Kalite İyileştirmede Veri Madenciliği Kullanımı ve Geliştirilmesi

Proje No: 105M138

Prof.Dr. Gülser KÖKSAL
Doç.Dr. İnci BATMAZ
Prof.Dr. Bülent KARASÖZEN
Prof. Dr. Sinan KAYALIGİL
Doç.Dr. Murat Caner TESTİK
Prof. Dr. Nur Evin ÖZDEMİREL
Prof. Dr. Gerhard Wilhelm WEBER
Berna BAKIR
Başak ÖZTÜRK

EYLÜL 2009
ANKARA

Önsöz

Sanayi kuruluşlarında ürün ve süreçlerin kalitesini iyileştirmeye yönelik veri madenciliği yaklaşımlarını belirlemek ve daha etkili yaklaşımlar geliştirmek için gerçekleştirilen bu proje, TÜBİTAK ve ODTÜ Bilimsel Araştırma Projeleri kaynaklarından desteklenmiştir. Proje çalışmaları için gerekli veri ERKUNT Döküm, TOFAŞ Otomotiv, VESTEL Elektronik ve DİMES Meyve Suyu kuruluşlarından elde edilmiştir. Ayrıca, AKSA Akrilik ve PETKİM Petrokimya kuruluşlarında incelemeler yapılmış ve SPAC Altı Sigma Danışmanlık firmasının deneyimleri hakkında bilgi alınmıştır. Projedeki veri analizi ve yöntem geliştirme çalışmalarına aşağıda isimleri verilen kişiler katkıda bulunmuştur.

Dr. Güvenç Aslan, Başkent Üniversitesi
Prof.Dr. Adil Bagirov, University of Ballarat
Doç. Dr. Regina Burachik, University of South Australia
Doç.Dr. Esra Karasakal, ODTÜ
Prof. Dr. Rafael Kasımbeyli, İzmir Ekonomi Üniversitesi
Dr. Süreyya Özöğür, Sabancı Üniversitesi
Dr. Pakize Taylan, Dicle Üniversitesi
Dr. Özlem Türker Bayrak, Çankaya Üniversitesi
Prof.Dr. Burhan Türkşen, TOBB ETÜ
Doç. Dr. Zeev Volkovich, Ort Braude College
Zeynep Anaklı, ODTÜ
Deniz Atınç, ODTÜ
Ezgi Avcı, ODTÜ
Dilber Ayhan, ODTÜ
Selcan Cansız, ODTÜ
Vuslat Çabuk, ODTÜ
Fatma Güntürkün, ODTÜ
İlker Arif İpekçi, ODTÜ
Masood Jabarnejad, ODTÜ
Elçin Kartal, ODTÜ
Tuna Kılıç, Çankaya Üniversitesi
Gizem Özer, ODTÜ
Barış Yenidünya, ODTÜ
Fatma Yerlikaya Özkurt, ODTÜ

Bununla birlikte, proje gelişme raporlarını değerlendiren hakemlerin önerileri yol gösterici olmuştur. Önemli destek ve katkılarından dolayı bu kurum, kuruluş ve kişilere teşekkür ederiz.

Gülser Köksal
Proje yürütücüsü

İçindekiler

Tablo Listesi

Şekil Listesi

Özet

Bu projede amaç, sanayi kuruluşlarında ürün ve süreçlerin kalitesini iyileştirmeye yönelik veri madenciliği (VM) yaklaşımlarını belirlemek ve daha etkili yaklaşımlar geliştirmektir.

Projede imalat sanayi kuruluşlarının ürün ve süreçlerinin kalitesini iyileştirme ile ilgili kalitenin tanımlanması, tahmin edilmesi, sınıflandırılması ve parametrelerinin optimizasyonu problemleri ele alınmıştır. Bu problemlerin çözümü için veri hazırlama ve önişlemenin yanısıra kümeleme, tahmin etme, sınıflandırma, birliktelik analizi ve optimizasyon VM işlevlerinin gerekli olabileceği belirlenmiştir. Bu kapsam dahilinde geniş bir literatür taraması yapılmış ve değişik imalat sektörlerinde etkinlik gösteren altı kuruluş ziyaret edilmiştir. Bunlardan üçünün sağladığı veriler üzerinde uygun VM metotları uygulanmış ve sonuçlar karşılaştırılmıştır. Bu karşılaştırma sonucunda belli VM işlevleri için kalite iyileştirme amaçlarına en uygun VM metotları belirlenmiş ve uygulayıcılara önerilmiştir.

Projenin yöntem geliştirme kısmında ise uygulama aşamasında karşılaşılan bazı problemlerin giderilmesi ve mevcut yöntemlerin kullanım kolaylığı ve/veya etkililiğinin artırılması yönünde çalışmalar gerçekleştirilmiştir. Sonuçta, kalite verilerinin yeniden örneklenmesi için bir yöntem; parametrik olmayan alternatif bir regresyon yaklaşımı (CMARS); ikili sınıflandırmada kullanımı kolay olan Mahalanobis Taguchi Sistemi metodunun çok sınıf ve ayrıca parametre optimizasyonu için uyarlamalar; bulanık sınıflandırmada kalite verilerine uygun alternatif yaklaşımlar (bulanık regresyona dayalı modeller) ve parametrik olmayan bulanık tahmin etme ve sınıflandırma fonksiyonları; parametre optimizasyonunda çekicilik fonksiyonlarının optimizasyonu için alternatif yaklaşımlar ve birliktelik kurallarının seçimi için bir yöntem geliştirilmiştir.

Bu sonuçların ve metotların kalite iyileştirme alanında uygulayıcıların çalışmalarına yön vermesi ve bunların kullanım kolaylığı ile etkililiğini artırması beklenmektedir.

***Anahtar Kelimeler*:** Kalite iyileştirme, kalite kontrol, veri madenciliği, kümeleme, tahmin etme, regresyon, sınıflandırma, parametre optimizasyonu, birliktelik analizi.

Abstract

The objective of this project is to identify the data mining (DM) approaches that can effectively improve product and process quality in industrial organizations, and to develop more effective approaches.

In the project, quality definition, prediction, classification and parameter optimization problems associated with product and process quality improvement in manufacturing industries are considered. For the solution of these problems, clustering, prediction, classification, association and optimization functions of DM as well as data preparation and preprocessing are determined as relevant. A comprehensive literature survey has been performed and six manufacturing companies operating in different sectors have been visited, within this context. Appropriate DM methods are applied on data sets obtained from three of these companies, and the results are compared. As a result, the most appropriate DM methods are suggested for specific DM functions and quality improvement purposes.

In the method development part of the project, studies are performed to overcome some problems encountered during the applications, and to increase ease of use and effectiveness of the VM methods. As a result, a resampling method for quality data; an alternative nonparametric approach (CMARS) for regression; adaptations of an easy to use binary classification method, Mahalanobis Taguchi system, to multiple classes and also to parameter optimization; alternative approaches for fuzzy classification of quality data (models based on fuzzy regression) and nonparametric fuzzy functions; alternative approaches for optimization of desirability functions in parameter optimization; and a method for reduction of association rules are developed.

It is expected that these results and approaches guide practitioners in quality improvement area, and incease the ease of use and effectiveness of them.

Keywords: Quality improvement, quality control, data mining, clustering, prediction, classification, parameter optimization, association analysis.

1. Giriş

Bu projede amaç, sanayi kuruluşlarında ürün ve süreçlerin kalitesini iyileştirmeye yönelik veri madenciliği (VM) yaklaşımlarını belirlemek ve daha etkili yaklaşımlar geliştirmektir. Buna yönelik olarak proje iki aşamada gerçekleştirilmiştir. Birinci aşamada, proje kapsamına giren kalite iyileştirme problemleri ile bunları çözmede kullanılabilecek VM işlevleri tanımlanmış ve mevcut VM metotları sanayi kuruluşlarından elde edilen bazı verilere uygulanmıştır. Böylece mevcut metotlardan hangilerinin kalite problemlerini çözmek için daha uygun olduğunu ve bunlar ile ilgili varsa iyileştirme gereksinimlerinin neler olduğunu belirlemek hedeflenmiştir. Projenin ikinci aşamasında ise belirlenen bu iyileştirme gereksinimleri doğrultusunda metotlar üzerinde iyileştirmeler gerçekleştirilmiş veya yeni metotlar geliştirilmiştir.

Projede imalat sanayi kuruluşlarının ürün ve süreçlerinin kalitesini iyileştirme ile ilgili kalitenin tanımlanması, tahmin edilmesi, sınıflandırılması ve parametrelerinin optimizasyonu problemleri ele alınmıştır. Bu problemlerin çözümü için veri hazırlama ve önişlemenin yanısıra kümeleme, tahmin etme, sınıflandırma, birliktelik analizi ve optimizasyon VM işlevlerinin gerekli olabileceği belirlenmiştir. Bu kapsam dahilinde 1997-2007 yıllarına ait literatür kapsamlı bir şekilde taranmıştır. Ayrıca değişik imalat sektörlerinde etkinlik gösteren altı kuruluş ziyaret edilmiştir. Bunlardan üçünün sağladığı veriler üzerinde uygun VM metotları uygulanmış ve sonuçlar karşılaştırılmıştır. Bu karşılaştırma sonucunda belli VM işlevleri için kalite iyileştirme amaçlarına en uygun VM metotları belirlenmiş ve uygulayıcılara önerilmiştir.

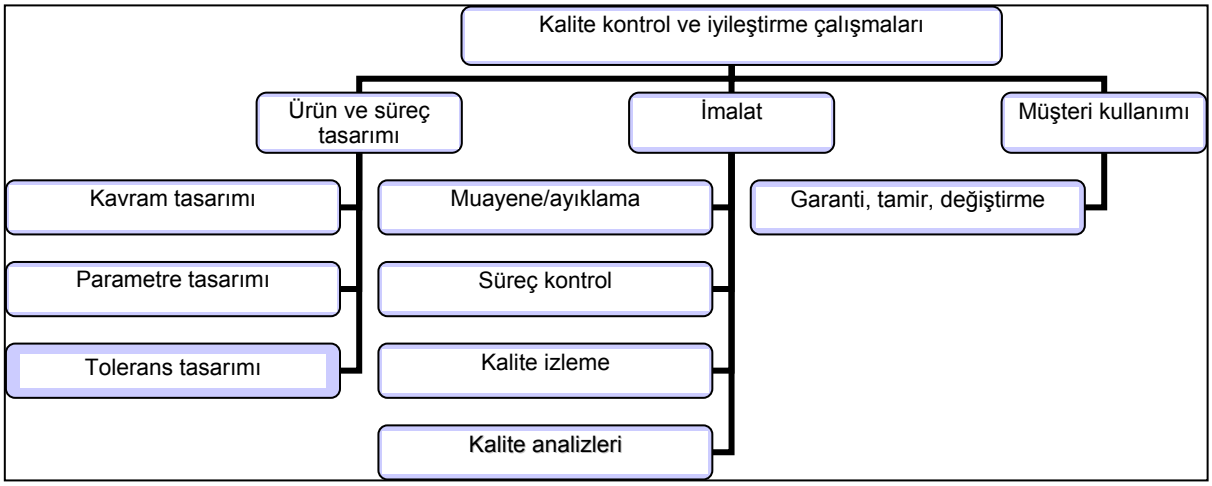
Projenin yöntem geliştirme kısmında ise uygulama aşamasında karşılaşılan bazı problemlerin giderilmesi ve mevcut yöntemlerin kullanım kolaylığı ve/veya etkililiğinin artırılması yönünde çalışmalar gerçekleştirilmiştir. Sonuçta, tipik olarak dengesiz bir yapı gösteren kalite verilerinin yeniden örneklenmesi yoluyla daha dengeli hale getirilmesini sağlayan bir yöntem geliştirilmiştir ve bu yöntemin açık kaynak kodu hazırlanmıştır. Başka bir çalışmada, parametrik olmayan bir regresyon yöntemi olan çok değişkenli uyarlanabilir regresyon eğrilerine (MARS) alternatif bir regresyon yaklaşımı, CMARS, geliştirilmiştir. MARS kalite verilerinin tahmin ve sınıflandırma modellerinin geliştirilmesinde başarılı olmakla birlikte veriye uyum ve modelin karmaşıklığını dengelemede katı bir yaklaşım izlemektedir. CMARS yaklaşımı bu konuda kullanıcıya esneklik sağlamaktadır. Öte yandan, ikili sınıflandırmada kullanımı kolay olan Mahalanobis Taguchi Sistemi (MTS) metodunun gerek duyduğu sınır değerin daha doğru belirlenmesi için bir öneri geliştirilmiştir. Bununla birlikte, MTS çok sınıflı kalite çıktıları için de kullanılabilecek hale getirilmiştir. Ayrıca parametre optimizasyonu için MTS metodunu kullanan bir yöntem ilk defa bu çalışmada geliştirilmiştir. Metot geliştirme çalışmaları kapsamında, bulanık sınıflandırmada kalite verilerine uygun alternatif yaklaşımlar da geliştirilmiştir. Bunlar yaygın kullanılan bir bulanık regresyon yaklaşımına dayandırılmıştır. Ayrıca bulanık tahmin etme ve sınıflandırma fonksiyonlarının parametrik olmayan versiyonları da geliştirilmiştir. Böylece bu fonksiyonların uygulamada etkili sonuçlar alacak şekilde kullanımı mümkün olmuştur. Başka bir çalışmada, parametre optimizasyonunda kullanılan ancak pürüzlü (türevlenemeyen) noktalar içerdiği için optimizasyonu sorunlu olan çekicilik fonksiyonlarının optimizasyonu için alternatif yaklaşımlar geliştirilmiştir. Bu yaklaşımlar, söz konusu optimizasyon problemlerinin çeşitli transformasyonlarını ve topolojik özelliklerini değerlendirmektedir. Bir doktora tezi kapsamında yürütülen bu çalışmalar henüz teorik aşamada bulunmakla birlikte tezin tamamlanması ile ümit verici sonuçların alınması beklenmektedir. Son olarak, birliktelik analizinden elde edilen çok sayıda kuralın anlamlı bir şekilde azaltılması için mevcut bir yaklaşım üzerinde önemli iyileştirmeler yapılmıştır.

