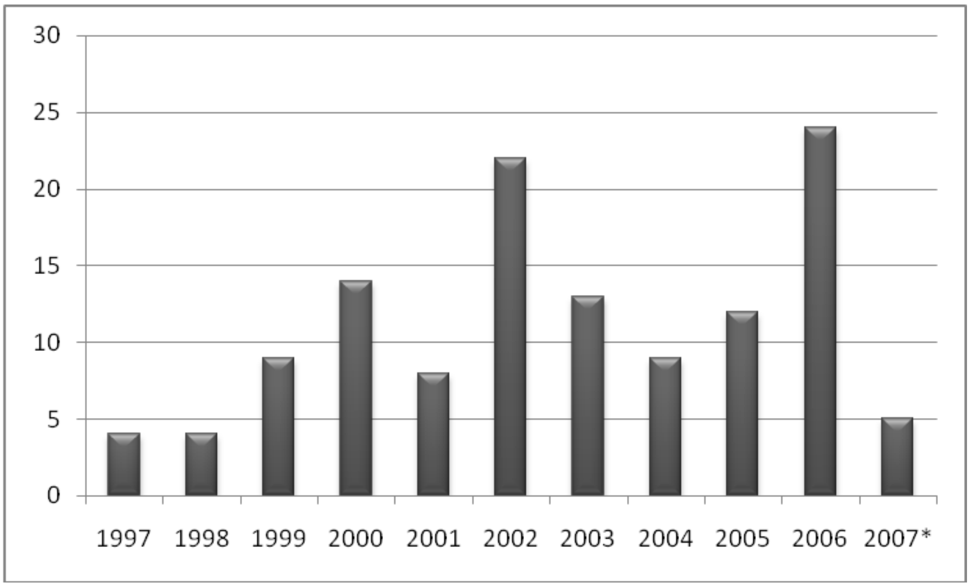


Figure 3.2 The number of articles using DM on quality improvements during the years

In applications, different softwares as Neuro Shell Predictor, Qnet for Windows, Neural Works Predict software package, Professional II/Plus 2000, Rosetta 2005, Fuzzy TECH etc. were used. For NNs, Matlab NN Toolbox was commonly used. Besides, some programming languages as C, C++, and Visual Basic 5.0 were also used for NNs. Moreover, several statistical package programs as SPSS, Minitab, SAS and Statistica were also used in applications.

Additionally, Quality task/ DM task and DM task/ DM tool tables are derived in order to see the relationship between them clearly.

Table 3.2 Quality task/ DM task

Quality Task \ DM Task	Data compression	Dimension reduction	Classification	Prediction	Clustering	Optimization
Process/product quality description	✓	✓			✓	
Classification of quality			✓			
Predicting quality				✓		
Parameter optimization						✓

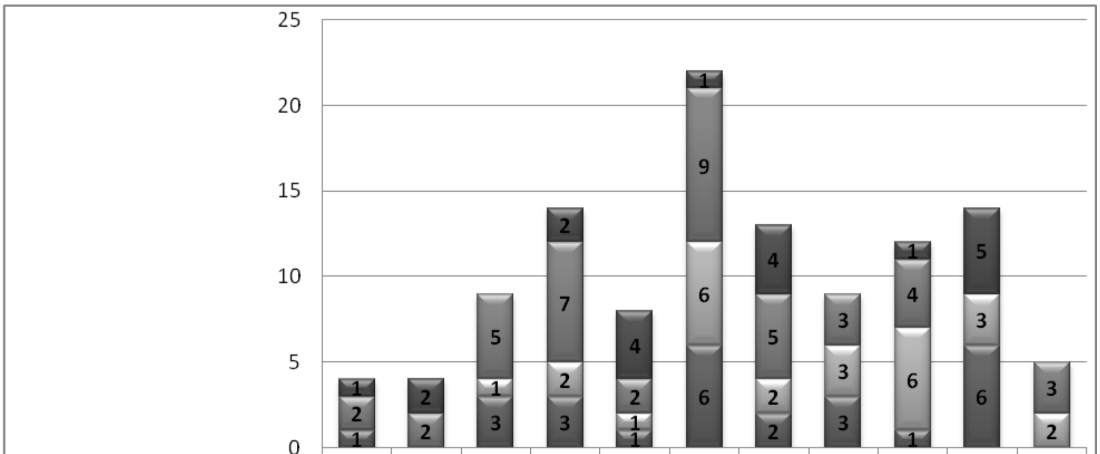
Table 3.3 DM task/ DM tool

DM task \ DM tool	S-based	DT-based	NN-based	R-based	CT	SVM	CBR	GA	SE	GSA	SA	SQP	Hierarchical	Partitional
Data compression	✓													
Dimension reduction	✓	✓	✓	✓				✓	✓	✓				
Classification	✓	✓	✓	✓	✓	✓		✓						
Prediction	✓	✓	✓		✓		✓							
Clustering													✓	✓
Optimization	✓		✓					✓			✓	✓		

S-based: Statistical based, DT-based: Decision tree based, NN-based: Neural network based, R-based: Rule based, CT: Combining techniques, SVM: Support vector machines, CBR: Case based reasoning, GA: Genetic algorithms, SE: Subjective and empirical approach, GSA: Grey superior analysis, SA: Simulated annealing, SQP: Sequantial quadratic programming method, Hierarchical: Hierarchical methods, Partitional: Partitional methods

In Table 3.3, DM tools are collected and presented in some general headings. These headings and the DM tools under these headings are shown in Appendix B. If we look the tables above, we see that most of the DM tools such as DT-based, NN-based and some of the traditional S-based techniques as R, PCA, etc. were used in manufacturing quality problems. In data compression one of the S-based techniques PCA was implemented. For dimension reduction purpose, besides S-based tools as ANOVA, R, correlation analysis (CA) and nonlinear regression (NLR), DT-based, NN-based and R-based tools as RST were also used in literature [e.g. 16, 17, 32, 49, and 64]. Similarly, the S-based tools as R, NLR and ANOVA, DT-based, NN-based, case based reasoning(CBR) and some combining techniques as fuzzy neural networks (FNN) were the tools used for prediction [e.g. 20, 26, 40, and 70]. For classification one of the S-based tools Naive Bayesian classifier, DT-based, NN-based, R-based, combining techniques as composite classifiers (CC), GA and SVM were implemented [e.g. 18, 23, 37, 44, and 51]. In clustering, partitional methods as K-means and SOM, and hierarchical methods as agglomerative clustering were used [e.g. 26, 29, and 46.] And lastly the S-based methods as RSM and TM, NN-based methods, GA, SA and SQP were used for optimization purpose [e.g. 25, 30, 42, 50, and 68].

Next, we draw a graph which shows the number of each quality tasks studied during the years. According to the Figure 3.3, predicting the quality is usually the most studied quality task in articles. Especially in year 2002 it is frequently studied. If we look at the parameter optimization, we see that it is mostly studied in years 2001, 2003 and 2006. Similarly, classification of quality is most studied in years 2002 and 2005 whereas process/product quality description is in years 2002 and 2006.



	1997	1998	1999	2000	2001	2002	2003	2004	2005	2006	2007*
Parameter optimization	1	2		2	4	1	4		1	5	
Predicting quality	2	2	5	7	2	9	5	3	4		3
Classification of quality			1	2	1	6	2	3	6	3	2
Process/product quality description	1		3	3	1	6	2	3	1	6	

Figure 3.3 The number of each quality tasks studied during the years

Lastly, DM tools used for the categories of the quality tasks were drawn. In Figure 3.4, it is clearly seen that statistical based techniques were commonly used for process/product description. Moreover we already know that predicting quality is the most frequently studied DM task in the articles. And, according to the Figure 3.5, in more than half of these studies NN-based methods were used. For classification of quality, again NN-based methods were most commonly used which is seen in Figure 3.6. But this time, the usage of the other methods as DT-based, R-based and CT are close to NN-based methods. And, according to the Figure 3.7 GA is the most often used DM tool for parameter optimization.

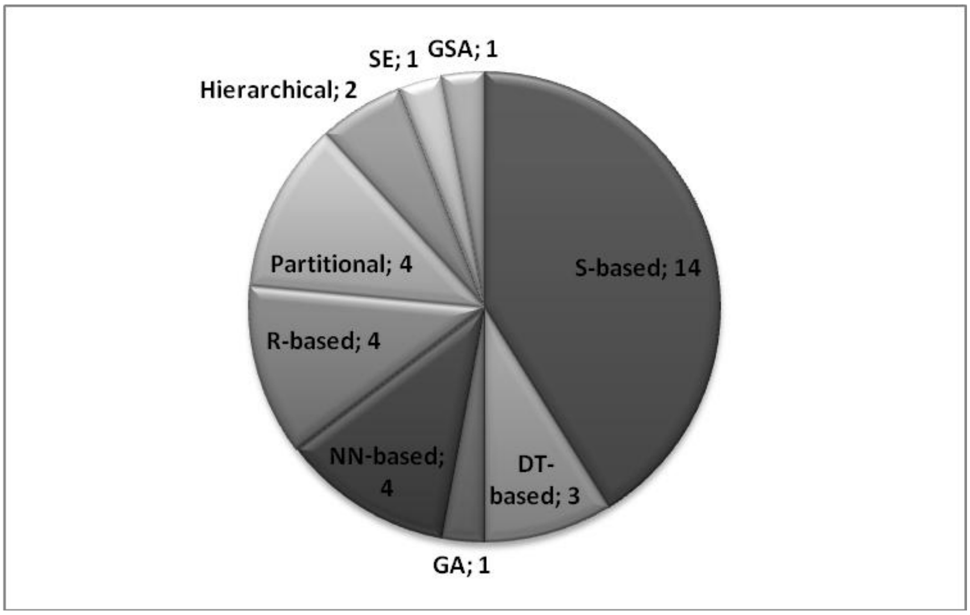


Figure 3.4 DM tools used for Process/product quality description

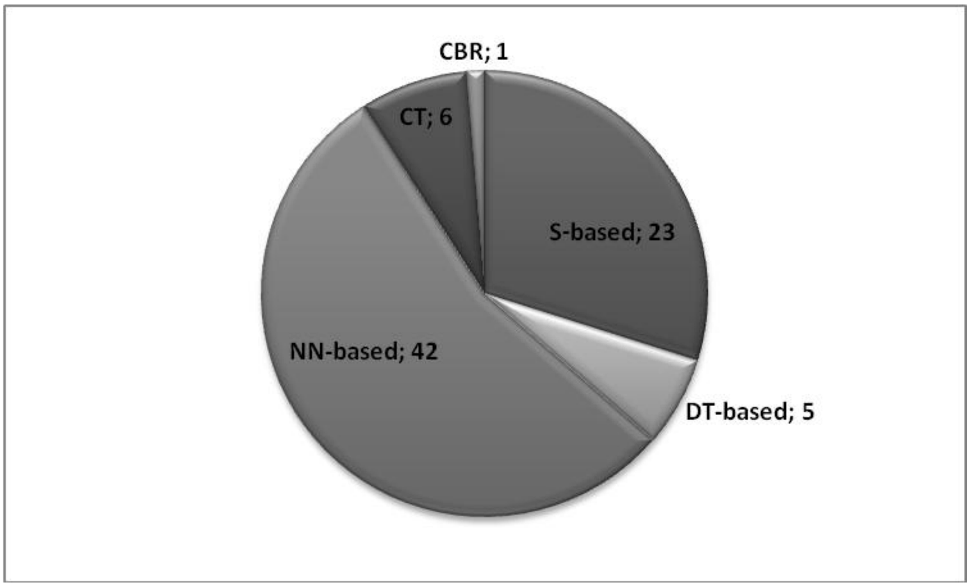


Figure 3.5 DM tools used for predicting quality

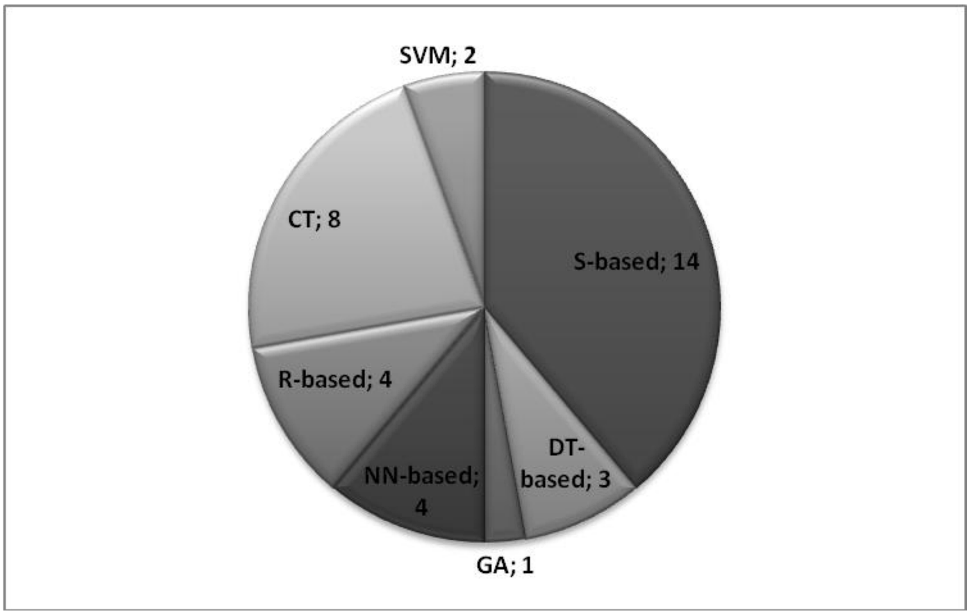


Figure 3.6 DM tools used for classification of quality

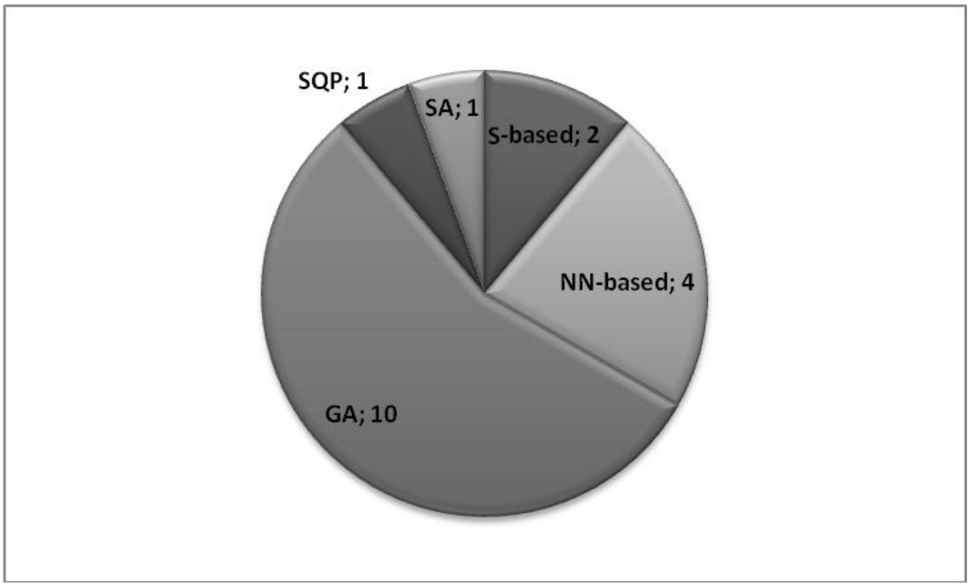


Figure 3.7 DM tools used for parameter optimization

In the last part of this study, some general information derived from the articles is presented. One of them is, data types were generally changed by discretization [e.g. 16, 19 and 37] or the interval of the data was changed by normalization in order to prepare the data for the analysis [e.g. 23, 94 and 95]. Another one is, S-based methods and NN-based methods [e.g. 21, 35 and 56] or regression and DT-based methods [e.g. 17, 26 and 79] were commonly used together in order to see the difference between the traditional statistical techniques and DM techniques. Moreover, FST was often used in combination with NN-based methods [e.g. 20, 40, 52 and 72]. And, the last derived information is, after predicting or classifying the quality by NN-based methods parameter optimization was usually done by GA [e.g. 28, 56, 58 and 76].

3.2.1 Thesis and dissertations on data mining applications on manufacturing

In this thesis, thesis and dissertations are searched between the years 2002 and 2007. To this end, three dissertations are found which are in the scope of this thesis. The first one was written by Stacey Lee Schertel in 2002 [95]. The purpose of that research is to understand the possible uses of DM in the Textile Industry, specifically in a spinning mill. Data was collected from a spinning mill operation. It was cleansed and merged to create a data warehouse. In case study it is aimed to determine the potential benefits afforded by using DM to analyze factors affecting process and product quality in a spinning mill. Data used in that study consisted of 55 input and one output variable. The number of records was 181. Data types of the inputs were continuous and nominal and it was integer for output. Regression, decision tree and neural network approaches were used. Besides, a new DM model was proposed by Schertel which contains 6 major steps with 28 sub steps.

The second dissertation was written by David R. Forrest in 2003 [96]. The purpose of that research is to build a model understanding the contributions of the subcomponents of a large complex system to the performance characteristics of the overall system. This model uses a top down approach to design the intermediate variables summarizing the contributions of subsystems to a high-level system variable. Each submodel estimates the variable of interest based on a set of predictor variables. A high-level model, the metamodel, then combines the estimates from the subsystems to produce an overall estimate of the variable of interest. Namely, methodology include identifying the variable of interest and the subsystems, estimating the variable of interest from the each of the sub-systems, estimating the variable of interest from the subsystem estimates and testing the resulting model. And, the metamodel results are compared with principal components logistic regression. Two types of the data were used in this research, one of which was simulated and the other was observational data coming from the semiconductor manufacturing plant. The aim of the case study is to find out the root causes of the defect rate. In real data, eight data sets whose data sizes were quite different were used to form submodels. Linear, generalized linear and logistic regression models were used to form the submodels.

The last dissertation was written by Shital Chamanlal Shah in 2005 [97]. The purpose of this research is mining noisy, temporal, and high dimensionality data in medical and energy applications. Since the medical applications are out of the scope of this research, only energy applications are mentioned. In these applications, the alarm system was developed for detection and avoidance of water chemistry faults at two commercial power plants. The proposed system consists of data preprocessing, learning, knowledge base, prediction, alarm generation, and display modules. A decision tree algorithm was applied and extracted rules from the preprocessed data. This system effectively identified normal and faulty operating conditions for two water chemistry systems.

CHAPTER 4

MATERIALS AND METHODS USED IN THIS THESIS

4.1 SPSS CLEMENTINE

Clementine is a special program that created by SPSS for DM applications. It provides us to use advanced modeling technology with easy use. It is useful for organizations. It uses CRISP-DM methodology to improve decision making. It works in tree steps as reading data, running the data and deriving the results. This pattern is named as a data stream and applications are done by creating and modifying data streams. Figure 4.1 shows an example of a simple stream.

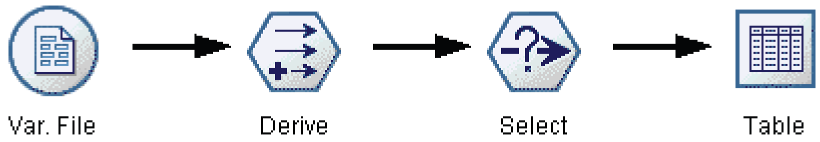


Figure 4.1 A simple stream

Clementine's visual interface is shown in Figure 4.2. It provides us to use data sources as var. file, SAS file, excel and database. It includes modeling algorithms, such as prediction, classification, segmentation, and association detection. Additionally it covers some record operations as aggregate, sort, merge and some field operation as filler, binning and reclassify. It also involves many graphs options. And lastly, in output part, there are many options to evaluate our models, data and results.

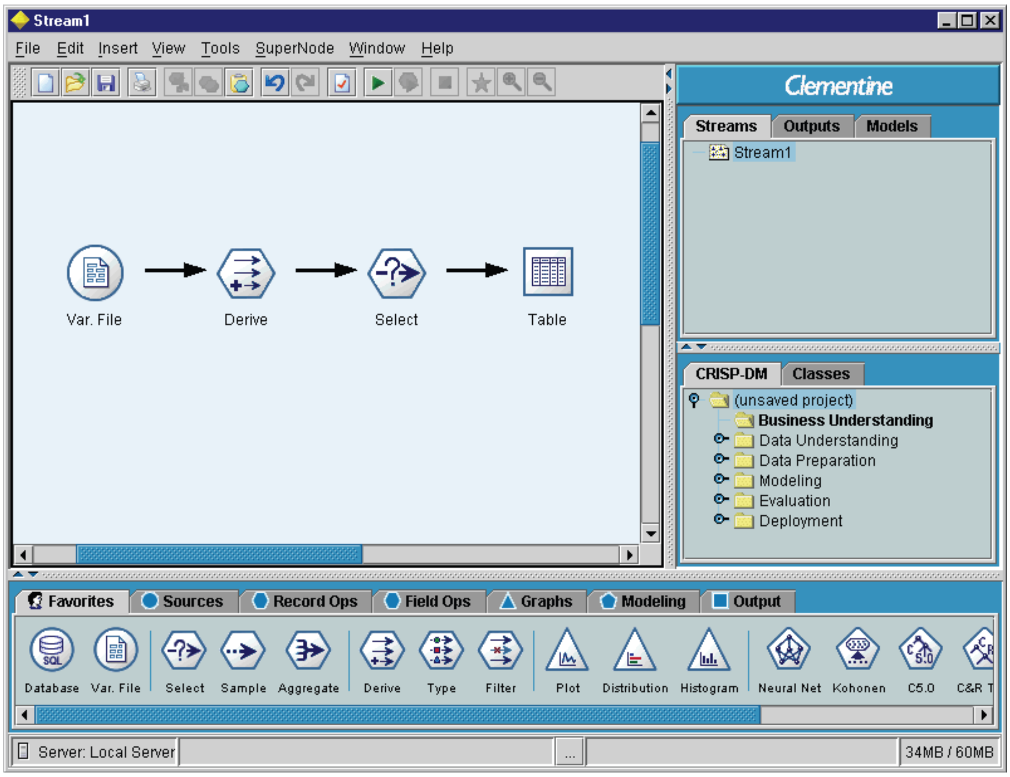


Figure 4.2 SPSS Clementine 10.1 User Interface

4.2 C5.0 ALGORITHM

C5 algorithm is proposed by Ross Quinlan in 1998. It is an extension of C4.5 algorithm. It generates tree using the concept of information entropy. It uses the normalized Information Gain which is difference in entropy to choose the best splitting. Entropy(S) is the measure of how random the class distribution is in S where S represents training data. During the construction of the tree, S divided into subsets S_a1, S_a2, S_a3,...,S_an. And the information gain of this splitting is then computed as Entropy(S) − Entropy(S_a1) − Entropy(S_a2) − ... − Entropy(S_an).

C5.0 has a number of improvements according to C4.5. Some of these are:

- Speed - C5.0 is significantly faster than C4.5 (several orders of magnitude)
- Memory Usage - C5.0 is more memory efficient than C4.5

- Smaller Decision Trees - C5.0 gets similar results to C4.5 with considerably smaller decision trees.
- Support For Boosting - Boosting improves the trees and gives them more accuracy.
- Weighting - C5.0 allows you to weight different attributes and misclassification types.
- Winnowing - C5.0 automatically winnows the data to help reduce noise.[10]

4.3 CLASSIFICATION AND REGRESSION TREES

The CART is a powerful method that provides us to predict or classify future observations. It is introduced by Breiman et al. (1984). It recursively splits the training records into two segments with similar output field values. At each step, it examines the input fields to find the best split which is found by measuring the reduction in an impurity index that results from the split. Both for the target and predictor fields, data types can be range or categorical. All splits are binary. It works successfully even though there are large number of fields and lots of missing values in data. The same predictor field can be chosen several times at different levels in the tree.

Different impurity measures are used depending on the type of the target type to find the best split. Gini or twoing impurity measures are the ones used for symbolic target fields. And for continuous targets, the least-squared deviation (LSD) method is used automatically.

The Gini index $g(t)$ at a node t in a CART tree is defined as,

$$G(t) = \sum_{j \neq i} p(j \mid t) p(i \mid t)$$

where i and j are categories of the target field, and

$$p(j \mid t) = \frac{p(j,t)}{p(t)}$$

$$p(j,t) = \pi(j) \frac{N_j(t)}{N(j)}$$

$$p(t) = \sum_j p(j,t)$$

Here $\pi(j)$ represents prior probability value for category j. $N_j(t)$ is the number of records in category j of node t and N_j is the number of records of category j in the root node.

The Gini index takes its maximum value when the records in a node are evenly distributed across the categories and it equals to 0 when all records in the node belong to the same category.

The Gini criterion function (s, t) for split s at node t is defined as

$$\Phi(s,t) = g(t) - p_L g(t_L) - p_R g(t_R)$$

where p_L is the proportion of records in t sent to the left child node, and p_R is the proportion sent to the right child node. The proportions p_L and p_R are defined as

$$p_L = \frac{p(t_L)}{p(t)}$$

and

$$p_R = \frac{p(t_R)}{p(t)}$$

The split s is chosen to maximize the value of $\Phi(s, t)$.

The twoing index is based on splitting the target categories into two superclasses, and then finding the best split on the predictor field based on those two superclasses. The superclasses $C1$ and $C2$ are defined as

$$C_1 = \{ j : p(j \mid t_L) \geq p(j \mid t_R) \}$$

and

$C_2 = C - C_1$

where C is the set of categories of the target field, and $p(j|t_R)$ and $p(j|t_L)$ are $p(j|t)$, as defined as in the Gini formulas, for the right and left child nodes, respectively.

The twoing criterion function for split s at node t is defined as

$$\Phi(s,t) = p_L p_R \left[\sum_j \left| p(j \mid \mathrm{t_L}) - p(j \mid \mathrm{t_R}) \right| \right]^2$$

where t_L and t_R are the nodes created by the split s. The split s is chosen as the split that maximizes this criterion.

The LSD measure $R(t)$ is simply the weighted within-node variance for node t, and it is equal to the resubstitution estimate of risk for the node. It is defined as

$$R(t) = \frac{1}{N_W(T) \sum_{i \in t} w_i f_i (y_i - \overline{y}(t))^2}$$

where $N_W(t)$ is the weighted number of records in node t, w_i is the value of the weighting field for record i (if any), f_i is the value of the frequency field (if any), y_i is the value of the target field, and $y(t)$ is the (weighted) mean for node t. The LSD criterion function for split s at node t is defined as

$$\Phi(s,t) = R(t) - p_L R(t_L) - p_R R(t_R)$$

The split s is chosen to maximize the value of $\Phi(s,t)$.

4.4 CHI-SQUARED AUTOMATIC INTERACTION DETECTOR

The CHAID which is introduced by Kass (1980) is a highly efficient statistical technique for segmentation. It works with all types of variables. It generates trees where more than two branches can attach to a single node. It deals with missing values by treating them all as a single valid category.

The statistical tests are used to identify significant predictors and optimal splits. *F* test is used if the target field is continuous whereas a chi-squared test is used if the target field is categorical. It merges values which are statistically homogeneous with respect to the target variable. Also it keeps all other values that are heterogeneous. Then it selects the best predictor to form the first branch in the decision tree. After the splitting process, the resulting child nodes are made of a group of homogeneous values of the selected field. At the end, after recursions this process terminates when a fully grown tree is constructed.

4.5 LOGISTIC REGRESSION

Logistic regression is a statistical technique that classifies observations based on a set of input fields. It is similar to linear regression. Linear regression works with numeric target fields whereas logistic regression works with symbolic target fields. Types of the data for input fields can be either symbolic or numeric. It generates a set of equations that link the input field values with the probabilities associated with each of the output field categories. These equations assign probabilities of membership between observations and possible output categories. Then one of these probabilities is the highest for a record. The category with the highest probability is selected as the predicted output value.

CHAPTER 5

CASE STUDY: ANALYSIS OF CUSTOMER QUALITY PERCEPTION DATA FOR DRIVER SEAT QUALITY IMPROVEMENT

5.1 DATA SET

Data used in this study is obtained from the other ongoing study [99]. The aim of gathering this data is to identify important factors in customer satisfaction from the driver seat. In this way, costumers' expectations and demands link with design and some technical properties of the seat. And hence by identifying the necessary improvable areas, more innovative and original designs can be developed for the driver seats. The result of this study is very important since driver seat is the one of the most important factors that affects both the driver comfort and buying decisions.

Data was gathered by questionnaire approach. Firstly one of the businesses, activating in automotive sector, was selected. Then the project members worked together with the staffs from different departments as design, marketing, manufacturing and quality in order to improve the design of the driver seats. First of all, they defined the profiles of the costumers by using the information from the marketing and service departments. Then data was gathered from different costumer profiles in some certain vehicle selling places and services. And some anthropometric measures were taken and detailed face to face interviews were done with the costumers.

Costumer segments to which the questionnaires were applied is shown in Table 5.1.

Table 5.1 Costumer segments to which the questionnaires were applied

Costumer segments	Categories	Number of people
Gender	Man	77
	Woman	3
Age interval	<35	33
	36-45	25
	45>	22
Education	Primary education	18
	High school	29
	Academy	9
	University or graduate	24
Job	Free job	30
	Working in an institution	41
	Others(retired and not workings)	9
Marriage	Living with family	75
	Living alone	5
Income	<1500	26
	1500-3500	37
	>3500	17
User of the specified vehicle or not	User or used before the specified vehicle	53
	Not used before	27
Function of the vehicle for the user	Carry load	7
	Carry passenger	31
	Carry both load and passenger	15
Weight(for men)	<55	2
	56-71	20
	72-99	51
	>100	7
Height(for men)	<157	0
	158-172	34
	173-185	42
	>186	4

The questionnaire used in this study is shown in appendix B. We can put there a small part of the questionnaire because of the privacy agreement between the business and the project group. It consists of five parts. In the first part there are some questions about customers' identity information. The questions in the second part are about the usage aim of the customers to the specified vehicle. In the third part customers' desires and expectations is written and recorded. In the fourth part there are some detailed questions about driver seat comfort, appearance and usage.

And in the fifth part there exist customers' anthropometrics measures which effect driver seat comfort.

Data obtained at the end of the study consists of 89 column/fields and 80 row/records. One of the fields is our output field and it shows the satisfaction of the customers from the driver seat. It takes values from one to seven. One represent customer isn't pleased with the driver seat whereas seven represents the highest customer satisfaction. Most of our input fields are in binary type. And the others are in nominal and interval types.

5.2 DATA PREPROCESSING

First of all we used the anomaly detection method to check whether there are outliers in data or not. This method is one of the modeling node references in Clementine. It implements clustering algorithms to find outlier cases. Method did not determine any outlier observation.

Later the inconsistencies within the attributes were examined. For nominal and ordered type variables, data was filtered and we checked whether there were any spelling errors or not. And, for interval type values, box plots were used.

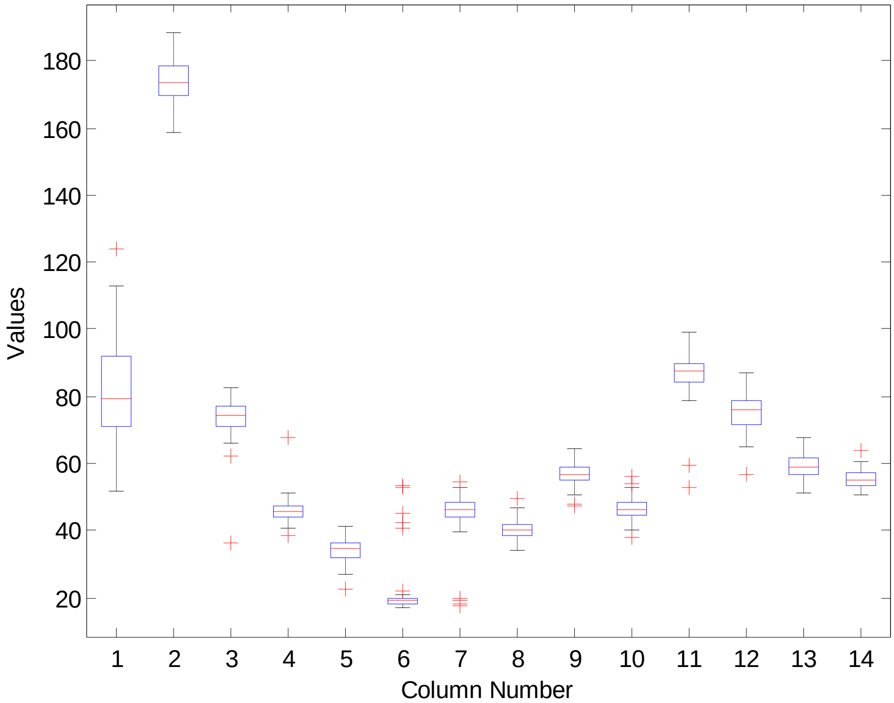


Figure 5.1 Box plots of continuous variables

Box plots showed that there were some inconsistencies. Then we examine the data and as expected we detect some spelling errors just as in the plots. For 25^{th}, 30^{th}, 31^{st}, 58^{th} and 73^{th} surveys the values in question 81 was replaced with the corresponding values in question 82. For the 27^{th} survey, the value in question 86 was changed considering the value in question 87. For the 49^{th} survey, the values in question 86 and 87 was changed. First surveys whose height is around 177,5 was taken. Then from these, six surveys with question 88 having value between 55 and 60 was chosen. And averages of the values in question 86 and 87 of these six surveys were taken. Finally we filled the missing values in the 49^{th} survey by using these averages. In addition to these, for the first survey, values in 78^{th}, 79^{th} and 80^{th} were entered wrongly as 80^{th}, 79^{th} and 78^{th} respectively, and then this fault was corrected. Moreover a new variable was derived by using the values in question 8 and 9. We gave one if a person uses the specified vehicle before, otherwise we gave zero. And, hence we removed the variables 8 and 9. Later, we removed variables 17, 23, 25 and 75 because they all had some missing values since they added to survey after 30^{th} survey and there were some questions in the survey which follow similar pattern with them. In addition to these, we examined the variability in attributes and we removed variables 19, 20, 21, 35, 40, 54, 57, 59, 61 and 62 since their variability is less.

After correcting the wrong spellings, we recognized that some variables had some missing values. These are questions 11^{th}, 12^{th}, 13^{th}, 53^{rd}, 55^{th}, 56^{th}, 58^{th} and 60^{th}. The common property of these questions is they all the questions that only answered by the specified vehicle users. Then we coded the missing values with 0 for questions 11, 12, 13 and with 2 for question 53, 55, 56, 58, 60. By this way for these fields these new codes represent the people who didn't use this vehicle before.