Bu rapor, proje aşamalarına uygun olarak, giriş ve sonuç dışında, iki kısımdan oluşmaktadır. İlk kısımda mevcut VM metotlarının uygulamaları, ikinci kısımda ise geliştirilen VM metotları aktarılmaktadır.

2. Kalite İyileştirmede Veri Madenciliğinin Kullanımı

2.1 Literatür Taraması Sonuçları

Bankacılık, sağlık, sigortacılık, pazarlama ve borsa gibi çeşitli başarılı uygulama alanları bulunan veri madenciliği (VM)'nin yakın bir zamanda kullanılmaya başlanılan uygulama alanlarından biri de imalat endüstrisidir. VM bu endüstrilerde kalite kontrol ve iyileştirme, lojistik, müşteri ilişkileri yönetimi gibi farklı amaçlar için kullanılabilmektedir. Bu amaçlardan biri olan kalite kontrol ve iyileştirme çalışmaları, imalat sürecine ait verilerin toplanıp analiz edilerek kalite üzerinde etkili değişkenlerin ve bu değişkenlerin hangi değer aralıklarında kaliteli/hatalı ürün oluşumunu etkilediğinin belirlenmesini içermektedir. Söz konusu bu çalışmalar ürünün ve imalat sürecinin tasarımında, imalat sırasında ve müşteri kullanımında yapılabilmektedir (Bkz. Şekil 2-1).



Şekil 2-1 Kalite kontrol ve iyileştirme çalışmaları (Phadke 1989'dan uyarlanmıştır).

Bu çalışmada, ürün ve süreç tasarımında kalite parametrelerinin optimizasyonu ele alınmıştır. Aynı zamanda, hataya yol açan temel nedenleri anlama ve kalitenin tahmini/sınıflandırılması çalışmaları üzerinde durulmuştur. Şekilde yer alan kalite izleme, süreç kontrol gibi diğer kalite kontrol ve iyileştirme etkinlikleri ise her biri ayrı ayrı geniş bir inceleme gerektirdiğinden dolayı çalışma kapsamı dışında bırakılmıştır (ayrıntılı açıklama Koksal v.d. (2009c) de verilmiştir). Çalışma kapsamında yer alan ve kalite işleri olarak adlandırdığımız kalite kontrol ve iyileştirme etkinlikleri aşağıdaki şekilde sınıflandırılmıştır:

i. *Süreç ve ürün kalitesinin tanımlanması:*
- Kaliteyi belirgin bir şekilde etkileyen değişkenlerin belirlenmesi,
- Kalite üzerinde etkili değişkenlerin önem derecelerine göre sıralanması,
- Veri içinde doğal olarak gruplanan düşük, orta ve yüksek kaliteli ürünlerin ve düşük/yüksek kaliteli ürünleri birbirinden ayıran olası etkenlerin belirlenmesi.

ii. *Kalitenin tahmini*
- Çeşitli girdi değişkenler ile sürekli değerlerle ölçülen kalite karakteristiklerinin arasındaki ilişkinin modellenmesi ve bu model yardımıyla kalite karakteristiklerinin değerlerinin tahminlenmesi.

iii. *Kalitenin sınıflandırılması*
- Çeşitli girdi değişkenleri kullanılarak kategorik değerlerle ölçülen kalite karakteristiklerinin sınıflandırılması ve modellenmesi,
- Hataların sınıflandırılması.

iv. *Parametre optimizasyonu*
- Kaliteyi en iyileyen süreç/ürün parametre değerlerinin belirlenmesi.

Literatürde 1997–2007 yılları arasında yayınlanmış ilgili makaleler araştırılmıştır. Bu araştırmada, belirli VM işleri için kullanılan VM teknikleri, uygulanan süreç ve kalite problemi, kullanılan verilerin özellikleri ve yazılımlar, uygulama sonuçları gibi konularda bilgi toplamak amaç edinilmiştir. Toplanan bilgiler değerlendirilerek kalite iyileştirme alanında çalışanlara yol gösterecek bulgular edinilmiştir. Taranan yayınlar ve bunların değerlendirmesi ile ilgili ayrıntılara Köksal v.d. (2008b) tarafından yayınlanan kapsamlı bir teknik raporda, ayrıca Köksal v.d. (2008c) ve Köksal v.d. (2009c)'de sunulmuştur. Burada söz konusu çalışmadan

elde edilen önemli bulgular özet halinde sunulmakta ve gerçekleştirdiğimiz tipik bir uygulamaya yer verilmektedir.

Aşağıda öncelikle literatürde yer alan ilgili VM süreçleri ve teknikleri tanıtılmaktadır. Daha sonra gelişmiş otomatik veri toplama ve kaydetme sistemlerine sahip olması nedeniyle uygulamalarda sıklıkla karşılaşılan elektronik parça imalatı endüstrisinde yapılmış olan uygulamalara ilişkin bilgilere yer verilecektir. Ek olarak, VM'nin bu endüstrinin dışında kalan metal parça imalatı, demir çelik endüstrisi, hadde endüstrisi gibi diğer endüstrilerdeki uygulamalarından da örnekler verilecektir.

Veri Madenciliği Süreci, İşleri ve Teknikleri

Literatürde VM için yapılmış çeşitli tanımlar mevcuttur. Fayaad'a göre VM 'önemli ilişkilerin ve düzenlerin çok açık görülmediği karmaşık bir gözlem verisinden daha önceden bilinmeyen bir bilgiyi çıkarmak için kullanılan analitik teknikler kümesidir' (Paolo, 2003). Yine literatürde yapılan diğer tanım göre, 'VM, açık ve faydalı sonuçlar elde etmek amacıyla başlangıçta bilinmeyen düzen ve ilişkileri bulmak adına büyük miktardaki verilerin seçimi, incelenmesi ve modellenmesi sürecidir' (Paolo, 2003). Bu çalışmada, kalite iyileştirmede VM, literatürde yer alan tanımlardan da yola çıkarak kalite verisi içinde saklanmış olan ve daha önceden bilinmeyen faydalı ve yenilikçi bilgilerin ortaya çıkarıldığı bir süreç olarak ele alınmaktadır. Bu süreç aşağıda da görüldüğü üzere 1. Veri hazırlama, 2. Veri önişleme, 3. Araştırmalı veri madenciliği, 4. Tahminleyici modeller ve/veya optimizasyon, 5. Değerlendirme ve yorum, olmak üzere beş ana adımdan oluşmaktadır (Bkz. Tablo 2.1). Uygulamalar bu adımların herhangi bir basamağından başlayabilmekte ve diğer adımları ihtiyaca göre birden fazla tekrarlayabilmektedir. Süreçte yer alan adımlar ve kullanılan tekniklere ilişkin ayrıntılı bilgiler için Köksal v.d. (2008b) teknik rapor çalışmasından yararlanılabilir.

Tablo 2-1 Kalite iyileştirmede kullanılan veri madenciliği işleri ve teknikleri

VM işleri	VM alt işleri		Yaygın kullanılan VM teknikleri
Veri hazırlama			Verisi toplanacak değişkeni belirleme, tablo haline getirme, örnekleme
Veri önişleme	Veri temizleme		Kayıp/aykırı/eksik/tutarsız veriyi silme veya yerine uygun değer koyma
	Veri dönüştürme		Düzeltme, normalleştirme, logaritmasını alma, merkeze getirme, kategorik veriyi sayısal hale getirme
	Veri indirgeme	Boyut indirgeme	ÇDR, ANOVA, KA, YSA, YKT, GA
		Veri sıkıştırma	TBA, DD, ÖD
	Kesikleştirme ve hiyerarşik yapı oluşturma		ÖD, KA, DA, k-ortalama, BS, NA
Araştırmalı veri madenciliği	Kümeleme		k-ortalama, değiştirilmiş k-ortalama, MÇB, bulanık c-ortalama, toplayan yöntemler, KDH, TBA, ANOVA
	Özetleme		nokta-dağılım grafiği, OLAP
	Birliktelik kurallarını çıkarma		Apriori
Tahminleyici modeller ve optimizasyon	Sınıflandırma		GDM, NBS, KA (C4.5, C5.0, ID5R), YSA (RBF, KDH, VKÖ), GA, YKT, DVM, BKT
	Tahminleme		ÇDR, DOR, ZSA, YYM, KA (CART), YSA (ÇKP-GYA, ÇKP-LM, RBF, BA), VDD, ANFIS
	Optimizasyon		YYM, TM, YSA (ÇKP-GYA, RBF, USA), GA, AP, BT, BM
Değerlendirme ve yorumlama			Görsel araçlar (çubuk diyagramı, pasta şeması, serpme diyagramı vb.), karar tablosu, karar haritası, karar atlası

VM sürecinin temel adımlarında, VM'nin farklı disiplinleri bir araya getiren bir yaklaşım olması nedeniyle çok çeşitli teknikler kullanılabilmektedir. Bu tekniklere örnek olarak OLAP gibi veri tabanlarından veri özetlemede kullanılabilecek teknikleri, k-ortalama gibi verileri benzerliklerine göre kümelemede kullanılabilecek teknikleri, regresyon gibi değişkenler arasındaki ilişkileri modellemede kullanılabilecek istatistiksel teknikleri, yapay sinir ağları ve genetik algoritmalar gibi optimizasyonda kullanılabilecek teknikleri verebiliriz. Literatürde bu teknikler farklı kıstaslar göz önünde bulundurularak sınıflandırılabilmektedir. Bu çalışmada; birlikteliği belirleme, sınıflandırma, kümeleme gibi bilgiye ulaşma şekli dikkate alınarak sınıflandırma yapılmıştır (benzeri bir sınıflandırma Dunham (2003) tarafından verilmiştir). Bu sınıflandırma sonucunda bir teknik birden fazla sınıfta yer alabildiği gibi bazı teknikler ise bu çalışma kapsamında ele alınmadığından dolayı hiçbir sınıfta yer almamaktadır. Bu çalışmada sadece kalite iyileştirme amaçlı kullanılmış teknikler ile şu ana kadar kullanılmamış fakat gelecekte kullanılabilecek teknikler dikkate alınmıştır.