5.3 DECISION TREE MODELING

In decision tree modeling we used five different train and test data. These data sets were selected randomly with replacement method. All training data sets include 57 records whereas test data sets include 23 records. The records in levels are quite different in our output field. For instance, there are only 3 records for level 3, 17 records for level 4, 5 records for level 5, 44 records for level 6 and 11 records for level 7. And there are no records belong to the levels 1 and 2. For this reason, we selected the records from each level randomly.

Moreover since some levels have small number of records and our data size is small we also tried developing trees by combining some output levels. We derived two more output fields. One of them includes three levels and the other includes two levels. In the derived output with three levels; levels 3 and 4 were combined and named as level 1; level 5 and 6 were combined and named as level 2 and level 7 was named as level 3. And in the other derived output with two levels; levels 3, 4 and 5 were combined and named as level 1 and the levels 6 and 7 were combined and named as level 2.

First of all, we constructed trees with the output field with seven levels in the original data. We developed trees for the five training data sets using C5, CART and CHAID algorithms and then test them with their test data. When constructing the trees we limited the minimum number of records in leaf nodes in order to prevent overfitting. We chose this value as two since there are only two records for level three in data. Otherwise we couldn't see the rules related to that level.

The obtained correct prediction rates for the train and the test datasets and the levels in the output field are shown in Table 5.2, 5.3 and 5.4. In tables there are also some other performance measures as stability, ease of use, depth of tree and computational efficiency. Here stability shows the performance of our model on both training and test datasets. It is computed by the difference between the correct prediction rates of the training and test data divided by the sum of these values i.e. $(CR_{TR} - CR_{TE}) / (CR_{TR} + CR_{TE})$. Ease of use shows the number of rules generated to reach a certain level of prediction accuracy, which reflects the tree size and complexity. Depth of tree is the number of levels in the constructed tree. The computational efficiency is the computation time that is spent by the algorithm to develop a model based on the training data set. And in below parts of the table there are some averages and standard deviations belong to the five replications/subsets.

Table 5.2 Results of the C5 algorithm for the output field with seven levels

C5		CORRECT PREDICTION RATE(%)									
		Overall	Level 3	Level 4	Level 5	Level 6	Level 7	Stability	Ease of use	Depth of tree	Computational efficiency
Replication 1	Train	92,98	100	83,33	100	93,55	100	0,51	18	9	1
	Test	30,43	0	20	0	38,46	33,33				
Replication 2	Train	84,21	100	100	50	93,55	37,5	0,23	11	6	1
	Test	52,17	0	40	0	69,23	33,33				
Replication 3	Train	84,21	0	91,67	100	83,87	87,5	0,66	12	8	1
	Test	17,39	0	60	0	7,69	0				
Replication 4	Train	84,21	0	91,67	100	87,1	75	0,42	11	9	1
	Test	34,78	0	20	0	46,15	33,33				
Replication 5	Train	78,95	0	83,33	0	93,55	75	0,34	10	6	1
	Test	39,13	0	60	0	38,46	33,33				
Avrg.	Train	84,91	40	90	70	90,32	75	0,43	12,4	7,6	1
	Test	34,78	0	40	0	40,00	26,66				
Std. Dev.	Train	5,05	54,77	6,97	44,72	4,56	23,39	0,16	3,21	1,52	0
	Test	12,68	0	20	0	22,03	14,91				

Table 5.3 Results of the CART algorithm for the output field with seven levels

CART		CORRECT PREDICTION RATE(%)									
		Overall	Level 3	Level 4	Level 5	Level 6	Level 7	Stability	Ease of use	Depth of tree	Computational efficiency
Replication 1	Train	87,72	100	83,33	75	93,55	75	0,22	11	6	5
	Test	56,52	0	40	0	76,92	33,33				
Replication 2	Train	80,7	100	75	50	93,55	50	0,30	9	5	4
	Test	43,48	0	20	0	61,54	33,33				
Replication 3	Train	94,74	100	100	75	93,55	100	0,33	13	6	5
	Test	47,83	0	20	0	69,23	33,33				
Replication 4	Train	85,96	50	100	50	93,55	62,5	0,24	11	6	5
	Test	52,17	100	0	0	76,92	33,33				
Replication 5	Train	92,98	0	91,67	100	100	87,5	0,46	11	6	5
	Test	34,78	0	20	0	53,85	0				
Avrg.	Train	88,42	70	90	70	94,84	75	0,31	11	5,8	4,8
	Test	46,96	20	20	0	67,69	26,66				
Std. Dev.	Train	5,63	44,72	10,87	20,92	2,88	19,76	0,09	1,41	0,45	0,45
	Test	8,36	44,72	14,14	0	10,03	14,91				

Table 5.4 Results of the CHAID algorithm for the output field with seven levels

CHAID		CORRECT PREDICTION RATE(%)									
		Overall	Level 3	Level 4	Level 5	Level 6	Level 7	Stability	Ease of use	Depth of tree	Computational efficiency
Replication 1	Train	82,46	100	66,67	50	93,55	75	0,41	17	6	3
	Test	34,78	0	0	0	53,85	33,33				
Replication 2	Train	87,72	0	100	50	93,55	87,5	0,34	15	7	5
	Test	43,48	0	20	0	69,23	0				
Replication 3	Train	85,96	0	100	0	100	75	0,21	15	7	3
	Test	56,52	0	80	0	61,54	33,33				
Replication 4	Train	91,23	0	100	50	100	87,5	0,40	15	7	4
	Test	39,13	0	40	0	46,15	33,33				
Replication 5	Train	85,96	100	91,67	0	96,77	75	0,48	12	5	3
	Test	30,43	0	40	0	30,77	33,33				
Avrg	Train	86,67	40	91,668	30	96,77	80	0,37	14,8	6,4	3,6
	Test	40,87	0	36	0	52,31	26,66				
Std. Dev.	Train	3,19	54,77	14,43	27,39	3,23	6,85	0,10	1,79	0,89	0,89
	Test	10,01	0,00	29,66	0	14,80	14,91				

After examining the results it is seen that any of the correct prediction rate of the trees for test datasets is in the desired level. One of the reasons for this result is our dataset is small. Another reason is there are some levels with zero cell or some levels with very small number of frequency in the dataset. Hence the results obtained from the developed trees aren't reliable. But in order to see the relationships within the data and able to compare results with the other results obtained from the two derived outputs, we put the decision tree graph of the tree which gave the best result. If we look at the averages of the different algorithms we see that CART algorithm gave better results than the other algorithms. Then we added the decision tree graph of the replication one which gave the best result for this algorithm to Appendix C (Figure C.1).

Later on, we tried to develop trees for the derived output field with three levels. We again developed trees for the five training data sets using C5, CART and CHAID algorithms and then test them with their test data. And, we constructed trees by choosing the minimum number of records in leaf nodes as two. The obtained results are shown in Table 5.5, 5.6 and 5.7.

Table 5.5 Results of the C5 algorithm for the output field with three levels

C5		CORRECT PREDICTION RATE (%)							
		Overall	Level 1	Level 2	Level 3	Stability	Ease of use	Depth of tree	Computational efficiency
Replication 1	Train	85,96	64,29	94,29	87,5	0,37	10	9	1
	Test	39,13	33,33	50	0				
Replication 2	Train	71,93	57,14	94,29	0	0,08	3	2	1
	Test	60,87	16,67	92,86	0				
Replication 3	Train	94,74	92,86	94,29	100	0,25	10	5	1
	Test	56,52	50	71,43	0				
Replication 4	Train	92,98	92,86	94,29	87,5	0,32	11	9	1
	Test	47,83	0	64,29	66,67				
Replication 5	Train	84,21	78,57	97,14	37,5	0,28	7	6	1
	Test	47,83	50	57,14	0				
Average	Train	85,96	77,144	94,86	62,5	0,26	8,2	6,2	1
	Test	50,44	30	67,144	13,334				
Standard deviation	Train	9,03	16,29	1,27	42,39	0,11	3,27	2,95	0,00
	Test	8,47	21,73	16,45	29,81573				

Table 5.6 Results of the CART algorithm for the output field with three levels

CART		CORRECT PREDICTION RATE (%)							
		Overall	Level 1	Level 2	Level 3	Stability	Ease of use	Depth of tree	Computational efficiency
Replication 1	Train	87,72	85,71	97,14	50	0,29	9	5	4
	Test	47,83	50	57,14	0				
Replication 2	Train	94,74	92,86	94,29	100	0,33	10	6	6
	Test	47,83	16,67	71,43	0				
Replication 3	Train	91,23	78,57	100	75	0,23	9	6	9
	Test	56,52	16,67	78,57	33,33				
Replication 4	Train	91,23	92,86	94,29	75	0,35	10	6	6
	Test	43,48	0	64,29	33,33				
Replication 5	Train	96,49	92,86	97,14	100	0,52	10	7	9
	Test	30,43	16,67	42,86	0				
Average	Train	92,28	88,57	96,57	80	0,35	9,6	6	6,8
	Test	45,22	20,00	62,86	13,33				
Standard deviation	Train	3,42	6,39	2,39	20,92	0,11	0,55	0,71	2,17
	Test	9,53	18,26	13,74	18,26				

Table 5.7 Results of the CHAID algorithm for the output field with three levels

CHAID		CORRECT PREDICTION RATE (%)							
		Overall	Level 1	Level 2	Level 3	Stability	Ease of use	Depth of tree	Computational efficiency
Replication 1	Train	94,74	100	97,14	75	0,37	15	7	4
	Test	43,48	66,67	35,71	33,33				
Replication 2	Train	92,98	92,86	97,14	75	0,36	11	5	5
	Test	43,48	50	50	0				
Replication 3	Train	98,25	92,86	100	100	0,43	14	6	5
	Test	39,13	16,67	50	33,33				
Replication 4	Train	91,23	100	91,42	75	0,45	10	5	5
	Test	34,78	0	57,14	0				
Replication 5	Train	96,49	100	94,29	100	0,42	12	5	4
	Test	39,13	50	35,71	33,33				
Average	Train	94,74	97,14	96,00	85	0,41	12,4	5,6	4,6
	Test	40,00	36,67	45,71	20,00				
Standard deviation	Train	2,77	3,91	3,26	13,69	0,04	2,07	0,89	0,55
	Test	3,64	27,39	9,58	18,26				

The results show us that output field with three levels also couldn’t give any desired conclusions. The reason of this is the records within the levels are again not enough to reach reliable conclusions. But in order to able to see the derived relationships within the data by using the output field with three levels, the decision tree graph of the replication three in C5 algorithm is added to Appendix C (Figure C.2).

Up to now, we couldn’t reach any reliable conclusion. And lastly, we tried to build tree for the output field with two levels. Here level one can be thought as not much satisfaction and level two as so much satisfaction. The results obtained from the constructed trees whose minimum number of records in leaf nodes is two, are shown in Table 5.8, 5.9 and 5.10.

Table 5.8 Results of the C5 algorithm for the output field with two levels

C5		CORRECT PREDICTION RATE (%)						
		Overall	Level 1	Level 2	Stability	Ease of use	Depth of tree	Computational efficiency
Replication 1	Train	92,98	83,33	97,44	0,21	6	6	1
	Test	60,87	42,86	68,75				
Replication 2	Train	96,49	94,44	97,44	0,19	9	6	1
	Test	65,22	14,29	87,5				
Replication 3	Train	92,98	88,89	94,87	0,21	7	4	1
	Test	60,87	57,14	62,5				
Replication 4	Train	94,74	94,44	94,87	0,12	8	6	1
	Test	73,91	28,57	93,75				
Replication 5	Train	96,49	94,44	97,44	0,13	7	6	1
	Test	73,91	71,43	75				
Average	Train	94,74	91,11	96,41	0,17	7,4	5,6	1
	Test	66,96	42,86	77,50				
Standard deviation	Train	1,76	4,97	1,41	0,04	1,14	0,89	0,00
	Test	6,59	22,59	12,96				

Table 5.9 Results of the CART algorithm for the output field with two levels

CART		CORRECT PREDICTION RATE(%)						
		Overall	Level 1	Level 2	Stability	Ease of use	Depth of tree	Computational efficiency
Replication 1	Train	92,98	77,78	100	0,18	6	4	5
	Test	65,22	71,43	62,5				
Replication 2	Train	92,98	83,33	97,44	0,14	5	4	3
	Test	69,57	42,86	81,25				
Replication 3	Train	96,49	94,44	97,44	0,34	7	4	3
	Test	47,83	57,14	43,75				
Replication 4	Train	92,98	83,33	97,44	0,18	6	4	3
	Test	65,22	14,29	87,5				
Replication 5	Train	94,74	88,89	97,44	0,12	7	6	4
	Test	73,91	57,14	81,25				
Average	Train	94,03	85,55	97,95	0,19	6,2	4,4	3,6
	Test	64,35	48,57	71,25				
Standard deviation	Train	1,57	6,33	1,14	0,08	0,84	0,89	0,89
	Test	9,91	21,66	18,01				

Table 5.10 Results of the CHAID algorithm for the output field with two levels

CHAID		CORRECT PREDICTION RATE(%)						
		Overall	Level 1	Level 2	Stability	Ease of use	Depth of tree	Computational efficiency
Replication 1	Train	100	100	100	0,28	11	6	4
	Test	56,52	57,14	56,25				
Replication 2	Train	96,49	100	94,87	0,16	8	6	2
	Test	69,57	28,57	87,5				
Replication 3	Train	96,49	94,44	97,44	0,26	7	4	3
	Test	56,52	42,86	62,5				
Replication 4	Train	98,25	100	97,44	0,27	8	5	2
	Test	56,52	14,29	75				
Replication 5	Train	96,49	100	94,87	0,26	8	8	3
	Test	56,52	57,14	56,25				
Average	Train	97,54	98,89	96,92	0,25	8,4	5,8	2,8
	Test	59,13	40,00	67,50				
Standard deviation	Train	1,57	2,49	2,15	0,05	1,52	1,48	0,84
	Test	5,84	18,62	13,55				

If we look at the averages of the correct prediction rates, for test data these values reached to 60%. And it took its maximum in C5 algorithm which was 67%. The decision tree graph of the replication three in C5 algorithm was seen in Appendix C (Figure C.3). Although 66% is not so high, it shows that the results obtained from the output with two levels can give us some considerable information about what affects customer satisfaction.

All analyses show that our data isn't enough to analyze customer satisfaction in detailed levels. By using this data we only draw general frontiers of customer satisfaction or no satisfaction. And we can identify influential variables on customer satisfaction.

After deciding to examine customer satisfaction in two levels, we redrew the tree using entire dataset. We used C5 algorithm because it gave us the best results in replications. And we limited the number of records in leaf nodes as four. The generated tree can be seen in Appendix C (Figure C.4). The rules derived from the three are shown below. There exist support and confidence level of the rules in parenthesis. Here, support shows us the number of records that the rule covers and confidence indicates how much percent of those records were predicted truly.

<u>*Rules for 1*</u> - contains 5 rule(s)

Rule 1 for 1.0 (4; 0,75)

♣ *if Soru33 = 1.0 and Soru43 = 1.0 and Soru13 in [0.000 1.000] and Soru22 = 1.0 and Soru7 in [1.000 2.000] and Soru27 = 0.0 and Soru80 <= 31.6 then 1.000*

Rule 2 for 1.0 (6; 0,667)

♣ *if Soru33 = 1.0 and Soru43 = 1.0 and Soru13 in [0.000 1.000] and Soru22 = 1.0 and Soru7 in [3.000] then 1.000*

Rule 3 for 1.0 (5; 1,0)

♣ *if Soru33 = 1.0 and Soru43 = 1.0 and Soru13 in [0.000 1.000] and Soru22 = 0.0 then 1.000*

Rule 4 for 1.0 (7; 0,714)

♣ *if Soru33 = 1.0 and Soru43 = 0.0 then 1.000*

Rule 5 for 1.0 (10,127; 0,79)

♣ *if Soru33 = 0.0 then 1.000*

<u>*Rules for 2*</u> - contains 3 rule(s)

Rule 1 for 2.0 (12; 1,0)

♣ *if Soru33 = 1.0 and Soru43 = 1.0 and Soru13 in [0.000 1.000] and Soru22 = 1.0 and Soru7 in [1.000 2.000] and Soru27 = 1.0 then 2.000*

Rule 2 for 2.0 (6; 1,0)

♣ *if Soru33 = 1.0 and Soru43 = 1.0 and Soru13 in [0.000 1.000] and Soru22 = 1.0 and Soru7 in [1.000 2.000] and Soru27 = 0.0 and Soru80 > 31.6 then 2.000*

Rule 3 for 2. 0 (29,873; 1,0)

♣ *if Soru33 = 1.0 and Soru43 = 1.0 and Soru13 in [2.000 3.000] then 2.000*

The C5 algorithm generated eight rules by using seven input variables. The input variables Soru33, Soru43, Soru13, Soru22, Soru7, Soru27 and Soru80 were chosen from the model as the most influential variable on customer satisfaction. The variable Soru33 shows whether a customer can easily reach the mechanism which adjusts the back of the driver seat or not. The variable Soru43 shows whether the roughness of a driver seat is appropriate or not. On the other hand, the variable Soru13 represents how frequently the customer uses that specified vehicle. In the variable Soru22, it is questioned that whether the back of the seat supports the customers' waist or not. The variable Soru7 represents the income range of a customer whereas Soru27 represents the expectation of a customer from the head of the driver seat to show a relaxing effect on head of the customer. And lastly, the variable Soru80 is the length of the customer's arm above the elbow. Five rules were extracted for level 1 whereas three rules were extracted for level 2. All rules have a sufficient support and confidence levels. We can accept Rule 5 for level 1, and rules 1 and 3 for level 2, whose both support and confidence levels are very high, as the most powerful rules generated by the model. Overall correct prediction rate of the model is 91, 25%. At the same time correct prediction rates for level 1 and 2, which are shown in Table 5.11, are also high.

Table 5.11 Coincidence Matrix for C5 model (rows show actuals)

	1.000000	2.000000
1.000000	25	0
2.000000	7	48

If we interpret the rules, the customers in the high level income group are less satisfied with driver seat. So before the new design of a driver seat, the expectations of the high level income customers can be collected. Then some extra properties might be added to the driver seat considering these expectations. Furthermore, the customers who frequently use or do not use the specified vehicle before are less satisfied with driver seat. This shows the design of the driver seat isn't suitable for the long usage. Similarly, the business can identify the problems that customers face with when using this vehicle frequently, and solve these problems in their new designs. The other information is the customers with arm length above the elbow is under 31,6cm are less satisfied. Then we can do further analysis to find the problems of that people whose this anthropometric measure is in this interval.

Moreover, gains charts, which are the visual evaluation tool which shows the performance of a specified model on predicting particular outcomes [9], are drawn. According to these charts we can say that the performance of the model is good for level 1 and 2. Because the gains achieved by the models are very close to the best line. These gain charts are illustrated in Figure 5.2 for level 1 and in Figure 5.3 for level 2.

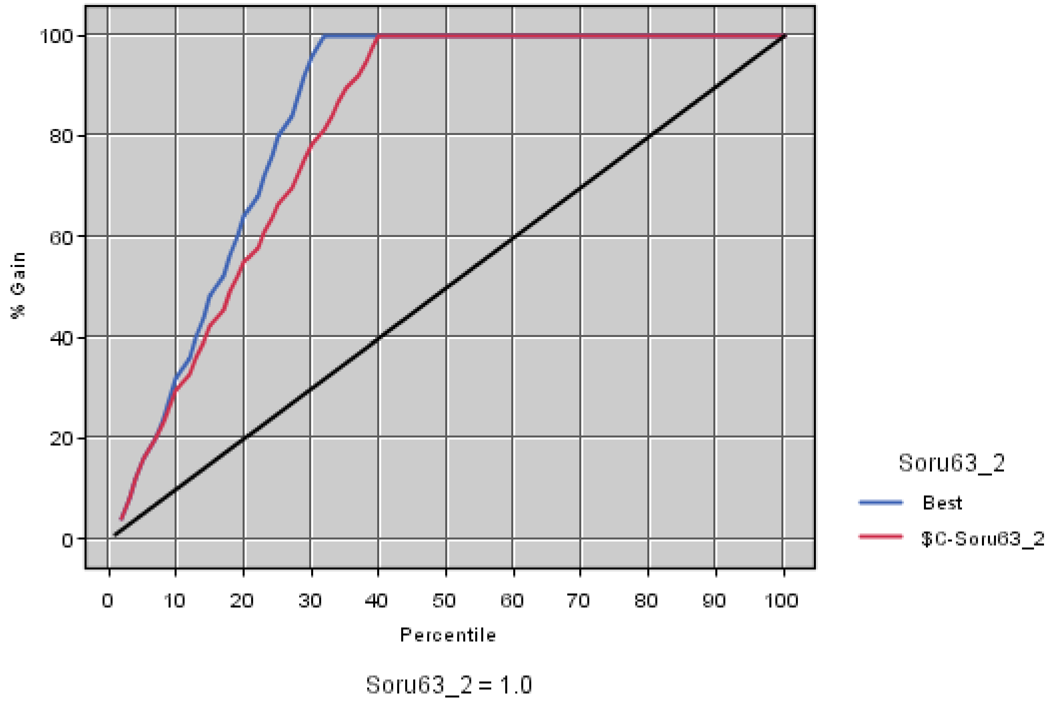


Figure 5.2 Gain chart of C5 model for the hit level ‘1’

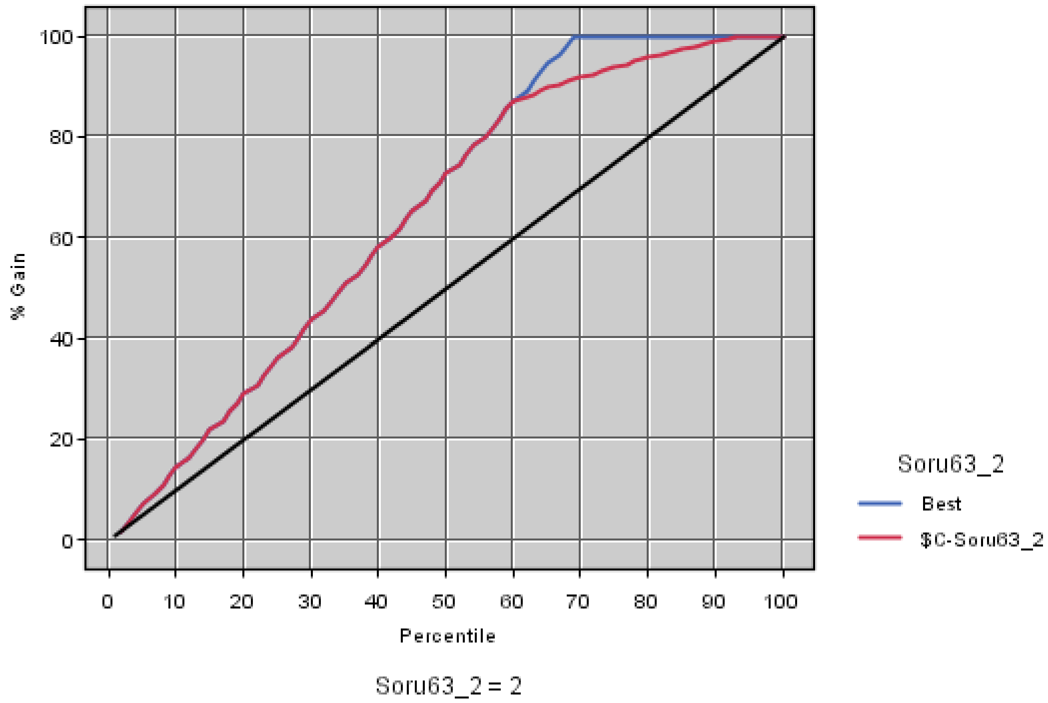


Figure 5.3 Gain chart of C5 model for the hit level ‘2’

5.4 LOGISTIC REGRESSION MODELING

The logistic regression modeling was used to check and compare the results of the decision tree modeling. This was already implemented in the ongoing study [99]. So that we use the results obtained from that study. In Table 5.12, 5.13 and 5.14, there exist some entries showing the results of the logistic regression results.

The stepwise procedure was used to develop the model. As it can be seen from Table 5.12, there are nine input variables in the model. All of them appear in the main effects. According to the Table 5.13, the overall model was found to be significant (Pearson = 0.549, Deviance = 1). And, the Pseudo R-Square statistics are also very high (Cox and Snell = 0,545, Nagelkerke = 0,777, McFadden = 0,651). From Table 5.14, we see that all the model parameters were found to be significant. The correct prediction rate of the generated model is 83,75% which is smaller than the decision tree model.

The input variables Soru43, Soru33, 13-2, Soru7, Soru42, Soru24, std85, std76 and Soru27 were selected by model as the most important variables. The variable 13-2 actually represents the variable Soru13. In decision tree algorithms the levels within the variables which are in ordered or nominal type can be combined automatically. But in logistic regression users have to combine them. Here, 13-2 is the variable which is obtained by combining some of the levels of the variable Soru13. Moreover the variables std85 and std76 show the standardized form of the variables Soru85 and Soru76.

Table 5.12 Step summary table of logistic regression modeling

Step Summary

Model		Action	Effect(s)	Model Fitting Criteria	Effect Selection Tests		
				-2 Log Likelihood	Chi-Square(a,b)	df	Sig.
Step 0	**0**	Entered	Intercept	86.92	.		
Step 1	**1**	Entered	Soru43	75.97	12.001	1	.001
Step 2	**2**	Entered	Soru33	64.15	13.672	1	.000
Step 3	**3**	Entered	13_2	56.78	6.638	1	.010
Step 4	**4**	Entered	Soru22	49.69	7.993	1	.005
Step 5	**5**	Entered	Soru7	45.32	4.284	1	.038
Step 6	**6**	Entered	Soru42	41.39	3.872	1	.049
Step 7	**7**	Entered	Soru24	38.54	2.928	1	.087
Step 8	**8**	Entered	std85	35.23	3.472	1	.062
	9	Removed	Soru22	36.67	1.447	1	.229
Step 9	**10**	Entered	std76	33.30	3.231	1	.072
Step 10	**11**	Entered	Soru27	30.30	2.884	1	.089
Stepwise Method: Forward Stepwise							
a. The chi-square for entry is based on the score test.							
b. The chi-square for removal is based on the likelihood ratio test.							

Table 5.13 Performance measures of logistic regression modeling

Model Fitting Information

Model	Model Fitting Criteria	Likelihood Ratio Tests		
	-2 Log Likelihood	Chi-Square	df	Sig.
Intercept Only	86.924			
Final	30.297	56.626	9	.000

Goodness-of-Fit

	Chi-Square	df	Sig.
Pearson	59.989	62	.549
Deviance	30.297	62	1.000

Pseudo R-Square

Cox and Snell	.545
Nagelkerke	.777
McFadden	.651

Table 5.14 Parameter estimates of logistic regression modeling

Parameter Estimates

Soru63_2(a)		B	Std. Error	Wald	df	Sig.	Exp(B)	95.0% Confidence Interval for Exp(B)	
								Lower Bound	Upper Bound
1.000000	Intercept	3.517	2.200	2.556	1	.110			
	[13_2=1.000000]	4.147	1.456	8.112	1	.004	63.215	3.644	1096.648
	[13_2=2.000000]	0(b)	.	.	0	.	.	.	.
	Soru24	-3.158	1.382	5.218	1	.022	4.25E-002	2.83E-003	.639
	[Soru27=0.000000E+000]	1.882	1.183	2.533	1	.111	6.570	.647	66.731
	[Soru27=1.000000]	0(b)	.	.	0	.	.	.	.
	Soru33	-5.079	1.831	7.696	1	.006	6.23E-003	1.72E-004	.225
	Soru42	-4.055	1.787	5.147	1	.023	1.73E-002	5.21E-004	.576
	Soru43	-6.015	2.022	8.845	1	.003	2.44E-003	4.64E-005	.129
	Soru7	2.393	1.143	4.382	1	.036	10.945	1.165	102.863
	std76	-1.046	.685	2.328	1	.127	.351	9.17E-002	1.347
	std85	1.101	.520	4.482	1	.034	3.008	1.085	8.337

a. The reference category is: 2.000000.

b. This parameter is set to zero because it is redundant.

5.5 FINDINGS AND DISCUSSION

In the case study, logistic regression and decision tree methods were implemented. If we examine the results of the two methods, we can see the similarities between them. The variables Soru43, Soru33, Soru13, Soru7 and Soru27 are commonly seen in both of the models. And the magnitudes of the effects are also same in the models. Correct prediction rate of the logistic regression model (83,75%) is smaller than the decision tree model (91,25%). 63 of the records were predicted correctly by both of the two models. Similarly 2 records were predicted wrongly by both of the two models. 4 records predicted wrongly only by decision tree model whereas 6 records predicted wrongly only by logistic regression model. 4 records couldn't be predicted in logistic regression but it was truly predicted in decision tree model. And one record couldn't be predicted in logistic regression and wrongly predicted by decision tree modeling.

Decision trees are the models which include both the main effects and the interaction effects of the variables. In small samples, they are useful for investigating the relationship within the data. And they can be used to identify the influential variables. Whereas in big samples, they are useful for both determining the important variables and prediction/classification of the quality. Because this time we can divide the dataset, then we obtain train and test data which both include sufficient number of records. Then we can confirm the accuracy and the power of our model by using test data.

On the other hand, logistic regression is a more advanced and accurate technique for small samples. It can be used for both determining the important variables and classification of the quality. When predicting the class of a new record, it gives us the probability of being of that record in the predicted class. By this way we can know the accuracy of our classification. Similarly it also gives us the significance level of the selected variables.

In conclusion, in small samples we can think decision trees as initial analyses before the logistic regression. We can use them to get some general information about the data. Especially, they can be very useful when there are too many input variables which are also the case in our study. Moreover decision trees can be helpful to combine the levels of the categorical input variables. In our decision tree model, the levels 0 and 1 within the variable Soru13 were combined automatically. If we look at the results of the logistic regression, we see that the same levels were combined in Soru13. So we can think it also as another beneficial pre-information about the data. However in big samples or when the number of the levels within the output field is high, decision trees can be more preferable than the logistic regression. The reason of this is interpretation of the decision tree models is easy whereas the logistic regression is hard and complicated.

CHAPTER 6

CONCLUSION AND FUTURE WORKS

In this thesis a comprehensive literature review of DM applications in quality improvements is presented. There is very little information about DM and its usage in a manufacturing environment in today's literature [96]. One contribution of this research is exploring that in which industries DM is used and in which quality improvement studies in those industries it is used up to now. Therefore we are able to determine the industries not used in DM studies and techniques which are not implemented up to now. And this presentation of the incomplete areas will be a guide for the future researchers.

Moreover there is an extended table in Appendix A for the people who want to look for detailed information about the articles. In this table information about the data and software used in articles and the DM processes followed in the articles can be found. On the other hand there isn't much information about the data preprocessing part of the DM process in the articles. This is because this part isn't mentioned in detail in the articles, even not mentioned in some of them. Whereas this part is the most time consuming and important part of the DM studies as mentioned in the DM sources. Hence in this thesis we emphasize that this part should have been explained in a more detailed manner.

In the second part of this thesis a case study approach is used to see how DM can be used in the customer satisfaction from the driver seat. In the case study the data which is obtained from another ongoing study [99] was gathered by questionnaire approach. We applied decision tree to this data in order to see the most important variables on the design of the driver seat. So we can think the decision tree modeling as an initial modeling. We choose this technique because it is suitable to the characteristics of our data and it is very user friendly. Also, it can quickly derive simple rules which can be interpreted easily. Because of these properties this technique is very acceptable and desirable by the firms.

In the case study C5.0 which is a popular algorithm used for classification is implemented. Then a model which has a correct prediction rate of 91,25% is generated. This model extracts meaningful rules. And the rules have high support and confidence level. At the end we identify seven input variables by using this model. These variables are the factors which should be considered by the firm while designing the seats.

After these we apply logistic regression to confirm the results obtained from the decision trees. We choose logistic regression because it is one of the traditional parametric statistical approaches. And it has been commonly used in manufacturing up to now. We take the results of the logistic regression from another ongoing study [99]. In the model constructed in the scope of that study, there are nine variables which are found to be significant. All of these variables appear in the main effects. The correct prediction rate of this model is also 83.75%. Five of the variables in the model are same with the decision tree modeling. As a consequence we can say that two models support each other.

Possible future work of this case study is modeling the significant input variables respectively. For instance in logistic regression modeling Soru33 is chosen as an important variable on customer satisfaction. This variable shows whether a customer can easily reach the mechanism which adjusts the back of the driver seat or not. Then by choosing this variable as an output variable we can describe the people who couldn't reach the mechanism which adjusts the back of the driver seat. For example after the analysis we get information as "Tall people couldn't reach the mechanism". By this way we can have a result as the firm should have a design which also satisfies the tall people. As a result, like in this example the other variables can be analyzed for further interpretations.

REFERENCES

[1] Berry, M. J. A., Linoff, G. (2000). Mastering data mining: the art and science of customer relationship management, Wiley Computer Pub., New York.

[2] Cabena, P., Hadjinian, P., Stadler, J. V., Zanasi, A. (1998). Discovering data mining: from concept to implementation, Upper Saddle River, Prentice Hall, N.J.

[3] Dunham, M. H. (2003). Data mining introductory and advanced topics, Upper Saddle River, Prentice Hall/Pearson Education, N.J.

[4] Han, J., Kanber M. (2001). Data mining: concepts and techniques, Morgan Kaufmann Publishers, San Francisco.

[5] Pyle, D. (1999). Data preparation for data mining, Morgan Kaufmann Publishers, San Francisco, Calif.

[6] Quinlan, J.R. (1993). C4.5 Programs for machine learning, Morgan Kaufmann Publishers

[7] Paolo, G. (2003). Applied Data Mining: Statistical Methods for Business and Industry, Wiley and sons, Newyork.

[8] Clementine® 10.1 Algorithms Guide (2006). USA:Integral Solutions Limited. http://www.spss.com/clementine/

[9] Clementine® 10.1 Node Reference (2006). USA:Integral Solutions Limited. http://www.spss.com/clementine/

[10] Clementine® 10.1 User's Guide (2006). USA:Integral Solutions Limited. http://www.spss.com/clementine/

[11] Data Mining Application areas (2005). Retrieved September 12, 2005, from http://www.bilgiyonetimi.org

[12] Data Mining Tasks (2007). Retrieved July 15, 2007, from http://www2.cs.uregina.ca/~dbd/cs831/notes/kdd/2_tasks.html

[13] Data Mining Tasks (2007). Retrieved July 15, 2007, from http://www.exinfm.com/pdffiles/intro_dm.pdf

[14] C4.5 Algorithm (2007). Retrieved July 15, 2007, from http://en.wikipedia.org/wiki/C4.5_algorithm

[15] Decision Tree Algorithm (2007). Retrieved August 1, 2007, from http://209.85.135.104/search?q=cache:zv0FQwsZkeIJ:www.megaputer.com/products/pa/algorithms/dt.php3+minumum+number+of+record+for+decision+tree&hl=tr&ct=clnk&cd=12&gl=tr

[16] Abajo, N., Diez, A. B., Lobato, V., Cuesta, S. R. (2004). ANN Quality Diagnostic Models for Packaging Manufacturing: An Industrial Data Mining Case Study, Proceedings of the Conference on Knowledge Discovery and Data Mining, Seattle, Washington, USA, 799-804.