İmalat Sektöründe Veri Madenciliği Uygulamaları

VM'nin kalite iyileştirme amaçlı kullanımında diğer VM çalışmalarında da görüldüğü gibi bazı zorluklarla karşılaşılmaktadır. Karşılaşılan zorluklardan birisi süreç verisi ve kalite verisi gibi, veri tabanlarının ayrı ayrı tutulması ve analizler için bunların birleştirilmesi gerektiğidir. Bir diğeri, halen kâğıt üzerinde tutulan verilerin yaygın olması ve bunların elektronik ortama taşınmasının zorluğudur. Ayrıca, Rokach ve Maimon (2006) tarafından da belirtildiği gibi kalite verileri genelikle az sayıda, dengelenmemiş (hatalı az, hatasız çok gibi) ve karışık tipte (sürekli ve kategorik veriler bir arada) olmaktadır. VM tekniklerinin büyük veri kümeleri üzerinde daha başarılı sonuçlar verdiği düşünülürse bu tarz veriler üzerindeki uygulamaların eleştirilmesi ve daha uygun yöntemlerin geliştirilmesi gerekmektedir. Ancak yarı iletken endüstrisi gibi bazı endüstrilerde diğerlerine kıyasla büyük veri kümeleri elde edilebilmektedir ve bu sektörde başarılı uygulamalar mevcuttur (Pham ve Afify 2005; Wang vd., 2007a). Ek olarak, VM sonucunda elde edilen bilgilerin yorumlanması ve uygulamaya konulması zor olabilmektedir (Kusiak, 2006; Harding vd., 2006, Wang v.d., 2007a).

Uygulamalarda kullanılan verinin yapısını incelersek, %37 ile en fazla olarak deney tasarımı verisinin kullanıldığı görülmüştür. Bu veri toplama yöntemiyle elde edilmiş verilerin gözlem sayıları 9-1,323 arasında değişmektedir. İkinci olarak %28 ile sayıları 27-16,381 arasında değişen eş zamanlı olarak toplanmış gözlem verileri gelmektedir. Ancak gözlem verilerinin bir kısmı kayıt verileri ile birleştirilip üretim ve kalite verisi olmak üzere farklı iki veri tabanında da tutulabilmektedir. Bu tarz geçmiş veriler tüm veri kümelerinin %20'sini oluşturup gözlem sayıları 27-58,076 arasında değişmektedir. Son olarak uygulamalarda %15 oranında benzetim verisinin kullanıldığı görülmektedir. Bu tarz veriler genellikle geliştirilmiş olan algoritmanın başarımını değerlendirmek amacıyla kullanılmaktadır. Kolay bir veri toplama yöntemi olması nedeniyle gözlem sayıları diğer toplama yöntemlerinin çok üzerindedir.

Uygulamalarda *Neuro Shell Predictor, Qnet for Windows, Neural Works Predict, Professional II/Plus 2000, Rosetta 2005, Fuzzy TECH* gibi çeşitli yazılımlar kullanılmıştır. Sinir ağları için *Matlab NN Toolbox* yaygın olmakla birlikte *C, C++* ve *Visual Basic* gibi programlama dillerinin de kullanıldığı görülmektedir. Uygulamalarda ayrıca *SPSS, Minitab, SAS* ve *Statistica* gibi paket programlarına da yerverilmiştir.

Uygulamalarda sıklıkla kullanılan sektörlerden biri olan bilgisayar ve elektronik parça imalatı üzerinde yapılan çalışmalar öncelikle ele alınmıştır. Bu sektörde yapılmış olan ürün ve süreç kalitesinin tanımlanması çalışmalarına bakarsak, düşük kaliteli silikon parçalarının nedenlerini araştıran çalışmaların yapıldığı görülmektedir (Gardner ve Bieker, 2000; Skinner vd., 2002). Bu çalışmalarda sırasıyla KDH ve hiyerarşik kümeleme tekniklerinden yararlanılmıştır. Benzer şekilde, KA ve birliktelik analizlerinden yararlanarak hatalı silikon parçalarının nedenleri araştırılmıştır (Bertino v.d., 1999). KDH (Karim vd., 2006), hiyerarşik kümeleme (Baek vd., 2005), KA ve YSA (Hsu ve Chien, 2007) kullanılarak silikon kalitesini iyileştirme çalışmaları gerçekleştirilmiştir. Hiyerarşik kümelemeden yararlanarak makinelerle silikon parça kalitesi arasındaki ilişki incelenmiştir (Hu ve Su, 2004). Yarı iletken endüstrisinde k-ortalama kullanılarak hataların ve üretim sürecindeki sapmaların olası nedenleri araştırılmıştır (Chien vd., 2007). Daha iyi süreç koşullarının oluşturulması için çalışmalar yapılmıştır (Kang vd., 1999).

Diğer sektörlerde olduğu gibi bilgisayar ve elektronik parça imalatı sektöründe de ürün ve süreç kalitesini tanımlama çalışmalarını sıklıkla kalite tahmini çalışmaları takip etmektedir. Bu çalışmalarda genellikle YSA tekniklerinden yararlanılmaktadır. Yapılmış olan çalışmalardan örnekler verirsek, kimyasal buhar çökeltme işleminde YSA'dan (Chen, 2007), nokta kaynak işleminde *genelleştirilmiş regresyon sinir ağları*ndan (Tseng, 2006), plastik optik fiber üretiminde parametrik ve parametrik olmayan ZSA, YSA ve VDD'den (Kim ve Lee, 1997) yararlanarak kalite tahmin modelleri kurulmuştur. Silikon bileşik üretim işleminde parametre ayarlarını belirlemek için YSA kullanılmıştır (Li vd., 2003a). Yine YSA'dan yararlanarak baskılı devre levhası imalatı sürecindeki sebep-sonuç ilişkileri ortaya konurken (Shi vd., 2004), ÇDR ve *bulanık uyarlamalı ağ* kullanılarak silikon yonga plakası üretimindeki torna işleminde yüzey pürüzlülüğü ve kesme parametreleri arasındaki ilişki modellenmiştir (Jiao vd., 2004).

Ürün ve süreç kalitesini tanımlama çalışmalarını takip eden diğer bir kalite işi kalitenin sınıflandırılmasıdır. Bütünleşmiş elektrik devresi imalatında YKT (Kusiak, 2000) ve KA'dan (Maimon ve Rokach, 2001; Rokach ve Maimon, 2006) yararlanarak sınıflandırma çalışmaları gerçekleştirilmiştir. YSA ayrıca baskılı devre levhası imalatı (Kusiak ve Kurasek, 2001) ve anakart montajında (Huang vd., 2006) kullanılmıştır. KA ise daha iyi süreç koşullarının oluşturulması (Kang vd., 1999), ürün kalitesinin iyileştirilmesi (Baek vd., 2005; Chien vd., 2006; Chien vd., 2007, Li vd., 2006) çalışmalarında kullanılmıştır. KA, YSA ve *entropi ağları*ndan faydalanarak transformatör demir kayıplarını doğru sınıflandırma yüzdesi artırılmıştır (Georgilakis ve Hatziargyriou, 2002).

Literatürde tanımlanan VM süreçlerinde yer almayan fakat yapılan çalışmalar incelendiğinde karşılaştığımız diğer bir kalite işi parametre optimizasyonudur. Genellikle kalite tahminini takip eden bu iş kapsamında hedeflenen kalite düzeyini verecek süreç/ürün parametrelerinin en iyi düzeylerini belirlenir. YSA, GA ve TM bu amaçla sıklıkla kullanılan tekniklerdir. Bilgisayar ve elektronik parça imalatından örnekler verilirse, YSA kullanılarak iyon implantasyon işleminde çoklu yanıtların eş zamanlı optimizasyonu gerçekleştirilmiştir (Hsieh ve Tong, 2001). Çoklu kuvantum kaynağı ve çığ fotodiyotlarının parametrik optimizasyonunda GA'dan yararlanılmıştır (Kim vd., 2001). Hung (2007) tel şerit tasarımı parametrelerinin optimizasyonunda TM'yi YSA ve GA ile birleştirerek uygulamıştır.

VM'nin literatürde ayrıca metal parça üretimi, plastik imalatı, cam imalatı, kâğıt üretimi, gıda ürünleri imalatı ve kimya endüstrisi gibi çeşitli endüstrilerde uygulaması bulunmaktadır. Bu endüstrilere ilişkin örnekleri incelersek, paketleme imalatında ÖD, GA, KDH ve KA'dan yararlanarak alüminyum folyo kalitesi üzerinde etkili değişkenler belirlenmiştir (De Abajo, 2004). Kalıp halindeki metal parçaların birleştirilmesi sırasında aracın iskeletindeki boyutsal değişimleri kontrol etmek için *korelasyon analizi*, *maksimal ağaç metodu* ve TBA'dan yararlanılmıştır (Lian vd., 2002). Silikon astar üretiminin polimerleşme işleminde (Chiang vd., 2002) ve plastik enjeksiyon kalıplama işleminde (Özçelik ve Erzurumlu, 2006) kalite değişkenlerinin belirlenmesinde ANOVA kullanılmıştır. Bükme yün eğirme işleminde ÖD, ÇDR ve *gri üstün analizi*nden faydalanarak önemli değişkenler seçilmiştir (Yin ve Yu, 2006).

Yukarıda belirtilen endüstriler içerisinde özellikle metal parça imalatı endüstrisinde çok sayıda kalite tahmini çalışması yapılmıştır. ÇDR, DOR ve YSA kaynak işlemindeki üst bilye genişliğini (Kim vd., 2003), YSA ve YYM ile birlikte kalıp yüzeyindeki pürüzlülüğün hata değerini tahminlemede kullanılmıştır (Erzurumlu ve Öktem, 2007). Galvenizlenmiş çeliğin mekanik özelliklerinin (Ordieres Mere vd., 2007), çelik levha üretiminde kaliteyi etkileyen değişkenlerin (Deng ve Liu, 2002), tornalama işleminde yüzey pürüzlülüğünün (Lin ve Wang, 2000), parmak freze kesme işleminde yüzey pürüzlülüğünün (Tsai vd., 1999) tahmininde ÇDR ve YSA'dan yararlanılmıştır. YSA ayrıca metal asal gaz kaynağı işlemi (Tay ve Butler, 1997), çelik kaynak işlemi (Cool vd., 1997; Vasudevan vd., 2002; Vasudevan vd., 2005), lazer kaynak işlemi (Olabi vd., 2006), çelik döküm işlemi (Perzyk vd., 2005), döküm işlemi (Batmaz, 2007), CNC tornalama işlemi (Suneel vd., 2002) gibi farklı işlemlerde de kalite tahmin amaçlı kullanılmıştır. YSA metal parça imalatı dışında ise, plastik enjeksiyon kalıplama işlemi (Sadeghi, 2000; Shen vd., 2007; Kurtaran vd., 2005; Özçelik ve Erzurumlu, 2006), ısıl sprey işlemi (Guessema vd., 2004), plazma sprey işlemi (Wang vd., 2007b), propandan propana oksidatif dehidrojenasyon (Holena ve Baerns, 2003), etilen proliz işlemi (Zhou vd., 2006) gibi farklı endüstrilerde uygulanmıştır.

Sınıflandırma çalışmaları incelendiğinde; alüminyum kaplama işlemi (Baek vd., 2002), kalıp halindeki metal parçaların montajı (Lian vd., 2002), döküm işlemi (Bakır vd., 2006), yüksek hassasiyetli üretim (Huang ve Wu, 2005) ve otomobil montajında (Wang, 2007) KA'dan yararlanıldığı görülmektedir. Benzer şekilde, YSA yardımıyla hadde endüstrisinde (Cser vd., 2001), alüminyum folyo üretiminde (De Abajo vd., 2004), basınçlı döküm işleminde (Krimpenis vd., 2006), kesme işleminde (Lee ve Dornfeld, 2007), öngermeli beton kazığı üretiminde (Tam vd., 2004) ve bir enerji santralinde (Tan vd., 2007) sınıflandırma çalışmaları gerçekleştirilmiştir. Bunlara ek olarak, plastik enjeksiyon kalıplama işleminde hata belirleme ve teşhisi için YSA ve DVM'den yararlanılmıştır (Ribeiro, 2005). Sarimveris vd. (2006) tarafından yapılan çalışmada ise YSA ve BKT kullanılarak kâğıt üretiminde ürün kalitesini sınıflandırma modeli kurulmuştur.