[17] Ali, O. G., Chen, Y. (1999). Design Quality and Robustness with Neural Networks, IEEE Transactions on Neural Networks, 10, 6, 1518-1527.

[18] Baek, J., Kim C., Kim, S. (2002). Online Learning of the Cause-and-Effect Knowledge of a Manufacturing Process, International Journal Of Production Research, 40, 14, 3275-3290.

[19] Braha, D., Shmilovici, A. (2002). Data Mining for Improving a Cleaning Process in the Semiconductor Industry, IEEE Transactions on Semiconductor Manufacturing, 15, 1, 91-101.

[20] Brinksmeier, E., Toe Nshoff, H. K., Czenkusch C., Heinzel C. (1998). Modeling and Optimization of Grinding Processes, Journal of Intelligent Manufacturing, 9, 303-314

[21] Chang, D., Jiang, S. (2002). Assessing quality performance based on the on-line sensor measurements using neural Networks, Computers&Industrial Engineering, 42, 417-424

[22] Chen, C.R., Ramaswamy, H.S. (2002). Modeling and optimization of variable retort temperature (VRT) thermal processing using coupled neural networks and genetic algorithms, Journal of Food Engineering, 53, 209–220

[23] Chen, W., Lee, A. H.I., Deng, W., Liu, K. (2007). The implementation of neural network for semiconductor PECVD process, Expert Systems with Applications, 32 , 1148–1153

[24] Cherian, R. P., Midha, P. S., Pipe, A. G. (2000). Modeling the Relationship Between Process Parameters and Mechanical Properties Using Bayesian Neural Networks for Powder Metal Parts, International Journal of Production Research, 38, 10, 2201–2214

[25] Chiang, T., Su, C., Li, T., Huang, R. C. C. (2001). Improvement of Process Capability Through Neural Networks and Robust Design: A Case Study, Quality Engineering, 14, 313–318

[26] Chien, C., Wang, W., Cheng, J. (2006). Data mining for yield enhancement in semiconductor manufacturing and an empirical study, Expert Systems with Applications, In Press.

[27] Cook, D.F., Ragsdale, C.T., Major, R.L. (2000). Combining a neural network with a genetic algorithm for process parameter optimization, Engineering Applications of Artificial Intelligence, 13, 391-396

[28] Cool, T., Bhadeshia, H.K.D.H., MacKay, D.J.C. (1997). The yield and ultimate tensile strength of steel welds, Materials Science and Engineering, A223, 186-200

[29] Cser, L., Gulyas, J., Szücs, L., Horvath, A., Arvai, L., and Baross, B. (2001). Different Kinds of Neural Networks in Control and Monitoring of Hot Rolling Mill, Proceedings of the Conference on Industrial and Engineering Applications of Artificial Intelligence and Expert Systems , Budapest , HONGRIE, 791-796

[30] Cus, F., Balic, J. (2003). Optimization of cutting process by GA approach, Robotics and Computer Integrated Manufacturing, 19, 113–121

[31] Deng B., Liu, X. (2002). Data Mining in Quality Improvement, Proceedings of the Twenty-first Annual SAS Users Group International Conference, Orlando, Florida, 111-127

[32] Dhond, A.,, Gupta, A., Vadhavkar, S. (2000). Data Mining Techniques for Optimizing Inventories for Electronic Commerce, Proceedings of the Conference on Knowledge Discovery and Data Mining, Boston, MA USA, 480-486

[33] Erzurumlu, T., Oktem, H. (2003). Comparison of response surface model with neural network in determining the surface quality of moulded parts, Materials and Design, 28, 459–465

[34] Feng, C.X. (J.), Wang, X. (2002). Development of Empirical Models for Surface Roughness Prediction in Finish, International Journal of Advanced Manufacturing Technology, 20, 348-356

[35] Feng, C.X. (J.) and Wang, X.F. (D.) (2003). Surface Roughness Predictive Modeling: Neural Networks versus Regression, IIE Transactions on Design and Manufacturing, 11-27

[36] Gardner, M., Bieker, J. (2000). Data Mining Solves Tough Semiconductor Manufacturing Problems, Proceedings of the Conference on Knowledge Discovery and Data Mining, Boston, MA USA, 376-383

[37] Georgilakis, P., Hatziargyriou, N. (2002). On the Application of Artificial Intelligence Techniques to the Quality Improvement of Industrial Processes, I.P. Vlahavas and C.D. Spyropoulos (Eds.): SETN(2002, LNAI 2308, pp. 473–484, 2002.

[38] Guessasma, S., Salhi, Z., Montavon, G., Gougeon, P., Coddet, C. (2004). Artificial intelligence implementation in the APS process diagnostic, Materials Science and Engineering B, 110, 285–295

[39] Han, L., Han, L., Liu, C. (1999). Neural network applied to prediction of the failure stress for a pressurized cylinder containing defects, International Journal of Pressure Vessels and Piping, 76, 215–219

[40] Ho, S., Lee, K., Chen, S., Ho, S. (2002). Accurate modeling and prediction of surface roughness by computer vision in turning operations using an adaptive neurofuzzy inference system, International Journal of Machine Tools & Manufacture, 42, 1441–1446

[41] Ho, G.T.S. Lau, H.C.W., Lee, C.K.M., Ip, A.W.H., Pun, K.F. (2006). An Intelligent Production Workflow Mining System for Continual Quality Enhancement, Intelligent Journal of Advanced Manufacturing Technology, 28, 792–809

[42] Holena, M., Baerns, M. (2003). A tool for the approximation of the dependency of yield on catalyst composition, and for knowledge extraction, Catalysis Today, 81, 485–494

[43] Hou, T., Liu, W., Lin, L. et, (2003). Intelligent Remote Monitoring and Diagnosis of Manufacturing Processes Using an Integrated Approach of Neural Networks and Rough Sets, Journal of Intelligent Manufacturing, 14, 2, 239-253

[44] Hou, T., Huang, C. (2004). Application of Fuzzy Logic and Variable Precision Rough Set Approach in a Remote Monitoring Manufacturing Process for Diagnosis Rule Induction, Journal of Intelligent Manufacturing, 15, 3, 395-408

[45] Hsieh, K., Tong, L. (2001). Optimization of multiple quality responses involving qualitative and quantitative characteristics in IC manufacturing using neural networks, Computers in Industry, 46, 1-12

[46] Hu, C., Su, S. (2004). Hierarchical Clustering Methods for Semiconductor Manufacturing Data, Proceedings of the IEEE International Conference on Networking, Sensing and Control, 2, 1063 – 1068

[47] Huang, M.L., Hung, Y.H. (2006). Combining radial basis function neural network and genetic algorithm to improve HDD driver IC chip scale package assembly yield, Expert Systems with Applications

[48] Huang, H., Wu, D. (2005). Product Quality Improvement Analysis Using Data Mining : A Case Study in Ultra-Precision Manufacturing Industry, Proceedings of the Conference on Fuzzy Systems and Knowledge Discovery, Changsha , CHINE, 577-580

[49] Huang, C., Li, T., Peng, T. (2006). Attribute Selection Based on Rough Set Theory for Electromagnetic Interference (EMI) Fault Diagnosis, Quality Engineering, 18, 161–171

[50] Ilumoka, A. A. (1998). A modular neural network approach to microelectronic circuit yield optimization. Microelectronic Reliability, 38, 571-580

[51] Jemwa, G. T., Aldrich, C. (2005). Improving Process Operations Using Support Vector Machines and Decision Trees, American Institute of Chemical Engineers, 51, 2, 526–543

[52] Jiao, Y., Lei, S., Pei, Z.J., Lee, E.S. (2004). Fuzzy adaptive networks in machining process modeling: surface roughness prediction for turning operations, International Journal of Machine Tools & Manufacture, 44, 1643–1651

[53] Kang, B. S., Choe, D. H., Park, S. C. (1999). Intelligent Process Control in Manufacturing Industry With Sequential Processes, International Journal of Production Economics, 60-61, 583-590

[54] Kim, S., Lee, C. M. (1997). Nonlinear Prediction of Manufacturing Systems Through Explicit and Implicit Data Mining, Proceedings of the Conference on Computer and Industrial Engineering, 33, 461-464

[55] Kim, E., Oh, C., Lee, S., Lee, B., Yun, I., (2001). Modeling and optimization of process parameters for GaAs/AlGaAs multiple quantum well avalanche photodiodes using genetic algorithms, Microelectronics Journal, 32, 563-567

[56] Kim, I., Son, J., Yarlagadda, P. K.D.V. (2003). A study on the quality improvement of robotic GMA welding process, Robotics and Computer Integrated Manufacturing, 19, 567–572

[57] Kurtaran, H., Ozcelik, B., Erzurumlu, T. (2005). Warpage optimization of a bus ceiling lamp base using neural network model and genetic algorithm, Journal of Materials Processing Technology, 169, 314–319

[58] Kusiak, A. (2000). Decomposition in Data Mining: An Industrial Case Study, IEEE Transactions on Electronics Packaging Manufacturing, 23, 4, 345-353

[59] Kusiak, A., Kurasek, C. (2001). Data Mining of Printed-Circuit Board Defects, IEEE Transactions on Robotics and Automation, 17, 2, 191-196

[60] Krimpenis, A., Benardos, P.G., Vosniakos, G.-C., Koukouvitaki, A. (2006). Simulation-Based Selection of Optimum Pressure Die-Casting Process Parameters Using Neural Nets and Genetic Algorithms, Intelligent Journal of Advanced Manufacturing Technology, 27, 509–517

[61] Lewis, R. W., Ransing, R. S. (1997). A Semantically Constrained Bayesian Network for Manufacturing Diagnosis, International Journal of Production Research, 35, 8, 2171–2187

[62] Li, M., Feng, S., Sethi, I. K., Luciow, J., Wagner, K., (2003). Mining Production Data with Neural Network & CART, Proceedings of the Third IEEE International Conference on Data Mining, 731-734

[63] Li, T.S., Su, C.T., Chiang, T.L. (2003). Applying robust multi-response quality engineering for parameter selection using a novel neural–genetic algorithm, Computers in Industry, 50, 113–122

[64] Lian, J., Lai, X. M., Lin, Z. Q., Yao, F.S. (2002). Application of Data Mining and Process Knowledge Discovery in Sheet Metal Assembly Dimensional Variation Diagnosis, Journal of Materials Processing Technology, 129, 1, 315-320

[65] Lin, W. S., Wang, K. S. (2000). Modeling and optimization of turning processes for slender parts, International Journal of Production Research, 38, 3, 587-606

[66] Mathews, P.G., Shunmugam, M.S. (1999). Neural-network approach for predicting hole quality in reaming, International Journal of Machine Tools & Manufacture, 39, 723–730

[67] Mieno, F., Sato, T., Slubnya, Y., Odagiri, K., Tsuda, H., Take, K. (1999). Yield Improvement Using Data Mining System, Proceedings of the Conference on Semiconductor Manufacturing, Santa Clara, CA, USA, 391-394

[68] Olabi, A.G., Casalino, G., Benyounis, K.Y., Hashmi, M.S.J. (2006). An ANN and Taguchi algorithms integrated approach to the optimization of CO2 laser welding, Advances in Engineering Software, 37, 643–648

[69] Ozcelik, B., Erzurumlu, T. (2006). Comparison of the warpage optimization in the plastic injection molding using ANOVA, neural network model and genetic algorithm, Journal of Materials Processing Technology, 171, 437–445

[70] Perzyk, M., Biernacki, R., Kochanski, A. (2005). Modeling of manufacturing processes by learning systems: The naive bayesian classifier versus artificial neural networks, Journal of Materials Processing Technology, 164–165, 1430–1435

[71] Raj, K. H., Sharma, R. S., Srivastava, S., Patvardhan, C. (2000). Modeling of manufacturing processes with ANNs for intelligent manufacturing, International Journal of Machine Tools & Manufacture, 40, 851–868

[72] Rallo, R., Ferre-Gine, J., Arenas, A., Giralt, F. (2002). Neural virtual sensor for the inferential prediction of product quality from process variables, Computers and Chemical Engineering, 26, 1735-/1754

[73] Ribeiro, B. (2005). Support Vector Machines for Quality Monitoring in a Plastic Injection Molding Process, IEEE Transactıons On Systems, Man, and Cybernetıcs—Part C: Applications and Reviews, 35, 3, 401-410

[74] Sarimveis, H., Doganis, P., Alexandridis, A. (2006). A classification technique based on radial basis function neural Networks, Advances in Engineering Software, 37, 218–221

[75] Sadeghi, B.H.M. (2000). A BP-neural network predictor model for plastic injection molding process, Journal of Materials Processing Technology, 103, 411-416

[76] Shen, C., Wang, L., Li, Q. (2006). Optimization of injection molding process parameters using combination of artificial neural network and genetic algorithm method, Journal of Materials Processing Technology.

[77] Shi, D., Tsung, F. (2003). Modeling and diagnosis of feedback-controlled processes using dynamic PCA and neural networks, International Journal of Production Research, 41, 2, 365–379

[78] Shi, X., Schillings P., Boyd, D. (2004). Applying artificial neural networks and virtual experimental design to quality improvement of two industrial processes, International Journal of Production Research, 42, 1,101–118

[79] Skinner, K. R., Montgomery, D. C., Runger, G. C., Fowler, J. W., McCarville, D. R., Rhoads, T. R., Stanley, J. D. (2002). Multivariate Statistical Methods for Modeling and Analysis of Wafer Probe Test Data, IEEE Transactions on Semiconductor Manufacturing, 15, 4,.523-530

[80] Suneel, T.S., Pandle, S.S., Date, P.P. (2002). A technical note on integrated product quality model using artificial neural Networks, Journal of Materials Processing Technology, 121, 77-86

[81] Tam, C.M., Tong, T. K. L., Lau, T. C. T., Chan, K.K. (2004). Diagnosis of prestressedconcrete pile defects using probabilistic neural Networks, Engineering Structures, 26, 1155–1162

[82] Tan, S.C., Lim, C.P., Rao, M.V.C. (2007). A hybrid neural network model for rule generation and its application to process fault detection and diagnosis, Engineering Applications of Artificial Intelligence, 20, 203–213

[83] Tay, K.M., Butler, C. (1997). Modeling and Optimizing of a MIG Welding Process—A Case Study Using Experimental Designs and Neural Networks, Quality And Reliability Engineering International, 13, 61–70

[84] Tsai, Y., Chen, J. C., Lou, S., (1999). An in-process surface recognition system based on neural networks in end milling cutting operations, International Journal of Machine Tools & Manufacture, 39, 583–605

[85] Tseng, T. (B.), Kwon, Y., Ertekin, Y. M. (2005). Feature-Based Rule Induction in Machining Operation Using Rough Set Theory for Quality Assurance, Robotics and Computer-Integrated Manufacturing, 21, 559–567

[86] Tsuda, H., Shirai, H., Takagi, O., Take, R., (2000). Yield Analysis and Improvement by Reducing Manufacturing Fluctuation Noise, Proceedings of International Symposium on Semiconductor Manufacturing, 249–252

[87] Vasudevan, M., Murugananth, M., Bhaduri, A.K. (2002). Application of Bayesian Neural Network for Modeling and Prediction of Ferrite Number in Austenitic Stainless steel Welds, Mathematical Modeling of Weld Phenomena, 6, The Institute of Materials, London, 1079–1099

[88] Vasudevan, M., Rao, B.P.C., Venkatraman, B., Jayakumar, T., Raj, B. (2005). Artificial neural network modeling for evaluating austenitic stainless steel and Zircaloy-2 welds, Journal of Materials Processing Technology, 169, 396–400

[89] Wang, R., Wang, L., Zhao, L., Liu, Z. (2006). Influence of Process Parameters on Part Shrinkage in SLS, Intelligent Journal of Advanced Manufacturing Technology

[90] Yang, T., Tsai, T. (2002). Modeling and implementation of a neurofuzzy system for surface mount assembly defect prediction and control, IIE Transactions, 34, 637–646

[91] Yang, T., Tsai, T., Yeh, J. (2005). A neural network-based prediction model for fine pitch stencil printing quality in surface mount assembly, Engineering Applications of Artificial Intelligence,18, 335–341

[92] Yin, X., Yu, W. (2006). The virtual manufacturing model of the worsted yarn based on artificial neural networks and grey theory, Applied Mathematics and Computation

[93] Zhai, L., Khoo, L., Fok, S. (2002, Feature Extraction Using Rough Set Theory and Genetic Algorithms—An Application for the Simplification of Product Quality Evaluation, Computers and Industrial Engineering, 43, 661–676

[94] Zhou, Q., Xiong, Z., Zhang, J., Xu, Y. (2006). Hierarchical Neural Network Based Product Quality Prediction of Industrial Ethylene Pyrolysis Process, Lecture Notes in Computer Science, 3973, 1132-1137

[95] Zuperl, U., Cus, F., (2003). Optimization of cutting conditions during cutting by using neural Networks, Robotics and Computer Integrated Manufacturing, 19, 189–199

[96] Schertel, S. L. (2002). Data Mining and its Potential Use in Textiles:A Spinning Mill. (Under the direction of Dr. George Hodge and Dr. William Oxenham)

[97] Forrest, D. R. (2003). High dimensional data mining in complex manufacturing processes. Faculty of the School of Engineering and Applied Science University of Virgina.