Çok sayıda sürecin kalite parametrelerinin optimizasyonu GA'dan yararlanarak yapılmıştır (öğütme (Brinksmeier vd., 1998), nokta kaynak (Tseng, 2006; Hamedi vd., 2007), makine işleminde kesme (Cus ve Balic, 2003), yonga levha üretimi (Cook vd., 2000) ve plastik enjeksiyon kalıplama (Shen vd., 2007; Kurtaran vd., 2005)). YSA da bu amaçla çok kullanılan tekniklerdendir (sıcak hadde endüstrisi (Cser vd., 2001), metal asal gaz kaynağı (Tay ve Butler, 1997; Meng ve Butler, 1997), çelik üretimi (Liu vd., 2004), sinterleme (Zhang vd., 2007) ve polimerizasyon (Chiang vd., 2002)). Ek olarak, lazer kaynak işleminin (Olabi vd., 2006) ve kesme işleminin (Lee ve Dornfeld, 2007) optimizasyonu TM kullanılarak gerçekleştirilmiştir.

Sonuç

Çalışmada VM yaklaşımlarının imalat endüstrilerinde kalite iyileştirmede kullanımını ortaya koymak ve yeni uygulamaları teşvik etmek amacıyla öncelikle VM süreci ve teknikleri tanımlanmıştır. Literatürde 1997 – 2007 yılları arasında yayınlanmış çalışmalar, tanımlanan VM süreci ve seçilen kalite iyileştirme işleri dikkate alınarak özetlenmiştir. Literatür çalışması, VM çalışmalarının imalat sektörünün hemen her alanında yapılmakta olduğunu göstermektedir.

Mevcut durumda kalite iyileştirme problemleri, alışılmış istatistiksel ve yönetim yaklaşımları (altı sigma, istatistiksel süreç kontrolü gibi) ile ele alınmaktadır. Ancak problemlere ilişkin toplanan veriler yüksek hacimli olduğunda ve/ya çok sayıda değişken içerdiğinde ve/ya bu değişkenler karışık tipte olduğunda veri analizi mevcut yaklaşımlar ile etkili bir şekilde yapılamamaktadır. Gerek literatürde yer alan çalışmalar gerekse bu raporda aktarılana benzer uygulamalar, VM yaklaşımlarının bu durumları da içeren kalite kontrol, iyileştirme, süreç optimizasyonu ve ürün tasarımı gibi alanlarda başarılı olabildiğini göstermektedir. Yine de bu uygulamaların yeterince yaygın ve kabul görmüş olduğunu söylemek mümkün değildir. VM tekniklerinin çoğunun yazılım ve uzman desteği gerektirmesi; kalite, süreç ve ürünler ile ilgili verilerin VM analizlerine uygun şekilde tutulmaması (bilgi sistemi altyapısı ile ilgili eksiklikler) uygulamaların az olmasının en önemli nedenleri olarak düşünülebilir. VM uygulamaları veri toplama ve analizi için uygun yazılım ve yetişmiş insan gücü desteği ile daha çok sayıda endüstri kuruluşunun kalite iyileştirme çalışmalarına hız ve etkililik kazandırabilir. Akademik çalışmalar özellikle kalite verilerinin az sayıda, dengesiz ve karışık tipte olduğu durumlar için daha etkili analiz ve çözüm yöntemleri geliştirme, elde edilen sonuçların yorumlanması ve değerlendirilmesini kolaylaştırma doğrultusunda ilerleyebilir.

İmalat sektörü dışındaki kuruluşlarda da kalite iyileştirme ve kontrol amaçlı benzer ya da farklı yeni uygulamalar olması kaçınılmaz gözükmektedir. Örneğin, müşteri servis verilerinin daha etkili analizi yoluyla müşteri memnuniyetinin artırılmasında VM'nin yeni uygulamalarından Metin Madenciliği, e-iş ortamında kalitenin iyileştirilmesinde Ağ Madenciliği gibi yaklaşımlar daha yaygın olarak kullanılabilecektir.

2.2 Saha Çalışması Sonuçları

2.2.1 Gözlemler

Proje kapsamında, imalat sanayinin kalite iyileştirme gereksinimleri ve buna yönelik veri toplama ve analiz uygulamaları konusunda bilgi almak ve proje için gerekli olan verileri toplamak için değişik alanlarında etkinlik gösteren altı firma ziyaret edilmiştir. Bunlar ERKUNT Döküm, TOFAŞ Otomotiv, VESTEL Elektronik, DİMES Meyve Suyu, AKSA Akrilik ve PETKİM Petrokimya firmalarıdır. Ayrıca SPAC Altı Sigma Danışmanlık firmasının kalite iyileştirme faaliyetlerinde çeşitli sanayi kuruluşlarımızın veri toplama ve analizi konusundaki deneyimleri hakkında bilgi alınmıştır.

Bu firmalarda yapılan incelemeler sonucunda proje ilgi alanımıza giren kalite iyileştirme amaçlarının genel olarak aşağıdaki gibi ifade edilebileceği anlaşılmıştır.

- kalite çıktılarını etkileyen ürün, süreç ve diğer değişkenleri ve bunlara göre çıktı değerini doğru belirleme,
- istenen kalite düzeyini verebilecek ürün ve süreç değişken (ayar veya nominal) değerlerini belirleme,
- bu belirlemeleri kolay ve hızlı yapma,
- sonuçları kolay ve doğru yorumlama ve kullanabilme.

Söz konusu amaçlara yönelik olarak döküm firmasında bir üründe gözlenen hataları azaltmak amacıyla değişik ürün ve süreç değişkenleri ile ilgili verinin toplanmakta olduğu ve bu verinin analizinde alışılmış istatistiksel yöntemlerin kullanıldığı gözlenmiştir. Bu analizler sonucunda süreçte yapılan değişikliklerin yeterli bulunmadığı anlaşılmıştır.

Otomotiv firmasında ise belli bir hafif ticari aracın sürücü koltuğunun mevcut tasarımdan memnuniyeti etkileyen faktörleri araştırmak üzere bir anket yapılmıştır. Anket verilerinin analizinde basit istatistiksel yöntemlerin uygulanmasıyla elde edilen sonuçların tasarımcıya yeterince yol gösteremediği belirlenmiştir.

Öte yandan, elektronik kart üreticisinin ürün tiplerinin siparişe göre sıklıkla değiştiği ve üretim hacminin yüksek olduğu, bu nedenle üretilen kartlar üzerindeki hataların kaynaklarını hızlı bir şekilde belirleme ve bunları gidermenin büyük önem taşıdığı anlaşılmıştır. Bu amaçla firma etkili bir veri toplama ve analiz yöntemine gereksinim duymaktadır. Mevcut durumda kısıtlı verinin elektronik ortamda toplanması sonucu

yapılan basit istatistiksel analizlere dayalı bir karar verme süreci işletilmektedir. Bu da zaman kaybına neden olmakta ve hataların kaynaklarını belirlemeyi zorlaştırmaktadır.

Meyve suyu üreticisi de benzer şekilde süreçlerinde, özellikle de kutulama sürecinde gözlenen hataların kaynaklarını belirlemek ve bunları gidermek istemektedir. Bu amaca yönelik kullanılabilecek verilerin daha çok hatalar ile ilgili çeşitli istatistikler olarak tutulduğu gözlenmiştir. İncelemeler sonucunda toplanan verinin, özet olması niteliğinden dolayı ancak hatalar ile ilgili belli odaklara işaret edebileceği, daha etkili sonuçlar elde edebilmek için süreçler ve kalite çıktıları ile ilgili daha detaylı veri toplanması gerektiği belirlenmiştir.

Elyaf üreticisinin ise bütünleşik ürün ve süreç tasarımına yönelik bir istatistiksel deney tasarımına göre elde edilen veriler toplandığı gözlenmiştir. Bu veriler sayıca çok az olduğundan alışılmış parametre optimizasyonu metotları ile incelenebilmekte ve etkili sonuçlar alınabilmektedir. Bu nedenle söz konusu veriler için ayrıca diğer veri madenciliği yaklaşımlarının uygulanması düşünülmemiştir.

Son olarak, incelenen petrokimya üretiminin sürekli akış tipinde olduğu ve geri beslemeli mühendislik kontrol sistemleri yardımıyla, tam otomasyon ortamında üretilen petrokimya ürünlerinin belli özelliklerinin istenen sınır değerler arasında sağlanmasına çalışıldığı belirlenmiştir. Bu ortamdan elde edilen veriler bir zaman serisine aittir ve kontrol sisteminin sürekli gerçekleştirdiği düzeltmeler süreçteki hata kaynaklarını gizlemektedir. Bu nedenlerle söz konusu verilerin proje amaçlarımıza uygun olmadığına karar verilmiştir.

Genel olarak, ziyaret ettiğimiz kuruluşların veri toplama ve hazırlama altyapısı ve yöntemleri ile ilgili sorunları bulunduğu belirlenmiştir. Tipik olarak kalite iyileştirme amacına yönelik gözlemsel verilerin girdi değişken değerleri ile çıktı değişken değerlerinin ayrı ortamlarda ve ayrı zaman periyotları için saklandığı ve bunları sonradan birbiri ile ilişkilendirmenin mümkün olmadığı veya çok güç olduğu gözlenmiştir. Deneysel veriler ise daha düzgün ve amacına yönelik olarak toplanmaktadır. Veri madenciliği için tercihan elektronik ortamda, süreç ve kalite çıktı verilerini bütünleşik saklayan sistematik ve güvenilir veri toplama düzenlerine gereksinim vardır.

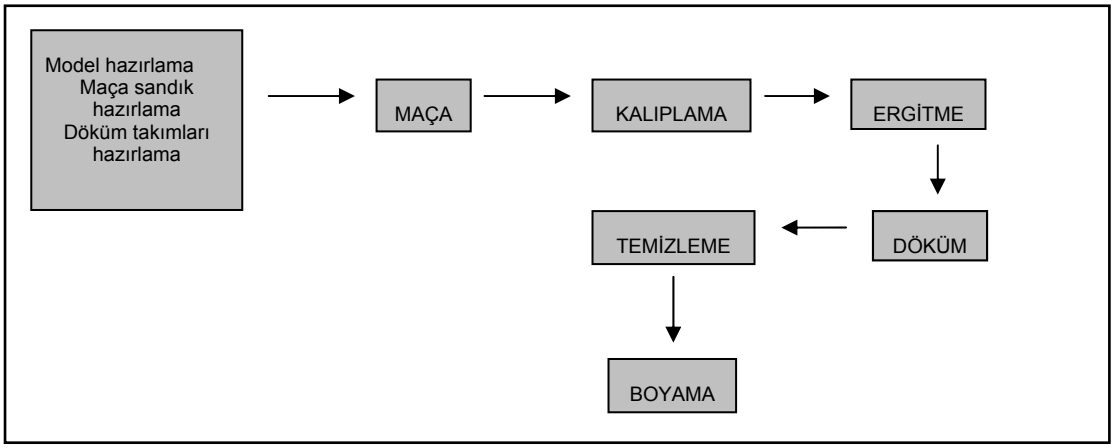
2.2.2. Kısımda döküm, koltuktan müşteri memnuniyeti ve elekrtonik kart verileri üzerinde yaptığımız VM uygulamaları ile sonuçları aktarılmaktadır.

2.2.2 Uygulamalar

2.2.2.1 Veri hazırlama ve Önişleme

2.2.2.1.1 Döküm Verisi

Ankara Sincan Organize Sanayi Bölgesinde faaliyetini sürdüren Erkunt Döküm Sanayi kuruluşundan özel bir ürün için, 2006 yılının ilk beş ayına ait süreç verisi alınmıştır. Seçilen ürün, firmanın en çok kalite problemi yaşadığı ürünlerden biridir. İlgili döküm sürecinin akışı Şekil 2-2'de gösterilmiştir. Bu akış içerisinde her bir evrede yapılan parametre ayarları çok çeşitli hata tiplerinin gözlemlenmesine neden olabilmektedir.