[98] Shah, S. C. (2005). Mining Noisy Data: A Prediction Quality Perspective, Graduate College of The University of Iowa.

[99] Cabuk, V. (2007). A comprehensive customer requirement analysis for a driver seat using logistics regression and optimization, M.Sc. Thesis, METU, IE, Ankara, to appear.

[100] Phadke, M. S. (1989), Quality Engineering Using Robust Design, Englewood Cliffs, NJ: Prentice-Hall

[101] Montgomery, D. C., Peck E. A., Vining G. G. (2001), Introduction to Linear Regression Analysis

[102] Regression analysis. (2007) Retrieved August 15, 2007, from http://en.wikipedia.org/wiki/Regression_analysis

[103] Analysis of variance. (2007) Retrieved August 15, 2007, from http://www.westburnpublishers.com/marketing-dictionary/a/analysis-of-variance-(anova).aspx

[104] Analysis of variance. (2007) Retrieved August 15, 2007, from http://www.chem.agilent.com/cag/bsp/products/gsgx/Downloads/pdf/two-way_anova.pdf

[105] Analysis of variance. (2007) Retrieved August 15, 2007, from http://en.wikipedia.org/wiki/Analysis_of_variance#Assumptions

[106] Köksal, G., Testik, M. C. (2007) Quality Problem Definitions and the Scope of QIwDM Project, METU, IE, Ankara

[107] Köksal, G. (2007) Data Mining Applications in Quality Improvement: A Tutorial and a Literature Review, EURO XXII, Prague

[108] Sharkey, A. J. K. (1996) On combining Artificial Neural Nets Department of Computer Science,. University of Sheeld, U.K..

[109] Braha, D. (2002) Data Mining for Design and Manufacturing: Methods and Applications, Massive Computing, 3

APPENDIX A

Table A.1 Literature review table

Researcher	***Aim of research***	***Product or process***	***Data source (number of records) (train+test+verification)***	***Input: Type (number of inputs)***	***Output: Type (number of outputs)***	***Software***	***Quality task***	***Data collection***	***Data preprocessing and DM task***	***DM tool***	***Result***
Abajo et al. 2004	Development of a new tinplate quality diagnostic models that provide the estimated quality of the final product	Packaging manufacturing	Online Observational (2600)(50%+25%+25%)	Cont.(21)	Cont.(2)		Process and product quality description Classification of quality		Discretization Dimension reduction Classification	GA DT (C4.5 & OC1) ANN(BNN and MLP)	S S S
Ali and Chen 1999	Presenting concise and accurate neural-network models for multiple quality characteristics in injection molding and finally modeling five critical to quality variables (CTQ's) simultaneously with high accuracy	Injection molding process	Experimental (1323)	Cont.(7)	Cont.(5)		Process and product quality description Predicting quality		Dimension reduction Prediction	MLR DT(C4.5&CART) MLR NLR ANN	S S S S S(more successful)
Beak 2002	Providing online quality control with the incremental cause	Aluminium coating process (used to	Online Observational(3000)	Cont.(6)	Nominal (1)		Classification of quality		Classification	DT (Statistical batch-based DT learning)	S(more successful)

Table A.1 (cont'd)

	and effect knowledge discovery using a novel DT learning method.	make TFT-Thin Film Transistor-arrays) a core part of (500) a TFL-LCD computer monitor								DT(ID5R)	S
Braha and Shmilovici 2002	Improving the wafer cleaning processes by identifying the significant attributes involved in the cleaning process and predicting how much a new item cleaned with the given conditions	Semiconductor manufacturing Advanced wafer dry cleaning process	Experimental	Cont.(11)	Cont.(2)		Classification of quality		Discretization Classification	Competitive ANN (SOM) DT (C4.5) ANN(BP) CC	 S S S(more successful)
Brinksmeier et al.1998	Modeling and optimization of grinding process	Cylindrical external grinding process	Experimental (48)	Cont.(2)	Cont.(4)		Predicting quality Parameter optimization	DOE(Full factorial)	Prediction Optimization	NLR ANN(BP) in combination with FST GA	S S S

Table A.1 (cont'd)

Chang and Jiang 2002	Developing a neural network model to probe the dependence between the quality of finished product and sensor measurements which were collected to monitor the failure of a tool in the manufacturing process	Cutting process	Online Observational(80)	Cont.(15)	Cont.(3)		Process and product quality description Predicting quality		Data compression Prediction	PCA MLR ANN	 S S (more successful)
Chen and Ramaswamy 2002	Developing prediction models and search for optimal variable retort temperature processing conditions for conduction heated foods	Canned foods Food processing	Simulated (250)	Cont.(10)	Cont.(4)	Neuro Shell Predictor (Wands System Group, Frederick MD 21703)	Predicting quality Parameter optimization	DOE(Full factorial)	Prediction Optimization	ANN GA	S S
Chen et al. 2007	Establishing a quality predictor for analyzing the relationship between manufacturing process parameter setting and final product quality	Semiconductor manufacturing Plasma-enhanced chemical vapor deposition	Simulated (650) (500+150)	Cont.(6)	Cont./ Binary(2)		Predicting quality Classificatio		Normalization Prediction Classification	 ANN(BP, NN parameters are selected using TM) ANN(BP,	 S S

Table A.1 (cont'd.)

	(TM was used for selecting network parameters like learning rate etc.)	(PECVD) process					n of quality			NN parameters are selected using TM)	
Cherian et al. 2000	Presenting a neural network based system for modelling mechanical behaviour of powder metal parts as a function of processing conditions	Powder metal parts manufacturing	Simulated(209)	Cont.(4)	Cont.(3)	CASIP Com.Aid Select Of Iron Powder for Exp. Data	 Predicting quality		Normalization Prediction	 ANN(BNN)	 S
Chiang et al. 2001	Optimizing multiple quality characteristics of a polymerization process	Silicon-filler manufacturing	Observational (to train and test NN) (32) Experimental (27*27)	Cont.(11)	Cont.(16)		Process/product quality description Predicting quality Parameter optimization		Dimension reduction Prediction Optimization	GLM (ANOVA) ANN ANN	S S S
Chien et al.2006	Developing a framework for mining production data to extract knowledge for manufacturing process monitoring and defect diagnosis in order to remove	Semiconductor manufacturing	Observational(71)	Nominal (168)	Cont.(1)		Process/product quality description Predicting quality		Clustering Prediction	Partitional methods (K-means) Nonparametric ANOVA (Kruskal-Wallis Test) DT (CHAID)	S S S

Table A.1 (cont'd.)

	assignable causes and thus improving the yield										
Cook et al. 2000	Modeling and optimizing parameters of a particle board manufacturing process	Particle board manufacturing process	Observational online(182) (127+55)	Cont.(26)	Cont.(3)	Neural Works Predict software package from NeuralWare/an easy-do-use add-in for Microsoft Excel	Predicting quality Parameter optimization		Prediction Optimization	ANN(RBF NN) GA	S S	
Cool et al.1997	Using an artificial neural network to empirically model and interpret the dependence of the yield and ultimate tensile strength of steel weld deposits as a function of many variables	Welding process	Experimental (770) Experimental (520)	Cont.(19) Cont.(21)	Cont.(1) Cont.(1)		Predicting quality		Missing value handling(with mean or zero) Normalization Prediction	ANN(BNN)	S	
Cser et al. 2001	Controling and monitoring of hot rolling mill	Hot rolling	Online Observational(16000)	Cont.(140)			Process/product quality description Parameter optimization		Clustering Optimization	Competitive ANN (SOM) ANN	S S	

Table A.1 (cont'd.)

Cus and Balic 2003	Determination of cutting parameters in machining operations to minimize production cost, time and quality problems.	Metal cutting processes	Experimental (20)	Cont.(3)	Cont.(3)		Parameter optimization		Optimization	GA	S
Deng and Liu 2002	Improving the quality of steel plate by adjusting the influential factors	Iron and Steel industry Steel plate production	Observational(1580) (70%+10%+20%)	Cont.(7)	Cont.(2)	SAS/EM	Predicting quality		Prediction	MLR ANN	S S(more successful)
Dhond et al. 2000	Highlights the use of NN-based DM techniques for forecasting hot metal temperature in a steel mill blast furnace	Iron and Steel industry	Observational Online(760)	Cont.(35) Cont.(11) Cont.(38) Cont.(49)	Cont.(1) Cont.(1) Cont.(1) Cont.(1)		Process/product quality description Predicting quality		Dimension reduction Prediction	ANN MLR ANN	S S S
Erzurumlu and Oktem 2007	Developing a model to predict surface roughness value error on mold surfaces	Milling process to make mold parts	Experimental (243) (236+7)	Cont.(5)	Cont.(1)	MATLAB for RSM	Predicting quality	DOE(3 level full factorial)	Prediction	RSM ANN	S S(slightly better than RS model)
Feng and Wang 2002	Devoloping an emprical model for surface roughness in finish turning	Metal processing industry	Experimental (64) (48+16)	Cont.(6)	Cont.(1)	MINITAB	Predicting quality	DOE (fractional factorial)	Prediction	NLR	S

Table A.1 (cont'd.)

Feng and Wang 2003	Devoloping an emprical model for surface roughness prediction	Metal processing industry Metal cutting process	Experimental (80) (64+16)	Cont.(5)	Cont.(1)	MINITAB	Predicting quality	DOE(2^5)	Prediction	NLR ANN	S S
Gardner and Bieker 2000	Identifying the critical poor yield factors from normally collected wafer manufacturing data	Semiconductor wafer manufacturing	Observational (17246)	Cont./Nominal (133)			Process/product quality description		Clustering	Competitive ANN (SOM) Rule Induction	S S
Georgilakis and Hatziargyriou 2002	Better understanding of different settings of process parameters and predict more accurately the effect of different parameters on the final product quality, quality improvement by increasing the classification success rate of transformer iron losses	Transformer manufacturing industry	Online Observational(768) (75%+25%) Online Observational(2595)(66,67% training)	Cont./Nominal (8) Cont.(9)	Cont.(1) Cont.(1)		Process/product quality description Classification of quality		Discretization Dimension reduction Classification	DT DT EN ANN (MLP) HDTNNC (Hybrid DT and ANN Classifier)	S S (fastest) S S (best classification but slowest) S(optimal time and accuracy)
Guessa	Presenting the	Termal	Experimental	Cont.(8)	Cont.(3)		Predicting		Prediction	ANN	S

Table A.1 (cont'd.)

sma et al. 2004	detailed procedure of the ANN implementation in the atmospheric plasma spray process and the control specifications dealing with the in-flight particle characteristics	spray process	(16*11)				quality				
Han et al. 1999	Predicting the failure stress of a cylinder with penetrating flaws under internal pressure and analyzing the sensitivity of parameters of a pressurized cylinder with defects	Pressurized cylinder manufacturing	Experimental (80) (70+10)	Cont.(5)	Cont.(1)	C language for ANN	Process/product quality description Predicting quality		Dimension reduction Prediction	ANN ANN	S S
Ho et al. 2006	Proposing an intelligent production workflow mining system (IPWMS) embracing online analytical processing (OLAP) and data mining	Slider manufacturing (DLC diamond coating like carbon process)	Online Observational(135)	Cont.(27)	Cont.(9)	NN software Qnet for Windows (proprietary neural network simulator)	Predicting quality		Filter out incomplete data (OLAP) Normalization Prediction	ANN	S

Table A.1 (cont'd.)

	technology, together with the use of artificial intelligence combining artificial neural networks (ANNs) and fuzzy rule sets to realize knowledge discovery and decision support in high-quality manufacturing										
Ho et al. 2002	Proposing a new method using an adaptive neuro-fuzzy inference system (ANFIS) to accurately establish the relationship between the features of surface image and the actual surface roughness, and consequently can effectively predict surface roughness using cutting parameters (cutting speed, feed rate,	Turning process	Experimental (73) (57+16)	Cont.(4)	Cont.(1)		Predicting quality		Prediction	ANN(Polynomial Network) FNN ANFIS (Adaptive neuro fuzzy infrasystem)	S S S

Table A.1 (cont'd.)

	and depth of cut) and gray level of the surface image										
Holena and Baerns. 2003	Dependency of propene yield on catalyst composition was approximated by the NN and logical rules were extracted with a prescribed consequent	Oxidative dehydrogenation of propane to propene	Experimental (226) (216+10)	Cont.(8)	Cont.(1)	MATLAB NN Toolbox	Predicting quality Parameter optimization		Prediction Optimization	ANN(Levenberg-Marquardt method) Sequantial quadratic programming method	S S
Hou and Huang 2004	Mining the casual relationship rules from the database of a remote monitoring and diagnosing manufacturing processes	Conveyor Belts Manufacturing	Online Observational (27)	Cont.(8)	Nominal (1)		Classification of quality		Discretization Classification	Rule induction (RST) Integration of FST and RST	S(sensitive to noisy data) S(more successful)
Hou et al. 2003	Monitoring and diagnosing manufacturing processes	Conveyor Belts Manufacturing	Online Observational (42) (27+15)	Cont.(8)	Nominal (1)		Classification of quality		Normalization Classification	ANN(BP) Rule Induction (RST)	S S
Hsieh and Tong 2001	Simultaneously optimize multiple responses including both	Semiconductor manufacturing	Experimental (36)	Ordinal/ Nominal (6)	Ordinal/ Cont.(2)		Parameter optimization		Optimization	ANN(BP)	S

Table A.1 (cont'd.)

	qualitative and quantitative quality characteristics	Ion implantation process									
Hu and Su 2004	Discussing the possible relationship between machines of the manufacturing process and the yield rates of wafers.	Semiconductor manufacturing	Observational (126 lots)	Binary (304) machines			Process/product quality description		Clustering	Hierarchical methods(Agglomerative)	S
Huang and Hung 2006	Optimizing the lower warpage properties for 0.65 mm. CSP assembly using a model based on RBFN-GA.	Chip scale package (CSP) manufacturing processes of micro hard disk drive (HDD) driver IC	Experimental (108) (80%+20%)	Cont.(9)	Cont.(1)	MATLAB used for RBFN simulations	Predicting quality Parameter optimization		Transformation Normalization Prediction Optimization	TM ANN(RBF NN) GA	S S S (RBF NN-GA is more successful)
Huang and Wu 2005	Analysing product quality improvement in ultra-precision manufacturing industry using data mining for	Ultra-Precision Manufacturing	Observational (11320)	Nominal/Binary(4)	Binary(1)		Classification of quality		Classification	DT (CHAID)	S

Table A.1 (cont'd.)

	developing quality improvement strategies										
Huang et al. 2006	Proposing a systematic approach, including data preprocessing, data discretization, data reduction, and rule generation, for selecting a group of attributes capable of representing the quality of motherboard assembly	Motherboard assembly	Online Observational(415)	Cont./ Nominal (11)	Nominal (1)	Rosetta 2005	 Process/product quality description Classification of quality		Discretization Dimension reduction Classification	 Rule induction (RST) Rule induction (RST)	 S S
Ilumoka 1998	Robust design of VLSI circuits	VLSI design	Simulated (500)	Cont.(26)	Cont.(3)		Predicting quality Parameter optimization		Prediction Optimization	ANN (Modular ANN, BP) RSM	S S
Jemwa and Aldrich 2005	Developing an online methodology for management of product and/or process quality	First-order reaction occurring in a continuous stirred tank reactor (CSTR)	Simulated (1000) (75%+25%)	Cont.(2)	Cont./Nominal(2)	MATLAB SVM Toolbox	Classification of quality		Classification	DT SVM	S S

Table A.1 (cont'd.)

	Comparing the online performance of the support vector classification (SVC) based and the DNN based systems	CSTR	Simulated (23000)	Cont.(2)	Cont./Nominal(2)					SVM DNN based systems	S (SVM is more rapid)
	To illustrate the use of the methodology in practice	An industrial manganese extraction plant	Observational	Cont.(2)	Cont./Nominal(2)					DT SVM	S S
Jiao et al. 2004	Modeling the relationships between surface roughness and cutting parameters in turning operations	Silicon wafers manufacturing	Experimental (186) (162+24)	Cont.(3)	Cont.(1)		Predicting quality	DOE(3^3)	Prediction	MLR FAN	U S(more successful)
Kang *et al* 1999	To develop a framework of intelligent process control system (in sequential manufacturing processes with automatic facilities) for the purpose of controlling and generating better	Semiconductor Manufacturing Color-CR Manufacturing	Online Observational(130) Online	Cont.(160) Cont.(31)	Cont.(1) Nominal (1)		Process and product quality description		Discretization Missing data handling Outlier handling Dimension reduction	Competitive ANN (SOM) ANN(BP)	S

Table A.1 (cont'd.)

	operating manufacturing conditions						Classification of quality		Classification	DT(C4.5)	S
Kim and Lee 1997	Comparing study of explicit and implicit methods to predict and control of general manufacturing process including nonlinear chaotic behavior	Plastic optic fibers manufacturing	Simulated	Cont.(3)	Cont.(4)		Predicting quality		Prediction	MLR Nonparametric TSA(Moving Average) Parametric TSA (Exponential smoothing) ANN CBR	S S S(more successful) S S(more successful)
Kim et al. 2001	Presenting a modeling and parameter optimization technique for GaAs/AlGaAs multiple quantum well(MQW) avalanche photodiodes(APDs)used for the image capture mechanism in a high-definition system	GaAs/AlGaAs MQW APD manufacturing	Experimental (31) (24+7) (The optimization results from GAs were verified by the simulated data)	Cont.(4)	Cont.(2)	MATLAB software package used for NN modeling ATLAS device simulation program used for APD simulation	Predicting quality Parameter optimization	DOE(D-optimal design)	Prediction Optimization	ANN (Feedforward propagation and BP) GA	S S

Table A.1 (cont'd.)