Şekil 2-2 Döküm sürecinin akış şeması

Elde edilen veriler maça, kalıplama ve ergitme gibi birbirini takip eden üç alt imalat sürecinde gözlemlenmiştir. Firma, bir parti ürünü temsil edecek şekilde özet veri toplamakta, dolayısı ile, üretilen her bir ürünün gerçek süreç değerleri bilinmemektedir. Tüm değerler örnekleme yoluyla bir veya birkaç kez ölçülmektedir. Ölçümlerin büyük bir bölümü üretim esnasında alınmaktayken, kalıplama hattına ait bazı değişkenlerin ölçümleri haftada bir kez yapılmaktadır. Bu değerler, ilgili haftada gerçekleşmiş tüm üretimler için geçerli sayılmaktadır. Değişkenlerin çoğu için ölçümlerin yapıldığı zaman kaydedilmemiştir. Bu sebeple, örnekleme ile elde edilen değerler arasında bir eşleştirme yapılması mümkün olmamıştır. Veri toplama prosedürünün ortaya koyduğu bu kısıtlardan ötürü, her partinin süreç değerleri o partiye ait örnekleme değerlerinin ortalamaları ile temsil edilmiştir. Bunun sonucu olarak da, belli bir ölçüde bilgi kaybı gerçekleşmiştir.

Süreç değişkenleri (46 değişken), döküm sürecinin 3 temel alt süreci olan maça, kalıplama ve ergitme hatlarından seçilmiştir. Çıktı değişkenleri ise bir parti üretimde gözlenen toplam hatalı ürün sayısını ve bunların 10 farklı hata sınıfına göre dağılımlarını ifade etmektedir. Gerçekte, bir üründe birden fazla hata tipi oluşabilmesine rağmen, firma sadece, ürünün atılmasına sebep olan tek baskın hatayı kaydetmektedir. Böylece, hata tipleri arasında varolabilecek ilişkiye yönelik veriye erişmek mümkün olmamaktadır. Partide toplam kaç ürün üretildiği de ayrıca kaydedilmektedir.

Elde edilen ilk veri kümesi her satır bir partiyi gösterecek şekilde, 95 satırdan oluşmuştur. Veri temizleme aşamasında, çok sayıda (%50'den fazla) kayıp değer içeren 11 süreç parametresi, uzak değer olarak belirlenen 3 satır veri kümesinden çıkarılmıştır. Ayrıca, firmadaki kalite grubu 4 hata tipinin seçilen süreç parametreleriyle ilişkisiz olduğunu, bu hataların parça uyumsuzlukları veya tasarım hatası gibi sebeplerle oluştuğunu belirtilmiştir. Bu sebeple, hata kategorileri 6'ya indirilmiş ve toplam hatalı ürün sayısı ve üretim miktarı da bu ölçüde azaltılmıştır. Firma ile yapılan görüşmeler sonucu, çalışmaların en önemli hata tipi olan azot gazı hatası için yürütülmesi kararlaştırılmıştır. İlk veri ön işleme işlemleri sonucunda ortaya çıkan veri kümesinde yer alan değişkenlerin açıklamaları aşağıdaki gibidir:

- Üretim miktarı
- Toplam fire sayısı
- N gazı boşluğu fire sayısı
- H gazı boşluğu fire sayısı
- Kum düşmesi fire sayısı
- Maça kırık fire sayısı
- Sızdırma fire sayısı
- Yüzey bozuk fire sayısı
- Kerntop Gaz Geçirgenliği
- Rheotech Gaz Geçirgenliği
- Boya Viskozite MC01
- Boya Viskozite MC02
- Aktif Kil
- Gaz geçirgenliği
- Kompaktibilite
- Nem
- Makas Mukavemeti
- Basma Mukavemeti
- Ayırma Mukavemeti
- Akışkanlık
- Yanma Kaybı
- Kum Sıcaklığı
- Alt mala
- Alt yan
- Üst mala
- Üst yan
- Topuk aşısı
- Ağız aşısı
- Döküm süresi (sn)
- Pota sıcaklığı
- Pota_Çil genişliği
- Pota_Çil derinliği
- Ocak sıcaklığı(dokum sıcaklıgı)
- Ocak_Çil genişliği
- Ocak_Çil derinliği
- %C

- %Si
- %Mn
- %Cr
- %Cu
- %Sn
- %S
- %Ti

Tahmin etme çalışmalarında kullanılan kesiksiz çıktı değişkeni her partide yer alan hatalı ürün oranıdır. Burada, hatalı ürün azot gazı hatasını (y2) ifade etmektedir. Bu değerler, azot gazı bakımından hatalı ürün sayısının üretim miktarına bölünmesi ile elde edilmiştir. Sınıflandırma çalışmalarında kullanılan çıktı değişkeni ise ayrıklaştırma (discretization) çalışması yapılarak elde edilmiştir. Aynı üretim koşullarında hem hatalı hem de hatasız ürün çıkması, yani bir partide yer alan hatalı ve hatasız ürünlerin hepsinin aynı süreç değerleri ile temsil edilmesi, ayrıca hatasız ürünlerin veri kümesinde sayısal olarak baskınlığı ürün bazında sınıflandırma yapmaya olanak vermemektedir. Kalite problemlerinde sıkça karsılasılan bu tip veri ile başarılı sınıflandırma modelleri oluşturulamamıştır. Bu yüzden, tahmin etme çalısmalarında kullanılan 92 güne ait döküm verisinde azot gazı hata yüzdesi olan çıktı değiskeni, hata oranı ve günlük üretim miktarı dikkate alınarak iyi (veya kabul edilebilir) parti ve kötü (veya kabul edilemez) parti olmak üzere 2 sınıfla ifade edilmistir. Döküm verisi ile sınıflandırma ve metotların karşılaştırılması çalışması olusturulan bu veri kümesi üzerinden yapılmıştır.

Veri kümesinin küçük olması sebebi ile, tahmin etme ve sınıflandırma yöntemlerinin karsılastırılmasında çapraz doğrulama yaklasımı kullanılmıs ve bölünme sayısı 3 olarak belirlenmistir. Ayrıca, sonuçların güvenilirliğini artırmak için 3 tekrar yapılmıştır. Böylece, her modelleme ve modele iliskin performans ölçümleri 3 kez tekrarlanan 3-katlı çapraz doğrulama sonucu 9 kez farklı bölünmeler kullanılarak yapılmıştır. Bölünmelerin orijinal kümenin dağılımını yansıtması için tabakalı örnekleme yöntemi kullanılmıştır.

Sonuç olarak, Erkunt Döküm Sanayi kuruluşu veri toplama prosedürü bakımından değerlendirildiğinde, etkili bir veri toplama ve işleme sistemlerinin olmadığı görülmüştür. Verilerin süreci daha iyi temsil edebilmesi için örnekleme ve kayıt yöntemlerinin yeniden planlanması gerekmektedir. Ayrıca, veriler kağıt formlara işlenmekte olup, düzenli bir şekilde elektronik ortama aktarılmadığından, her hangi bir analiz çalışması uzun bir veri hazırlama aşamasından sonra gerçekleşebilmektedir. Bu da, hızlı çözümler elde etmeyi engellemektedir. Tüm bu problemler firmaya aktarılmış ve veri madenciliği çözümlerinin etkili bir şekilde kullanılabilmesi için gerekli olan veri toplama prosedürü konusunda öneriler sunulmuştur.

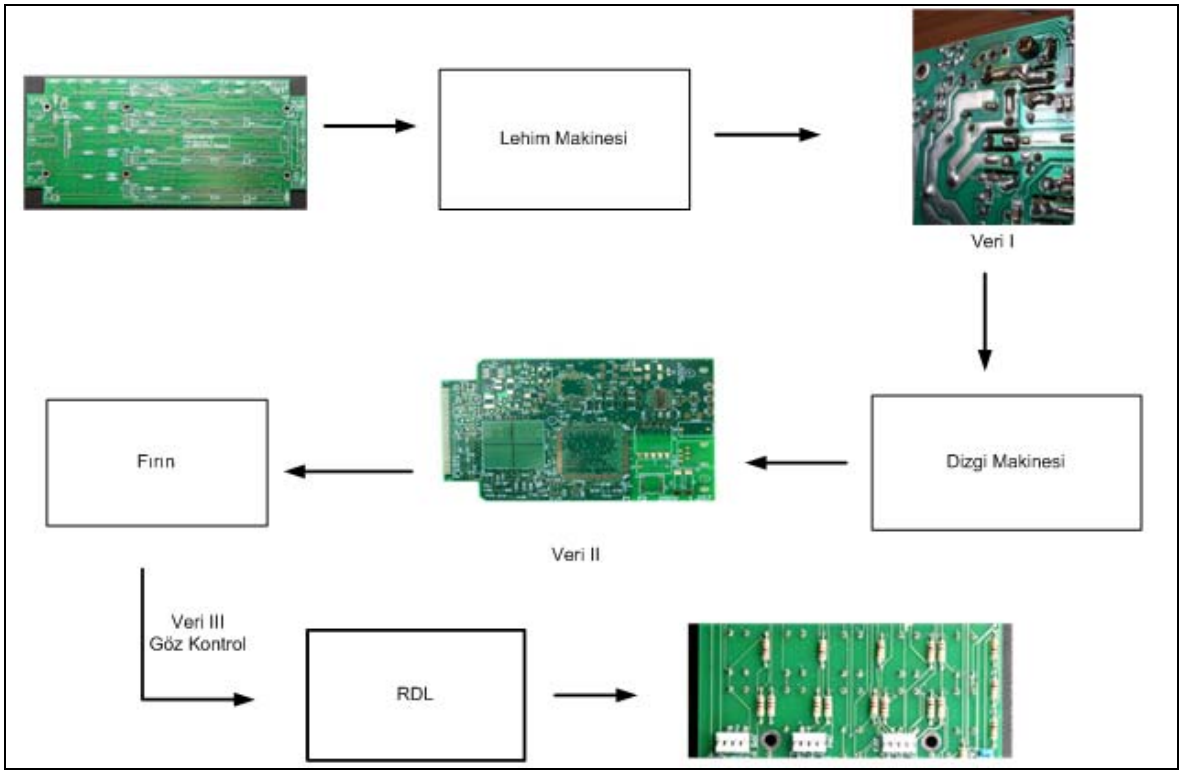
2.2.2.1.2 Müşteri Memnuniyeti Verisi

TOFAŞ'ın ürettiği bir hafif ticari aracın sürücü koltuğunun müşteri memnuniyetini artıracak şekilde yeniden tasarımlanması kapsamında müşteri memnuniyeti etkileyen faktörlerin ve seviyelerinin belirlenmesi amacıyla, seçilen aracın müşteri segmentinden rassal seçilen 80 kişiye bir anket formu uygulanmıştır. Buna göre, yaş, cinsiyet, araç ile katedilen yol, antropometrik ölçüler, kullanım kolaylığı, estetik görünüm gibi girdi değişkenleri ile sırt rahatlığı, basen rahatlığı, genel memnuniyet, kullanım kolaylığı, göze hoş görünmesi gibi çıktı değişkenleri belirlenmiştir. Çalışmalarda çıktı değişkeni olarak genel memnuniyet ele alınmış ve bu değişkenin seviyeleri, müşterilerin cevapları da dikkate alınarak, iki sınıflı değişken oluşturacak şekilde birleştirilmiştir.

Koltuktan memnuniyetin sınıflandırılması probleminde, veri üçe bölünürken memnuniyet değerinin (çıktı) dağılımı yine her parçada aynı olacak sekilde tabakalı örnekleme yöntemi kullanılmıştır. Döküm verisinde olduğu gibi 3 tekrar yapılmıştır.