Kim et al. 2003	Presenting a new algorithm to establish a mathematical model for predicting top-bead width through a neural network and multiple regression methods, to understand relationships between process parameters and top-bead width	Robotic gas metal arc (GMA) welding process	Experimental (33) (18+9+6)	Cont.(3)	Cont.(1)	MATLAB for ANN SAS for R	Predicting quality	DOE(Fractional factorial)	Prediction	MLR NLR ANN	S S S(more successful)
Krimpenis et al. 2006	Modeling the effect of die-casting process parameters on process output using simulation runs, carefully selected using DOE and hence obtaining the optimal process conditions	Pressure die-casting process	Simulated Experimental (16) (12+4)	Cont.(4)	Binary/Cont.(2)	PRO-CAST used for finite element simulation MATLAB NN Toolbox	Process/product quality description Predicting quality Classification of quality Parameter optimization	DOE(orthogonal array Taguchi's design)	Dimension reduction Prediction Classification Optimization	GLM (ANOVA) ANN(Feedforward) ANN(LVQ) GA(multi response)	S S S S
Kurtaran et al. 2005	Determining optimum values of process parameters in injection molding of a bus	Plastic injection molding process	Simulated (243)	Cont.(5)	Cont.(1)	MATLAB NN Toolbox	Predicting quality Parameter optimization	DOE(3^3)	Prediction Optimization	ANN GA	S S

Table A.1 (cont'd.)

	ceiling lamp base to achieve minimum warpage										
Kusiak 2000	Improving the quality of wafers Eliminating the waste at the down stream production stages, where the actual integrated circuits are produced	Semiconductor Industry Integrated Circuit Manufacturing	Observational (40000)	Nominal/Cont.(15)	Cont.(2)		Process/product quality description Classification of quality		Discretization Dimension reduction Classification	Rule Induction (RST) Rule Induction (RST)	S S
Kusiak and Kurasek 2001	Solving a quality engineering problem in electronics assembly	Semiconductor Industry Printed Circuit Board Manufacturing	Observational (4104) (2052+2052)	Binary/Nominal/Cont.(14)	Binary(1)		Classification of quality		Discretization Classification	Rule Induction (RST)	S
Lewis and Ransing1997	Proposing a new network architecture which should overcome many of the disadvantages of the existing manufacturing diagnostic tools.	Pressure die casting process	Observational	Cont./ Binary(14)	(43)		Process/product quality description			ANN (Bayesian Network)	S
Li et al. 2003	Determine parameter settings in a manufacturing process with	Silicon compound manufacturing process	Observational(500) (400+100)	Cont.(14)	Cont(7)		Predicting quality Parameter		Prediction Optimization	ANN(BP) GA (multi	S S

Table A.1 (cont'd.)

	multiple responses						optimization			response)	
Li et al.2003	Improving the glass coating process by giving guidance on adjustments to the machine settings, resulting in shorter setup time and better glass coating quality	Glass manufacturing	Online Observational	Cont.(74)	Cont.(1)				Data Cleaning(ignore the tupple) Smoothing Normalization		
							Process and product quality description		Data compression	PCA	S
							Predicting quality		Prediction	DT(CART) ANN	S S (ANN slightly better than CART)
							Process/product quality description		Dimension reduction	Neural network sensitivity analysis	S
Lian et al.2002	Improving the quality of vehicles by controlling the dimensional deviation of body-in-white	Assembly of sheet metal products		Cont. /Binary	Nominal (1)		Process/product quality description		Dimension reduction Clustering Data compression	CA Maximal tree method(MT) PCA	S S S

Table A.1 (cont'd.)

							Classification of quality		Classification	DT	S
Lin and Wang 2000	An abductive network is adopted in order to construct a prediction model for surface roughness and error-of roundness in the turning operation of slender parts	Turning process	Experimental (97) (81+16)	Cont.(4)	Cont.(2)		Predicting quality Parameter optimization	DOE (3level full factorial)	Prediction Optimization	MLR ANN(Abductive Network) SA optimization	S S(more successful) S
Mathews and Shunmugam 1999	Developing a new approach for the condition monitoring in reaming.	Reaming process	Experimental (80%+100%)	Cont. (more than 8)	Cont.(3)		Predicting quality		Prediction	ANN (BP)	S
Mieno et al. 1999	Specifying the failure cause and improving yield	Semiconductor manufacturing	Observational (58)	Nominal	Real (1)		Predicting quality		Prediction	DT	S
Olabi et al. 2006	Optimization of CO_2 laser welding process.	Welding processes	Experimental (14)	Cont. (3)	Cont.(2)		Predicting quality Parameter optimization		Prediction Optimization	ANN TM	S S
Özçelik and Erzurumlu 2006	Minimizing warpage of thin shell plastic parts	Plastic Injection molding processes	Simulated(81)	Nominal/Cont.(7)	Cont.(1)	For Taguchi exp. Design Minitab	Process/product quality description Predicting	DOE(orthogonal array Taguchi's design)	Dimension reduction Prediction	GLM (ANOVA) ANN	S S

Table A.1 (cont'd.)

						2002 For Optimization MATLAB 6.5	quality Parameter optimization		Optimization	GA	S
Perzyk et al. 2005	Comparing modeling capabilities of two types of learning systems: the naive Bayesian classifier (NBC) and artificial neural networks (ANNs), based on their prediction errors and relative importance factors of input signals	Ductile cast iron Steel casting process	Observational(790) (700+90) Observational(172) Simulated (1200) (1000+200) Simulated (1200) (1000+200) Simulated (1200) Simulated (172)	 (12) Cont.(5) Cont.(12) Cont.(12) Cont.(12)	Cont.(1) Binary (1) Binary/ Cont.(4) Cont.(1) Binary (1) Binary (1)	Programmed in the VBA for Excel	 Predicting quality Classification of quality		Normalization Discretization Prediction Classification	 ANN(BP and SA) Naive Bayesian classifier	 S S(in some applications it is better than ANN)

Table A.1 (cont'd.)

Raj et al. 2000	Utilizing the function approximation capabilities of ANNs in the modeling of different manufacturing processes	Hot upsetting process Hot extrusion process Metal cutting process	Simulated (331) (320+11) Simulated (195) (180+15) Experimental (12)	Cont.(4) Cont. (4) Cont. (3)	Cont. (1) Cont. (1) Cont. (2)		Predicting quality		Prediction	ANN(BP) FEM	S(more successful) S
Rallo et al. 2002	Inferential prediction of product quality from process variables	Low-density polyethylene (LDPE) process	Online Observational (3143) (3923) (740) (1057) (1545) (4395) (14803)	Cont.(25)	Cont.(1)		Process/product quality description Predicting quality		Normalization Clustering Prediction	Competitive ANN (SOM) GLM ANN(Fuzzy ARTMAP NN) ANN(Dynamic RBF NN)	S S S S(All methods are better than LM)
Ribeiro 2005	SVMs are applied within the framework of an industrial problem for fault detection and diagnosis in an	Automotive industry Plastic injection molding	Observational(200) (120+80)	Cont. (6)	Nominal (1)		Classification of quality		Classification	ANN(RBF NN) SVM	S S(SVMs slightly better than

Table A.1 (cont'd.)

	injection molding process	process									RBF NNs)
Sadeghi 2000	Developing a neural network model for predicting the quality or soundness of the injected plastic parts based on key process variables and material grade variations	Plastic injection molding process	Simulated (2000)	Cont.(4)	Cont./Binary(3)	CAE for simulation	Predicting quality Classification of quality		Prediction Classification	ANN ANN	S S
Sarimveis et al. 2006	Proposing a new classification method for classifying the product quality in real time	Paper manufacturing Issue making process	Observational Online(258) (150+108)	Cont.(5)	Binary (1)		Classification of quality		Classification	FST(combined with RBF NN) ANN(Feed forward propagation) ANN(RBF NN)	S S S(more successful)
Shen et al. 2006	Selecting the optimal control variables in injection molding under certain given constraints to obtain best part quality	Plastic injection molding process	Simulated (252) (163+89)	Cont.(5)	Cont.(1)	The program for the process opt. of injection molding developed using MATLAB	Predicting quality Parameter optimization	DOE(Taguchi's method)	Prediction Optimization	ANN GA	S S

Table A.1 (cont'd.)

Shi and Tsung 2003	Presenting a new integrated scheme to diagnose the root cause of faults in a feedback-controlled processes by using dynamic PCA and neural networks		Simulated	Cont. (16500)			Process/product quality description Classification of quality		Data compression Classification	Dynamic PCA ANN	S S
Shi et al. 2004	Achieving a better understanding of process behavior and improving the process quality of two complex manufacturing processes	Chemical manufacturing process Printed Circuit Board Manufacturing	Experimental (37) Experimental (outputs were obtained from above established ANN)(729) Experimental (32) Experimental (outputs were obtained from above established ANN)(432)	Cont.(6) Cont.(6) Cont.(5) Cont.(5)	Cont. (2) Cont.(2) Cont.(3) Cont.(3)	C Language for ANN SAS	Predicting quality Process/product quality description	DOE(2^{6-1}) DOE(3^6) DOE(4^2*2^3)	Normalization Prediction Dimension reduction	ANN(MLP, BP) GLM (ANOVA) RSM	S S S
Skinner et al. 2002	Determining the quality or yield of the wafers and the cause of low yield	Semiconductor manufacturing	Observational (1122)	Cont.(23)	Cont.(1)	SAS Minitab	Process/product quality description		Data compression Clustering	PCA Hierarchical	U U

Table A.1 (cont'd.)

	wafers					SAS Proc GENMOD	Predicting quality		Prediction	clustering(Agglomerative) MLR GLM DT(CART)	 U U S
Suneel et al. 2002	Proposes a NN based model which predict form dimial errors and the surface finish on parts produced during CNC turning process	CNC turning process	Experimental (100) (50+50)	Cont.(4)	Cont. (3)	NEURAL program written in Practical Neural Network Recipes in C++	Predicting quality		Normalization Prediction	 ANN(MLP and BP)	 S
Tam et al. 2004	Presenting probabilistic neural network architecture for diagnosing the causes of prestressed concrete pile damages	Pre-stressed concrete piles	Observational (240)	Binary (18)	Binary (12)		Classification of quality		Classification	Probabilistic Neural Network (PNN)	S

Table A.1 (cont'd.)

Tan et al. 2007	Extracting meaningful and useful rules from the hybrid ANN model for undertaking fault detection and diagnosis (FDD) problems	Circulating water (CW) system in a power generation plant	Observational Online (2500) (1000+1000+500)	Cont.(12)	Nominal (1)		Classification of quality		Classification	ANN(Fuzzy ARTMAP (FAM)-Rectangular basis function network (RecBFN))	S
Tay and Butler 1997	Modeling and optimizing a metal inert gas (MIG) welding process.	Metal inert gas (MIG) welding process.	Experimental (17) (80%+20%+5)	Cont.(5)	Cont.(3)		Predicting quality Parameter optimization		Prediction Optimization	ANN(Gaussian RBF NN) ANN	S(quite time consuming) S(quite time consuming)
Tsai et al. 1999	Developing an in-process based surface recognition system that is capable of predicting the surface roughness of end milling on aluminum type materials	End milling cutting process	Experimental (492) (450+42)	Cont.(4)	Cont.(1)	SPSS for R C language for ANN	Predicting quality	DOE	Normalization Transformation Prediction	MLR ANN(BP)	S S(more successful)
Tseng et al.2005	Predicting acceptance level of surface roughness	Machine parts operations Machining process	Online Observational(1000) (63,2%+36,8%)	Nominal/Cont.(8)	Cont.(1)	Developed with C++ the Common Gateway Interface (CGI)	Process/product quality description		Discretization Dimension reduction	Discriminant Analysis MLR	S(more successful)

Table A.1 (cont'd.)

						Minitab Statistica	Classification of quality		Classification	Rule induction (RST)	S
Tsuda et al. 2000	Developing a method to clarify the correlation between yield and various wafer parametrical data value by reducing the influence of manufacturing fluctuation	Semiconductor manufacturing	Observational (1000)	Nominal/Cont.	Cont.(1)		Predicting quality		Prediction	DT(CART)	S
Vasudevan et al. 2002	Presenting work to reveal the influence of compositional variations on ferrite content for the austenitic stainless steel base compositions and to study the significance of individual elements on ferrite content in austenitic stainless steel welds	Welding process	Experimental (924)	Cont.(13)	Cont.(1)		Predicting quality		Normalization Prediction	ANN(BNN)	S
Vasudevan et	Demonstrating the use of ANN for	Welding process	Online Observationa	Cont.(13)	Cont.(1)	A software package in	Predicting quality		Prediction	ANN(BNN)	S

Table A.1 (cont'd.)

al. 2005	assessing the quality of welds in terms of weld microstructure, mechanical properties and weld defects		l(1283) (1020+263)	(12) (6) (6)	Binary (1) Binary (4) Cont.(2)	corporating the standart MLP based ANN alg. was developed in Visual Basic 5.0	Classification of quality		Classification	ANN	S
Wang et al. 2006	Investigating the relationship between the shrinkage and the process parameters of SLS a rapid prototyping system to improve dimensional accuracy of its products	Selective laser sintering process	Experimental (33) (27+6)	Binary/ Cont.(7)	Cont.(1)		Predicting quality	One factor at a time experiments	Prediction	ANN	S
Yang and Tsai 2002	Proposing a neurofuzzy system for surface mount assembly defect prediction and control	Surface Mount Technology (SMT) assembly process in electronics industry	Hybrid data from both Experimental and Observational/Online (81+1603) (1684+30)	Cont.(8)	Binary (4)	Fuzzy TECH (Anon, 98)	Classification of quality	DOE(3^{8-4})	Classification	Neurofuzzy system	S

Table A.1 (cont'd.)

Yang et al. 2005	Proposed a NN based quality control system for the solder stencil printing process	Solder paste stencil printing process	Experimental (243) (80%+20%)	Cont.(8)	Cont.(2)	Commercial package used Neural-Works professional II/Plus (2000)	Predicting quality	DOE(3^{k-p})	Prediction	ANN	S
Yin and Yu 2006	The ANN is combined with grey superior analysis to select important variables	Worsted spinning process	Experimental (77 lots) (69+8)	Cont.(18)	Cont. (4)	MATLAB 6.5 for GSA	Process/product quality description Predicting quality		Normalization Dimension reduction Prediction	SE MLR GSA ANN	S S S(more successful) S
Zhai et al.2002	Describing a prototype future extraction system for simplifying the product quality evaluation	Electronic Devices Manufacturing	Observational(170)	Cont.(12)	Binary (1)		Process/product quality description Classification of quality		Discretization Dimension reduction Classification	Rule induction (RST) GA	S S
Zhou et al.2006	Proposing a two-layer hierarchical neural network to predict the product qualities of an industrial KTI GK-	Petrochemical industry Ethylene pyrolysis process	Observational(7036) (5300+1736)	Cont.(26)	Cont.(2)				Outlier handling(deleted) bad samples deleted; data		

Table A.1 (cont'd.)

	V ethylene pyrolysis process						Process/product quality description Predicting quality		filtered Normalization Dimension reduction Prediction	CA ANN (BANN) (Feedforward propagationNN,trained by Leverberg-Marquart algorithm)	S S
Zuperl and Cus 2003	Optimization of cutting parameters	Turning process	Experimental (40)	Cont.(3)	Cont.(3)		Parameter optimization		Normalization Optimization	ANN (feedforward, RBF NN)	S(feedforward more accurate but more time consuming)

APPENDIX B

SAMPLE OF THE QUESTIONAIRE

BÖLÜM I

1. **soru1:** Cinsiyetiniz?

Erkek ☐ **(1)** Kadın ☐ **(2)**

2. **soru2:** Yaşınız?

18-20	☐	**(1)**
21-25	☐	(2)
26-30	☐	**(3)**
31-35	☐	**(4)**
36-40	☐	**(5)**
41-45	☐	(6)
46-50	☐	**(7)**
51-55	☐	(8)
55 ve üzeri	☐	(9)

3. **soru7:** Ailenizin/sizin toplam aylık geliriniz?

1500 YTL'nin altı	☐	**(1)**
1500-3500 YTL arasında	☐	**(2)**
3500 YTL'nin üstü	☐	**(3)**

BÖLÜM II

1. **soru13:**Aracı hangi sıklıkla kullanacaksınız / kullanıyorsunuz?

Tüm gün boyunca sürekli ☐ **(1)**

Gün içerisinde ihtiyacım oldukça ☐ **(2)**

Nadir olarak (haftasonu ve tatillerde)☐ **(3)**

4. **soru14:** Aynı segmentte / benzeri bir araç kullandınız mı?

Evet ☐ **(1)** Markası /Markları:

Hayır ☐ **(0)**

BÖLÜM IV

1. **soru22:**Koltuk arkalığı belinizi yeterince destekliyor mu?