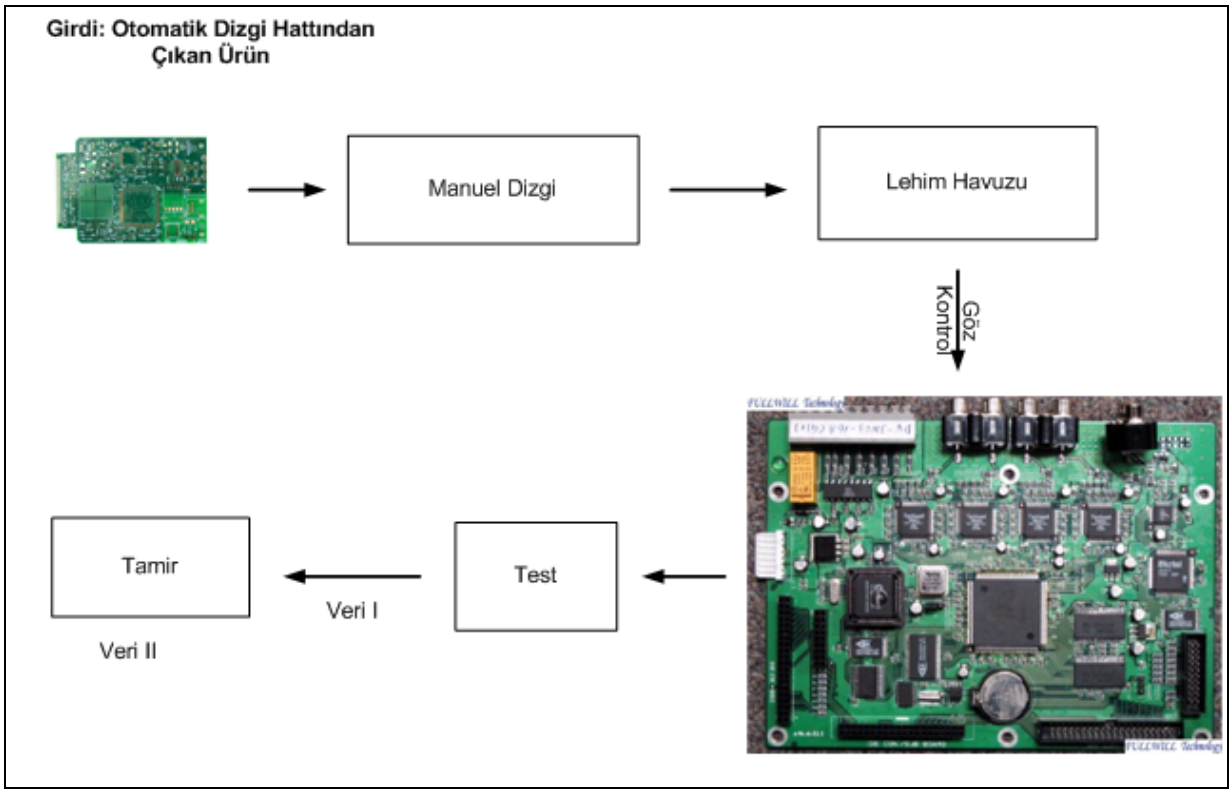
2.2.2.1.3 Elektronik Kart Verisi

Manisa'da faliyet gösteren VESTEL firmasından, DVB ürünleri için üretilen PCB (printed circuit board) kartları ile ilgili 2006 yılına ait veriler alınmıştır. Otomatik ve manuel dizgi olmak üzere iki ana hattan oluşan elektronik kart imalat sürecinin akışı Şekil 2-3 ve Şekil 2-4'de verilmiştir. Günde yaklaşık 35 – 40 bin, ayda 1,5 – 2 milyon kart üretilen firmada aylık kayıp 70 – 80 bin seviyelerinde gerçekleşmektedir.



Şekil 2-3 Otomatik Dizgi Hattı

Otomatik dizgi hattının üç noktasında otomatik kontrol makineleri ile veri toplama işlemi yapılmaktadır. Bu aşamada, hatanın gerçekleştiği ürün ile eşleştirme yapılmaksızın, çeşitli hata tiplerinin (örneğin; lehimler arası kısa devre, kayık malzeme, ters malzeme, vb.) gözlenme sıklığı kayıt altına alınmaktadır. Üç veri toplama noktasında da, gözlenen hatalar düzeltilmekte ve bir sonraki üretim aşamasına iletilmektedir. Üçüncü veri toplama noktasından sonra, hatasız hale getirilmiş ürünlere barkod numarası verilmekte ve bu aşamadan sonra manuel dizgi hattında kaydedilen hatalar ürün ile eşleştirilebilmektedir.



Şekil 2-4 Manuel Dizgi Hattı

Kartların tamir hattında tutulan arıza kayıtları 28778 gözlem içermektedir. Söz konusu veriler, test aşamasından itibaren kaydedilmektedir. Burada, testi gerçekleştiren görevli kartta gözlediği hatayı veya hataları (görüntü hatası, ses hatası gibi) kartın barkod numarasına göre sisteme girmektedir. Bir kart üzerinde birden fazla hata kaydedilebileceği gibi bazı hata tipleri diğer hataların ölçümünü engelleyebilmektedir. Tamir hattına gönderilen kart detaylı incelenerek bildirilen arızaların sebepleri bulunur (hat kopuk, parça dik, lehimsiz, kısa devre gibi). Arızaya sebep olan parça(lar) onarılır veya yenisi ile değiştirilir. Kimi durumlarda kart tamir edilemeyecek şekilde zarar görmüş olabilir. Kartta gözlenen hatalara ilişkin bilgiler ve yapılan düzeltmeler veritabanına kaydedilir. Tamirden gelen kartlar ikinci kez testlere tabi tutulur, gerekiyor ise tekrar tamire gönderilir. Ürün bazlı ham veri kümesi için kayıt altına alınan bilgiler şunlardır:

- Ürün türü (ÜT): Kartın hangi ürün türüne ait olduğunu gösterir
- Yarı mamul kodu (YMK): Kartın tasarım bilgilerini içerir
- İş emri (İE): Müşteri sipariş bilgilerini içerir
- Seri no (SN): Karta verilen numaradır
- Arıza açıklaması (AA): Kartın arızasına ilişkin test personelin yaptığı açıklamadır
- Arıza sebebi (AS): Kartın arızasına ilişkin tamir personelin teşhis sonucudur
- Arıza konumu (AK): Arızanın kart üzerindeki konumunu gösterir
- Personel: Testi ve tamiri gerçekleştiren personel bilgisidir
- Tarih: Testin ve tamirin gerçekleştirildiği tarihtir
- Tamir edilen yer (TEY): Tamirin gerçekleştirildiği hattır
- Üretim no (ÜN): Manuel sipariş numarasını gösterir
- Tamire ayrılan bant (TAY): Kartın hangi bölümden tamire ayrıldığını gösterir
- Kart tipi (KT): Kartın nereden geldiğini ve hangi bölüme ait olduğunu gösterir
- Tedarikçi: Kartın ham olarak hangi firmadan tedarik edildiğini gösterir

Firma ile yapılan görüşme sonucu, mevcut veri kümesi de dikkate alınarak, aynı arızaların genelde aynı sebeplerle birlikte gözlenip gözlenmediği, belirli arızaların birlikte gözlenip gözlenmediği ve arıza ve sebeplerinin belirli bir tedarikçi ile birlikte gözlenip gözlenmediğinin irdelenmesi kararlaştırılmıştır. Ortaya konan bu sorulara cevap verebilecek bir yaklaşım birliktelik analizidir. Bu sebeple veri ön işleme çalışmaları bu yaklaşıma yönelik gerçekleştirilmiştir. Öncelikle, analizlerde yer almayacak değişkenler veri kümesinden çıkarılmıştır. Örneğin, personel, tamir edilen yer ve üretim no bilgileri firma tarafından önemsiz görülmüştür ve analizlere dahil edilmemiştir. Bazı değişkenler ise eşleştirme amacı ile kullanılmıştır. Örneğin, yarı mamul kodu, iş emri ve seri no bilgileri bir ürünü tek olarak belirlemektedir. Ayrıca tarih ve arıza konumu bilgileri de bu aşamada ilgisiz görülmüştür. Dolayısı ile söz konusu sekiz değişken veri kümesinden çıkarılmıştır.

Veri önişleme aşamasında karşılaşılan en önemli iki sorun kartlarda saptanan arıza açıklamaları ve arıza sebepleri ile ilgilidir. Birinci problem, tamir ve test personelinin, gözlemlediği arızaları sisteme belli bir standartta girmemesinden kaynaklanmıştır. Bu sebeple, gerçekte 51 farklı değer alan arıza açıklaması ve 18 farklı değer alan arıza sebebi değişkenleri için, sistemde aynı hatayı ifade eden farklı bir çok ifade yer almıştır. Firma ile görüşülerek, gerçekte aynı anlama gelen ifadeler saptanmış ve standart bir değerle değiştirilmiştir. Diğer problem ise hata ve hata sebebi yerine barkod numaralarının girilmiş olmasıdır. Bu şekilde kaydedilmiş gözlemlerin gerçek değerleri belirlenememiştir. Dolayısı ile, bu kayıtların arıza açıklaması ve arıza sebebi değerleri eksik veri olarak değerlendirilmiştir.

Veri kümesi, uygulanacak analizlere ve kullanılacak algoritmalara göre yeniden yapılandırılmıştır. Örneğin, ürün bazlı birliktelik analizlerinde kategorik olarak ifade edilen arıza açıklaması, arıza sebebi, ürün türü, tedarikçi, kart tipi ve tamire ayrılan bant değişkenleri, kategori sayıları kadar, ikili değişkene dönüştürülmüştür. Elde edilen veri kümesinde, 28778 kayıt 87 tane ikili değişken ile ifade edilmiştir.

2.2.2.2 Kümeleme

Kümeleme bir veri kümesindeki homojen grupların bulunması işlemidir, sınıflandırma ve tahminden farklı olarak bu grupların etiketi bilinmemektedir. Kümeleme işlemiyle kalite iyileştirmede ihtiyaç duyulan kural kümeleri ve tahmin modelleri gibi çıktılar elde edilmez, verideki ilişkili veri grupları tanımlanır. Bu nedenle önemli bir veri madenciliği yaklaşımı olan kümelemenin kalite iyileştirmede tek başına bir yöntem olarak kullanılamasa bile, sınıflandırma ve tahmin amaçlı yöntemler öncesi veri de önişleme yapmak amacıyla kullanılmasını önermekteyiz. Çünkü kümeleme analizi sırasında, veri kümesindeki uç değerler ve azınlıkta olan veriler saptanabilmektedir. Kalite iyileştirmede veri madenciliği çalışmalarında kümeleme kullanacaklar için bu çalışma şunu göstermiştir: sınıflandırma ve tahmin amaçlı yöntemlerin, özellikle gerçek veriler için kesin sonuçlar vermediği ya da tahmin gücü yüksek modeller üretemediği durumlarda, veride kümeleme yapılarak homojen grupların belirlenmesinden sonra incelenmesi gerekir.

Proje kapsamında yürüttüğümüz kümeleme çalışmalarında döküm verisine uygun gördüğümüz 6 kümeleme yöntemi bu veri kümesi üzerinde test edilmiştir: **k-ortalamalar (k-means), medoidler etrafında bölümleme (MEB), değiştirilmiş k-ortalamalar, kendi kendini düzenleyen haritalar (SOM), bulanık c-ortalamalar (BCO), aşamalı tam bağlantı yöntemi (H/C)**. Kümeleme işleminin, diğer veri madenciliği yöntemleri öncesinde veri ön işleme amaçlı kullanılabileceği ve karar ağaçları gibi yöntemlerin tüm veri kümesi yerine bulunan bir küme üzerinde uygulanarak hata oranlarının azaltılabileceği gösterilmiştir. Bu projede SPSS Clementine 10.1'deki k-Ortalamalar yazılımı ve diğer tüm uygulamalar için MATLAB ile kodlanmış kümeleme yazılımları kullanılmıştır. SPSS Clementine paket programı kümeleme yöntemlerinden sadece k-Ortalamalar ve 2 adımlı kümelemeyi içermektedir ve kümeleme yöntemlerinin geçerliliğini ölçmede kullanılan indisleri içermemektedir. Bizim çalışmamızdaki k-ortalamalar, MEB, SOM ve H/C uygulamaları için MATLAB'da kodlanmış Cluster Validity Analysis Platform (CVAP) v. 3.42_7 yazılımı kullanılmıştır. Bu ücretsiz yazılım, kümeleme ve kümeleme geçerlilik indisleri ile ilgili literatürde bulduğumuz en kapsamlı ve kullanımı kolay yazılımdır. Bulanık c-ortalamalar uygulaması için "Fuzzy Clustering and Data Analysis Toolbox" yazılımı kullanılmıştır (Balasko, 2006). Değiştirilmiş k-ortalamalar için Adil Bagirov'un geliştirdiği yazılım kullanılmıştır. Kullandığımız yöntemler bu metodları kullanacak uygulamacılar internette yeni ve gelişmiş versiyonları olan bu yazılımları kullanabilirler.

Kümeleme sonuçlarının yöntem bazında geçerliliği, dahili indisler: Dunn (Halkidi,2001), Davies-Bouldin (Halkidi,2001), Silhoutte (Caat,2005), Calinski-Harabasz (Calinski v.d.,1974), R-Kareler (Sharma,1996), Bölünme Katsayısı (Windham,1981) ve Sınıflandırma Entropisi (Windham,1981) yardımı ile tespit edilmiştir. Dunn, Silhoutte, Calinski-Harabasz (CH), Bölünme Katsayısı (PC) indislerinin en büyük, Davies-Bouldin (DB), R-Kareler (RS), Sınıflandırma Entropisi (CE) indislerinin en küçük değeri aldığı küme sayısı optimaldir**.** Bu indislerden Dunn, DB, Silhoutte, CH indisleri k-Ortalamalar, MEB ve SOM metodlarının, RS indisi H/C yönteminin, PC ve CE indisleri ise BCO yönteminin geçerliliğini tespit etmede başarılıdır.

Farklı yöntemler kullanılarak elde edilen kümeleme sonuçlarının birbiriyle ne ölçüde uyuştuğu harici indisler (Rand (Halkidi,2001), Folkes-Mallows (Halkidi,2001), Jaccard (Halkidi,2001)) yardımı ile tespit edilmiştir. Yine harici indisler kullanılarak döküm verisindeki y2 ve y3 kalite çıktı değerleri hatalı ve hatasız olarak kodlanarak kümeleme sonuçlarımızın bunlarla alakası bulunmaya çalışılmıştır (bkz. 3.2.3.1). Tüm indis uygulamaları için yukarıda bahsedilen CVAP v. 3.42_7 yazılımı kullanılmıştır.

2.2.2.2.1 k-Ortalamalar

Toplam n nesneyi (veri satırını), k kümeye bölmek için, her nesnenin rastgele seçilen küme merkezlerine olan Öklit uzaklıklarını hesaplar ve en yakın uzaklığa göre kümelemeyi yapar. İlk adımda bu şekilde yaptığı kümelemeyi, sonraki adımlarda her kümedeki noktaların ortalamasını alarak güncellediği küme merkezleriyle tekrarlamaktadır. Her yinelemede noktaların hangi küme merkezine daha yakın olduğunu bulmak için bu uzaklıkların karesini en aza indirerek, optimum küme merkezlerini bulmaya çalışır (MacQueen,1967).

Tablo 2-2 k=2-4 için k-ortalamalar ile döküm verisinde bulunan kümeler ve küme merkezlerinin birbirlerine uzaklıkları (kümelerin benzemezlik değerleri)

k=2	*küme_1* (**70** Nesne) – *küme_2* (**22** Nesne)	1.113769
k=3	*küme_1* (**68** Nesne) – *küme_2* (**22** Nesne)	1.111567
	küme_1 (**68** Nesne) – *küme_3* (**2** Nesne)	1.593595
	küme_2 (**22** Nesne) – *küme_3* (**2** Nesne)	1.968277
k=4	*küme_1* (**68** Nesne) – *küme_2* (**6** Nesne)	1.44533
	küme_1 (**68** Nesne) – *küme_3* (**2** Nesne)	1.593595
	küme_1 (**68** Nesne) – *küme_4* (**16** Nesne)	1.104353
	küme_2 (**6** Nesne) – *küme_3* (**2** Nesne)	2.197992
	küme_2 (**6** Nesne) – *küme_4* (**16** Nesne)	1.055844
	küme_3 (**2** Nesne) – *küme_4* (**16** Nesne)	1.95292

Döküm veri kümesinde k-Ortalamalar ile k=2 için bulunan 70 ve 22 nesneli kümeler birbirinden ayrık ve nesne sayıları itibari ile kompakt'tır. K=3 ve k=4 için olan kümelemelerde 2 nesneli (yani aslında küme sayılamayacak) kümeler de en doğal kümelemenin k=2 için olduğunu göstermektedir. K-Means ile bulduğumuz kümelerden 70 nesneli kümeye girdi ve çıktı değişkenleri arasındaki ilişkiyi modellemek üzere yaygın olarak tercih edilen, karar ağacı metotlarından CART (SPSS Clementine® 10.1) uygulanmıştır. Tüm veri kümesi için hata sebeplerinin CART ile modellenmesi ile elde edilen sonuçları 70 nesneli kümede bulduğumuz sonuçlar ile karşılaştırdığımızda, 70 nesneli kümeden elde edilen sonuçların öğrenme ve kararlılık açısından tüm veri kümesinden elde edilen sonuçlardan daha iyi olduğu bulunmuştur.

k-Ortalamalar ile bulduğumuz kümeleme sonuçları Dunn, Davies-Bouldin, Silhoutte ve Calinski-Harabasz indisleri ile geçerlilik testinden geçirilerek bu yöntemin döküm verisinde bulduğu küme sayısının 2 olduğuna karar verilmiştir.

2.2.2.2.2 Medoidler Etrafında Bölümleme (MEB)

Bu yöntem veri kümesinde k medoid bularak, bu k medoide olan uzaklıklarına göre toplamdaki n nesneyi kümelemektedir. Medoid, bir kümedeki tüm nesnelere olan ortalama uzaklığı (benzemezlik ölçüsü) en küçük olan küme elemanıdır. MEB, medoidleri hesaplarken uzaklık matrisini kullanmaktadır. Uzaklık matrisi veri kümesindeki tüm nesnelerin birbirlerine olan Öklit uzaklığını gösterir (Kauffman v.d.,1990).

Tablo 2-3 k=2-4 için MEB ile döküm verisinde bulunan kümeler ve küme medoidlerinin birbirlerine olan uzaklıkları (kümelerin benzemezlik değerleri)

2 küme	*küme_1* (**40** Nesne) – *küme_2* (**52** Nesne)	1.2838
3 küme	*küme_1* (**33** Nesne) – *küme_2* (**34** Nesne)	1.2838
	küme_1 (**33** Nesne) – *küme_3* (**25** Nesne)	1.2729
	küme_2 (**34** Nesne) – *küme_3* (**25** Nesne)	1.1242
4 küme	*küme_1* (**20** Nesne) – *küme_2* (**34** Nesne)	1.2838
	küme_1 (**20** Nesne) – *küme_3* (**25** Nesne)	1.2729
	küme_1 (**20** Nesne) – *küme_4* (**13** Nesne)	1.1374
	küme_2 (**34** Nesne) – *küme_3* (**25** Nesne)	1.1242
	küme_2 (**34** Nesne) – *küme_4* (**13** Nesne)	1.5336
	küme_3 (**25** Nesne) – *küme_4* (**13** Nesne)	1.5523

Bu tablodan döküm verisinde 4 gruplu (k=4) kümelemede benzemezliğin, 2 ve 3 kümeli ayrıştırmalara göre daha fazla olduğu görülmektedir. 4 gruplu kümeleme ile bulunan 20, 34, 25 ve 13 tane nesne içeren kümelerin nesne sayısı da oldukça uygundur. Bu durum, MEB ile yaptığımız çalışmada, 4 gruplu kümelemenin tercih edilmesine yol açmaktadır.

MEB ile bulduğumuz kümeleme sonuçları Dunn, Davies-Bouldin, Silhoutte ve Calinski-Harabasz indisleri ile geçerlilik testinden geçirildiğinde indisler MEB yönteminin k=2 için yaptığı kümelemenin k=4'e göre daha iyi olduğunu söylemektedir. Kümeleme çalışmalarında sonuçlar esnek olarak değerlendirilebileceğinden her iki sonuç da doğru kabul edilebilir.

2.2.2.2.3 Değiştirilmiş k-Ortalamalar (pürüzlü optimizasyon ile)

K-ortalamalar yöntemi, bir veri kümesindeki anlamlı küme sayısını bilmediğimiz/tahmin edemediğimiz durumlarda kümeleme problemini çözmede yeterli olamamaktadır. Değiştirilmiş k-Ortalamalar yöntemi kümeleme işlemine başlarken k-ortalamalar yönteminde rastgele seçilen küme merkezlerini, rastegele seçmek yerine bir pürüzlü optimizasyon alt problemini çözerek tespit etmektedir (Bagirov v.d.,2003).

Tablo 2-4 k=2-4 için Pürüzlü optimizasyon ile değiştirilmiş k-ortalamalar ile bulunan kümeler

k=2	k=3	k=4
küme_1: 61 nesne küme_2: 31 nesne	küme_1: 59 nesne küme_2: 31 nesne küme_3: 2 nesne	küme_1: 45 nesne küme_2: 24 nesne küme_3: 2 nesne küme_4: 21 nesne

k=4 durumunda, k-ortalamaların iki grubu 10'dan az üye ile kurulmuşken, değiştirilmiş k-Ortalamalar ile tek bir küme dışındakilerin hepsi 20'nin üzerinde üyeye sahiptir. Değiştirilmiş k-Ortalamalar yönteminin k-Ortalamalar'a göre kümeyi daha dengeli böldüğünü görüyoruz, bu durum özellikle k değeri arttıkça ortaya çıkmaktadır.

2.2.2.2.4 Kendi Kendini Düzenleyen Haritalar (SOM)

Yapay sinir ağlarının özel bir türü olan ve ilk olarak Teuvo Kohonen tarafından ortaya konan kendi kendini düzenleyen haritalar (SOM), çok boyutlu veri kümeleri için etkin bir görselleştirme ve kümeleme yöntemidir. SOM, çok boyutlu veri kümesindeki karmaşık ilişkileri, düşük boyutlu topolojik uzayda (genellikle 1 veya 2 boyutlu) basit geometrik ilişkilere dönüştürerek görselleştirir (Kohonen,1995).

Tablo 2-5 k=3-4 için SOM ile bulunan kümeler ve küme merkezlerinin birbirlerine uzaklıkları

k=3	*küme_1* (**50** gözlem) – *küme_2* (**18** gözlem)	0.4631
	küme_1 (**50** gözlem) – *küme_3* (**24** gözlem)	0.7015
	küme_2 (**18** gözlem) – *küme_3* (**24** gözlem)	0.5161
k=4	*küme_1* (**22** gözlem) – *küme_2* (**12** gözlem)	0.4546
	küme_1 (**22** gözlem) – *küme_3* (**23** gözlem)	0.4144
	küme_1 (**22** gözlem) – *küme_4* (**35** gözlem)	0.8943
	küme_2 (**12** gözlem) – *küme_3* (**23** gözlem)	0.4319
	küme_2 (**12** gözlem) – *küme_4* (**35** gözlem)	0.7862
	küme_3 (**23** gözlem) – *küme_4* (**35** gözlem)	0.6027

SOM sonuçlarını Dunn ve Davies-Bouldin indisleri ile test ettiğimizde her iki indeks de en doğru küme sayısını 3 olarak belirlemiştir. SOM yöntemi k=2 için kümeleme yapmamaktadır.

2.2.2.2.5 Bulanık c-Ortalamalar (BCO)

Bir nesnenin birden fazla kümeye ait olabileceği şekilde veriyi kümeleyen kümeleme yöntemlerine bulanık (fuzzy) kümeleme denmektedir. Bulanık kümeleme yöntemlerinde, herbir nesnenin, herbir küme için aitlik değeri (üyelik derecesi) vardır. Bu üyelik değeri 0 ile 1 arasında sonsuz değer alabilmektedir (Dunn,1973).

Bulanık olmayan kümeleme yöntemleri k-Ortalamar ve MEB ile döküm verisinde 2 küme olduğu fikri oluşmuştur. Verinin yapısını daha iyi anlamak ve bulanık kümelemenin 2 kümeden farklı bir yapı bulup bulamayacağını görmek amacıyla Bulanık c-Ortalamalar döküm verisine uygulanmıştır. Bu yönteme uygun olan Bölümleme Katsayısı (PC) ve Sınıflandırma Entropisi (CE) adlı indisler ile sonuçlar geçerlilik testinden geçirildiğinde bulanık anlamda da veri de 2 homojen grup bulunmaktadır. PC'nin en büyük, CE'nin en küçük değeri aldığı küme sayısı optimaldir.