Evet ☐ **(1)** Hayır ☐**(0)**

2. Koltuk başlığından beklentileriniz nelerdir?

Soru26:Çarpma anında başı koruması☐ **→ işaretli ise (1) değil ise (0)**

Soru27: Başı dinlendirme ☐ **→ işaretli ise (1) değil ise (0)**

Diğer:........

3. Koltuk ayar mekanizmalarına rahatlıkla erişebiliyor musunuz?

Evet ☐

Hayır ☐

Hangi ayarlar?

Soru31:İleri-geri ayarı ☐ **→ işaretli ise (0) değil ise (1)**

Soru32:Yüksekli ayarı ☐ **→ işaretli ise (0) değil ise (1)**

Soru33: Koltuk arkası ayarı ☐ **→ işaretli ise (0) değil ise (1)**

Soru34: Bel desteği ayarı ☐ **→ işaretli ise (0) değil ise (1)**

Soru35: Başlık ayarı ☐ **→ işaretli ise (0) değil ise (1)**

4. **soru43:**Koltuk uygun sertlikte mi?

Evet ☐ **(1)** Hayır ☐ **(2)** Neden?

5. **soru63:**Koltuk hakkındaki genel düşünceniz nedir?

Çok Kötü **(1)**	Kötü **(2)**	Biraz Kötü **(3)**	Normal **(4)**	Biraz iyi **(5)**	İyi **(6)**	Çok İyi **(7)**

23. Koltukla ilgili iyileştirme önerileriniz nelerdir?

BÖLÜM IV

	Vücut Ölçüleri	**Açıklama**	**Elde edilen ölçümler**
Soru76:	ağırlık		
Soru77:	boy		
Soru80:	üst kol uzunluğu	kol dirsekten 90 derece büküldüğü zaman omuzun dış ucundan dirseğe kadar olan mesafe	
Soru81:	el uzunluğu	parmaklar düzgün olarak uzatıldığında bilekten orta parmak ucuna kadar olan mesafe	

APPENDIX C

DECISION TREES

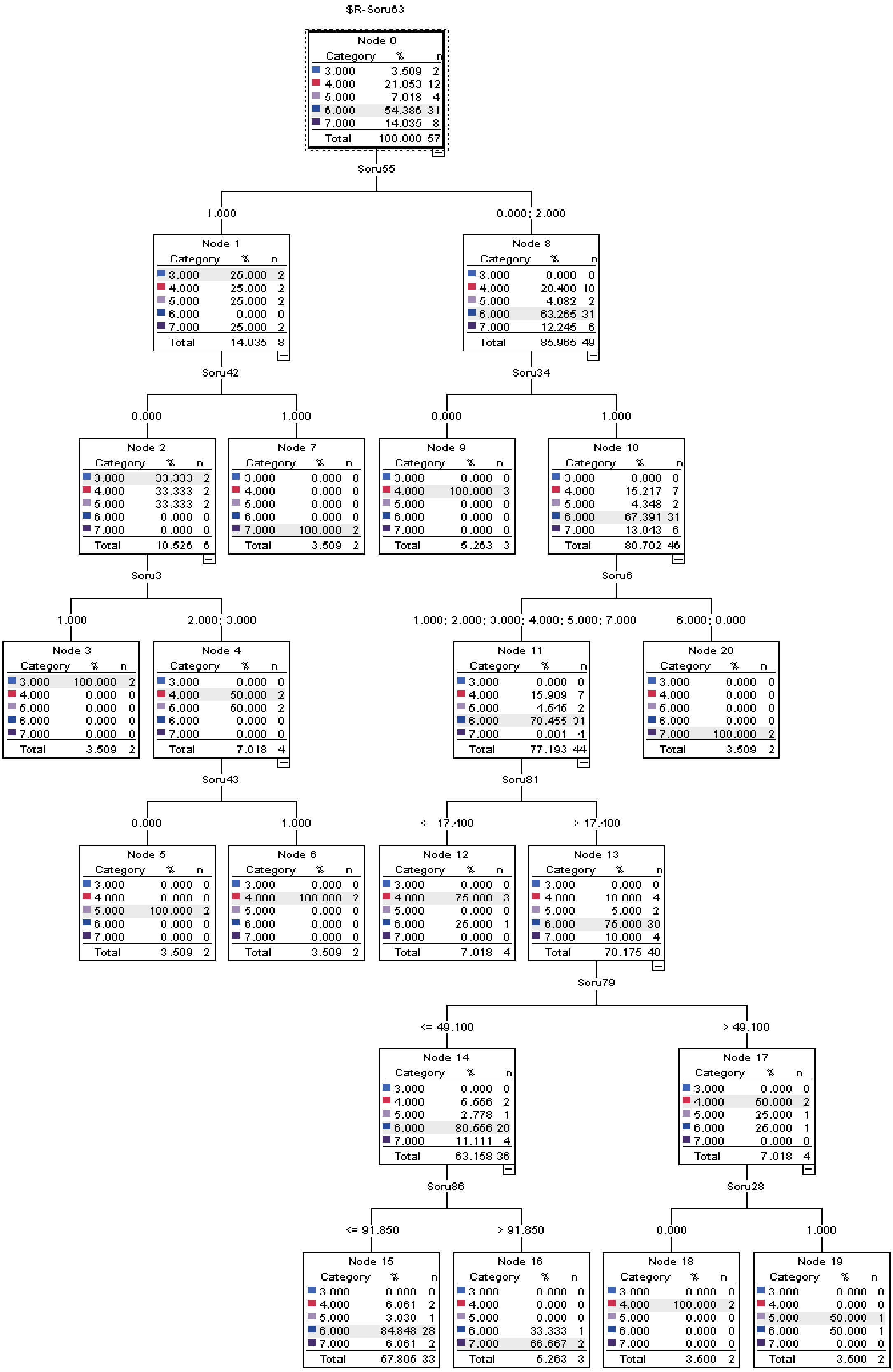


Figure C. 1 Decision tree of CART algorithm for output field with seven levels

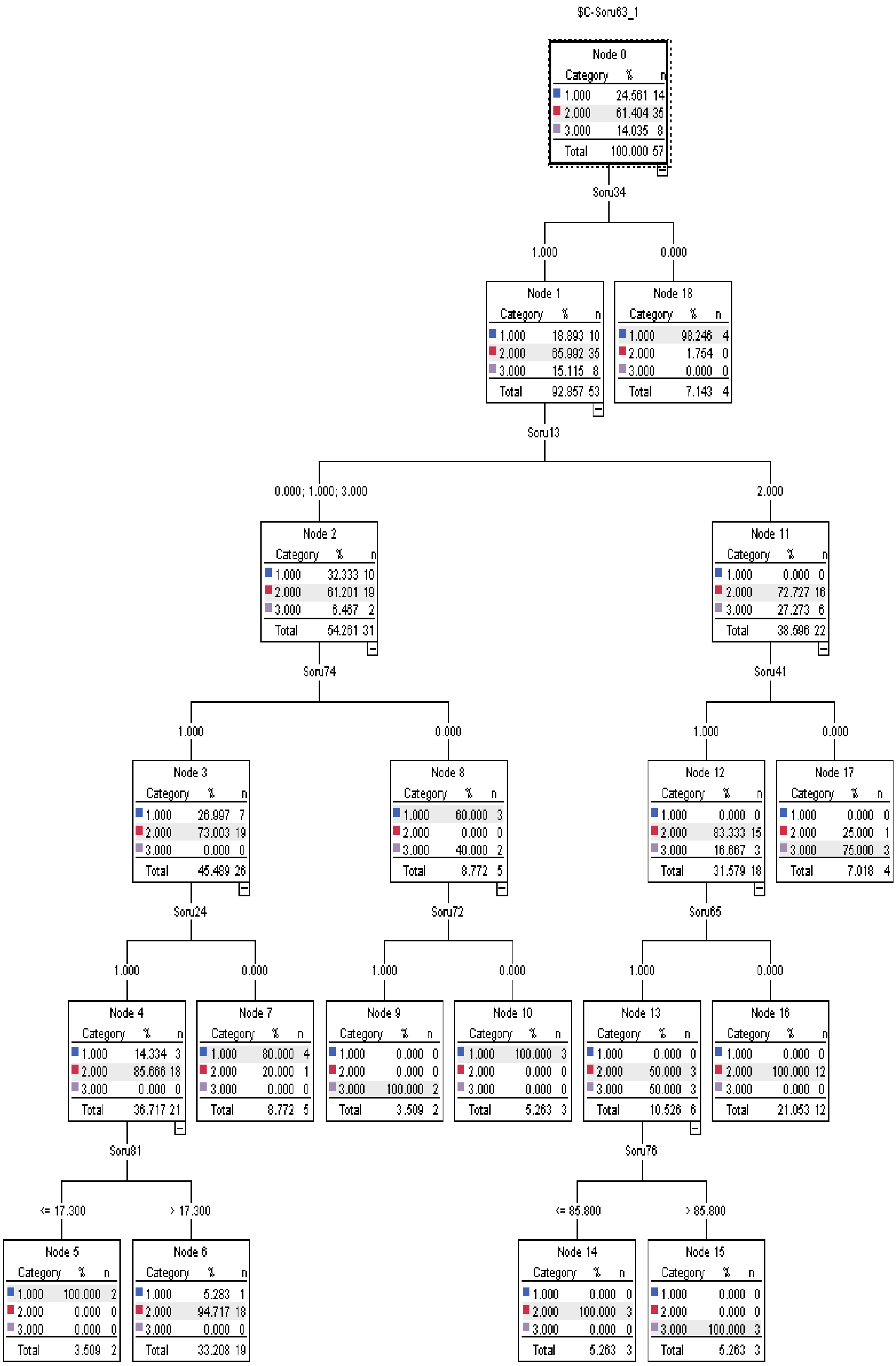


Figure C.2 Decision tree of C5 algorithm for output field with three levels

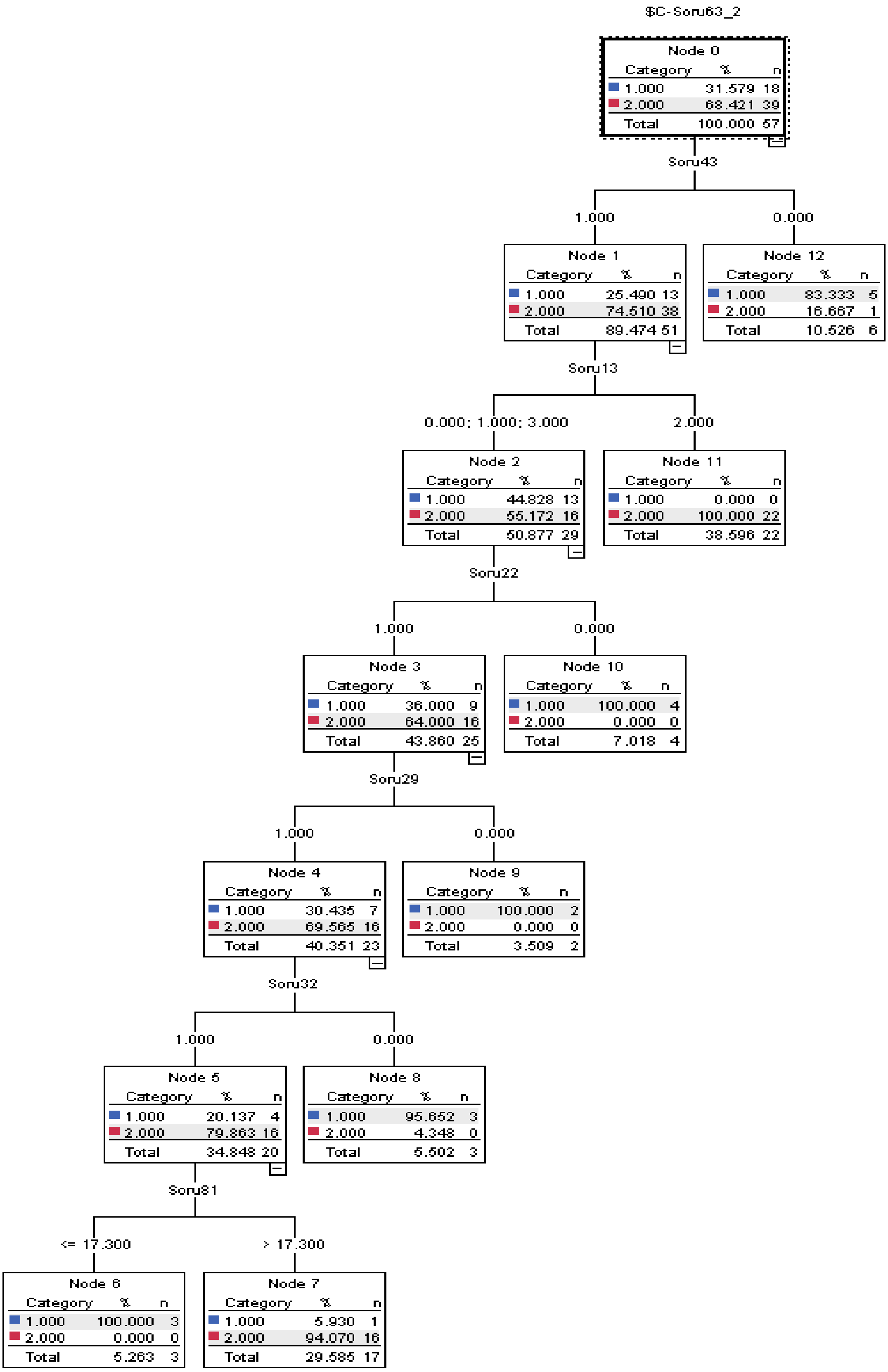


Figure C.3 Decision tree of C5 algorithm for output field with two levels

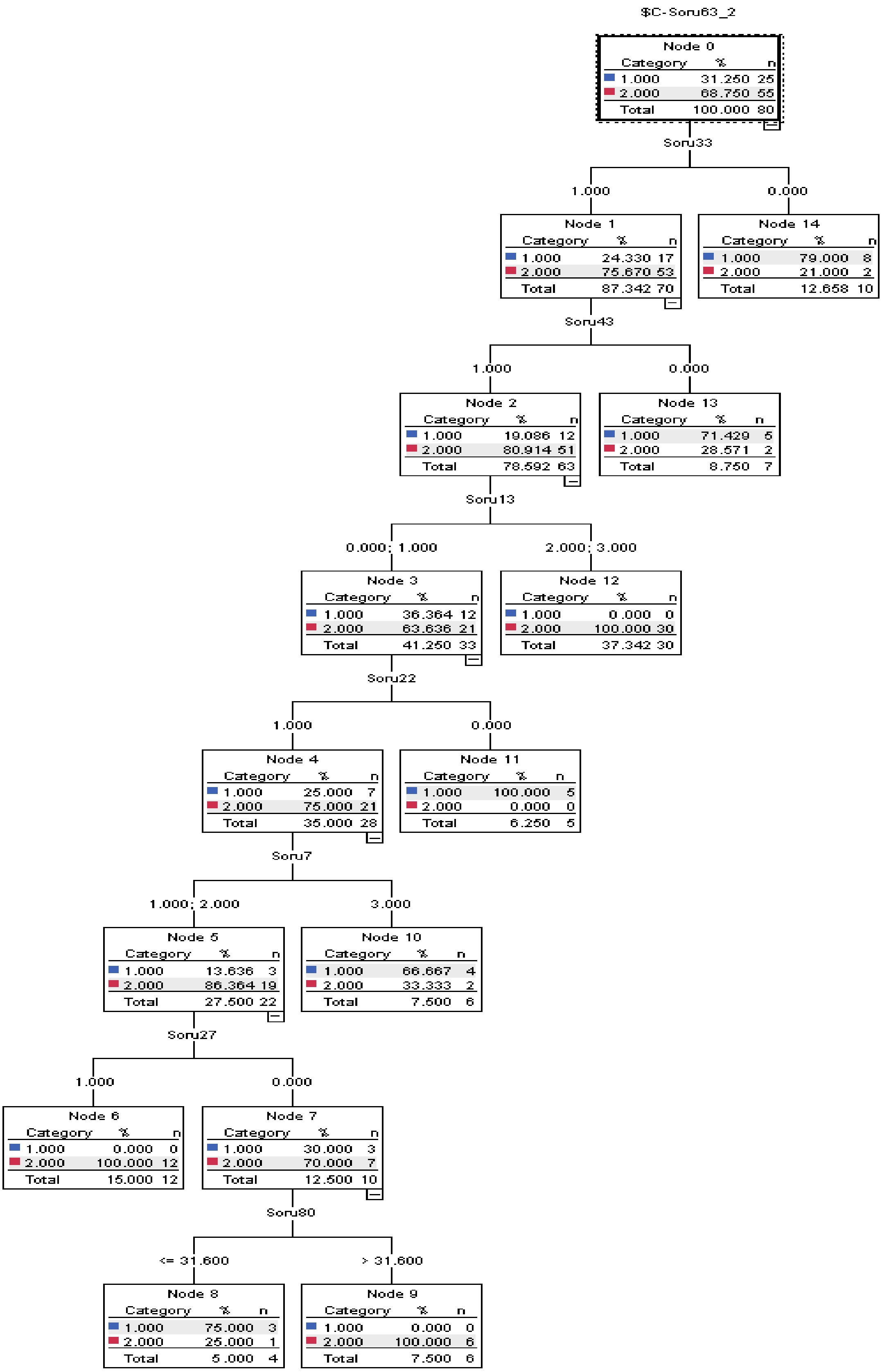


Figure C.4 Decision tree of C5 algorithm using entire dataset