2.2.2.2.6 Aşamalı Tam Bağlantı Yöntemi (H/C)

Bu yöntemde öncelikle nesneler arasındaki uzaklıklar hesaplanır. İki küme arasındaki uzaklık, bu iki kümede bulunan birbirinden en uzak iki nesnenin uzaklığı olarak tanımlanır. Daha sonra oransal uzaklıklar dendrogram adı verilen ağaç grafiği üzerinde gösterilir. Dendrogram yardımıyla birbirine yakın nesneler birbirlerine yakınlık oranları bakımından guruplanırlar, gibi ilk adımda tüm nesneler tek kümededir (Ward, 1963).

Döküm verisini yukarıda anlatılan bulanık ve bulanık olmayan bölümleme yöntemleri ile incelemiştik. Veri kümesinin yapısını daha iyi anlamak ve kümeleme yöntemlerinin döküm verisinde birbiriyle ne oranda ölçüştüğünü saptamak istediğimizden bölümleme yöntemlerinden farklı bir yöntem ile kümeleme yapan bir aşamalı bağlantı yöntemi olan tam bağlantı yöntemi uygulanmıştır. Bu uygulama sonuçları R-kareler indisi ile test edildiğinde en iyi küme sayısı yine 2 olarak görünmektedir. R-kareler indisinin en küçük değeri aldığı küme sayısı optimaldir.

Tablo 2-6 k=2-4 için yöntemlerin dahili indis değerleri

Measures	Models	k=2	k=3	k=4
Dunn	MEB	**1.5288**	1.1644	1.0982
	k-Ortalamalar	**1.4842**	1.1754	0.92415
	SOM	**-**	1.1599	0.66589
DB	MEB	**1.2384**	1.6126	1.348
	k-Ortalamalar	**1.3388**	1.6126	1.6194
	SOM	**-**	1.2464	1.7362
S	MEB	**0.28837**	0.21172	0.16846
	k-Ortalamalar	**0.27461**	0.21112	0.17416
	SOM	**-**	0.18334	0.10311
CH	MEB	**35.9767**	31.1511	23.4789
	k-Ortalamalar	**39.9764**	31.649	26.3452
	SOM	**-**	30.5739	19.6063
PC	BCO	**0.6792**	0.5630	0.5312
CE		**0.4932**	0.7634	0.8801
RS	HC	**0.2686**	0.3509	0.4235

Dunn: Dunn index
DB: Davies-Bouldin index
S: Silhoutte index
CH: Calinski-Harabasz index
PC: Bölünme Katsayısı
CE: Sınıflandırma Entropisi
RS: R-Kareler
BCO: Bulanık C-Ortalamalar

2.2.2.2.7 Sonuç

Sonuç olarak, yürüttüğümüz kümeleme çalışmaları ile döküm veri kümesinin yapısını homojen grupları bularak anlamak, kümeleme yöntemlerinin veride bulduğu farklı kümelerin ne oranda ölçüştüğünü görmek ve bu yöntemlerin en iyi küme sayısı açısından birbirini destekleyip desteklemediğini saptamak mümkün olmuştur. Bu çalışmada indisler kullanılarak döküm verisine uygulanan bütün kümeleme yöntemlerinin en iyi küme sayısı olarak 2 bulduğunu görülmüştür. Bütün kümeleme yöntemlerinin bulduğu kümeler nesnelerin küme aidiyetleri anlamında birbirinden farklıdır. Bu kümelerin birbiriyle ne ölçüde örtüştüğü Kısım 3.2.3.1'de sunulmuştur. Bu kısımda anlatılan çalışmalarımız (Akteke-Ozturk v.d., 2007) bildirisinde sunulmuştur.

2.2.2.3 Birliktelik Analizi

Birliktelik analizleri, veri kümesindeki değişkenler arasında ilginç ilişkileri ve birliktelikleri ortaya çıkarmakta kullanılan bir veri madenciliği yöntemidir. Birliktelik kuralları, belirli bir sonucu bir koşul kümesi ile ilişkilendirir. Bu yaklaşımın kullanıldığı en tipik örnek "alışveriş sepeti analizi"dir. Bu analizde, müşterilerin alışveriş alışkanlıkları, satın aldığı ürünler arasında ilişkiler kurularak (birlikteliklere bakarak) belirlenmeye çalışılmaktadır. Örneğin, "A ürünün satın alınması, B, C ve D ürünlerinin alınması ile ilişkilidir" şeklindeki bir kural, alışveriş sepeti analizi ile elde edilebilir. Daha genel bir ifade ile, birliktelik kuralları aşağıdaki formda sonuçlar üretir:

EĞER *koşul* İSE *sonuç*

Burada, koşul ve sonuç kümeleri değişken-değer ikililerinin kesişmeyen birleşimleridir. Birliktelik kurallarını değerlendirmede kullanılan en yaygın iki ölçüt destek (support) ve güven (confidence) değerleridir. "X İSE Y" şeklinde ifade edilen bir kural için destek değeri X ve Y nin birlikte gerçekleşmesi olasılığı olan Pr(X ve Y) değeri, güven ise Y'nin X gerçekleşmişken koşullu olasılığı olan Pr(Y|X) değeridir. Buradaki olasılıklar veri kümesinde gözlenen frekanslara göre hesaplanmaktadır. X ve Y'nin birlikteliğinin önemli olması için hem destek, hem de güven kriterinin olabildiğince yüksek olması gerekmektedir. Bu değerler için alt sınırlar kullanıcı tarafından belirlenebilmektedir. En çok bilinen ve kullanılan birliktelik analizi algoritması Apriori

algoritmasıdır (Agrawal & Srikant, 1994). Bu algoritmada temel mantık, tüm sık geçen öğelerin bulunması ve bu sık geçen öğelerden güçlü birliktelik kurallarının üretilmesidir. Sık geçen öğelerin bulunmasında veri tabanı bir çok kez tarandığı için, algoritmaların performansını belirleyen asıl adım budur.

Bu projede, birliktelik analizleri elektronik sektöründe karşılaşılan problemlere uygulanmıştır. Elektronik sektöründe faaliyet gösteren firmalar için, hata sebeplerinin üretim sürecinin başlarında belirlenerek giderilmesi maliyet açısından çok önemlidir. Üretim sürecinin başında birkaç YTL tutarında olan bir baskılı devre kartı (printed circuit board – PCB), üretim sürecinin sonunda üzerine monte edilen parçalarla birlikte ele alındığında onlarca YTL tutarında bir ara ürüne dönüşmektedir. VESTEL firması yetkilileri ile yapılan görüşme sonucunda, özellikle üretim hattının sonunda belirlenen bir hatanın giderilmesinin çok zor ve bazı durumlarda imkansız olduğu ve bu kartların atılarak firmaya önemli bir maliyete sebep olduğu öğrenilmiştir.

Yapılan birliktelik analizlerinde amaç PCB kartlarında tesbit edilen arıza açıklamaları ile bu arızaların sebepleri arasında olası ilişkilerin aranmasıdır. Ayrıca, arıza açıklaması ve arıza sebebi değişkenlerinin tedarikçiler ile ilişkilendirilmesi de amaçlanmıştır. Analizlerde, Apriori, GRI (Generalized Rule Induction) ve Carma olmak üzere üç farklı algoritma kullanılmıştır (Clementine® 11.0 Algorithms Guide, 2007). Özetle, cevabı aranan sorular aşağıdaki gibi olmuştur:

- Belirli arızalar, belirli bir grup arıza sebebinden mi kaynaklanmaktadır?

  Örneğin,

  EĞER Arıza sebebi=Dik İSE Arıza açıklaması=Ses hatası

- Arızalı ürünlerin önemli bir kısmı belirli bir üretim hattından mı gelmektedir? Belirli ürün türlerinde daha çok mu arıza gözlenmektedir? Belirli kart tiplerinde daha çok mu arıza gözlenmektedir?

  Örneğin,

  EĞER Ürün türü=1 VE Kart tipi=5 VE Üretim bandı=2 İSE Arıza sebebi=Kırık

- Belirli arızalar belirli tedarikçilerden alınan kartlarda daha mı sık gözlenmektedir?

  Örneğin,

  EĞER “Tedarikçi3” İSE “Besleme hatası”

- Belirli arızalar belirli ürün türünde daha mı sık gözlenmektedir?

  Örneğin,

  EĞER Ürün türü=1 İSE Arıza açıklaması=Görüntü bozuk

Yapılan çalışmalar sonucu bir takım ilişkiler ortaya koyan kural kümeleri elde edilmiştir. Örneğin,

1. *Tamir hattından alınan 28778 kayıttan **2248’ine software flash boş teşhisi** konulmuş ve bunların **% 95’i cihaz çalışmıyor açıklaması** ile tamir hattına gelmiştir.*
2. *Tamir hattından alınan 28778 kayıttan **135’i** hem **9 nolu ürün türü** imiş hem de **birinci banttan tamire ayrılmış** ve bunların **% 100’ünün arıza sebebine EKSİK** teşhisi konulmuştur.*
3. *Tamir hattından alınan 28778 kayıttan **139’u** hem **görüntü yok açıklaması** ile tamir hattına gelmiş hem **7 nolu ürün türü** imiş hem de **9 nolu kart tipi** imiş ve bunların **% 66,9’unun arıza sebebine KISA DEVRE** teşhisi konulmuştur.*

Elde edilen kurallar kümesi firma yetkilileri ile incelendiğinde, kuralların ortaya koyduğu bazı sonuçların firmanın deneyimlediği bir takım üretim sorunlarına karşılık geldiği anlaşılmıştır. Daha açık bir ifade ile, firma yetkilileri, bir dönem ürünlerde belli bir hata saptamaya başladıklarını ancak hata sebebinin tam olarak anlaşılıp gerekli önlemlerin alınması için geçen süre boyunca bir çok hatalı ürün üretmek durumunda kaldıklarını ifade etmişlerdir. Bu tip durumlarda yönlendirici kural kümelerinin otomatik olarak oluşturulabileceği bir sistemin, problem çözme sürecini kısaltacağını, dolayısı ile, hata maliyetlerini düşürebileceğini söylemişlerdir. Ayrıca, kuralların ortaya koyduğu en olası arıza sebeplerinin tamir hattında geçen sürenin kısaltılması ve verimliliğin artırılması amacı ile de kullanılabileceği önerisi sunulmuştur. Örneğin, 1 numaralı kuralda ortaya çıkarılan bilgiye göre, tamir hattına *cihaz çalışmıyor* açıklaması ile gelen bir üründe ilk önce *software flash boş* arıza sebebine bakılması faydalı olacaktır.

Analizler sonucunda ayrıca, belli bir tedarikçi firmanın üretimde kullanılan ürün sayısı diğer tedarikçi firmalara oranla az olmasına karşın, üretilen bir çok kuralda yer alması dikkat çekmiştir. Bu sebeple, bu tedarikçi firmanın hata verileri bireysel olarak analiz edilerek bir takım kurallara ulaşılmıştır. Örnegin,

1. *Tamir hattından alınan **Tedarikçi5'e** ait 969 kayıttan **459'u software hatası açıklaması** ile tamir hattına gelmiş ve bunların **% 96,95'inin arıza sebebine SW teşhisi** konulmuştur.*
2. *Tamir hattından alınan **Tedarikçi5'e** ait 969 kayıttan **249'u görüntü yok açıklaması** ile tamir hattına gelmiş ve bunların **% 63,45'inin arıza sebebine KISA DEVRE teşhisi** konulmuştur.*

Firmaya sunulan kurallardan özellikle örnekte verilen 1. numaralı kural dikkat çekmiştir. VESTEL firması bu bilginin çok faydalı olduğunu dile getirmiş ve bu konuda daha detaylı incelemeler yapacaklarını ve tedarikçi firma ile bu sorunu paylaşacaklarını söylemişlerdir.

Birliktelik analizi çalışmalarında yaşanan en önemli sorun algoritmaların ürettiği kural sayısının çok fazla olması olmuştur. 300'ün üstünde bulunan bu birliktelik kurallarından bir kısmının destek ve güvenilirlik gibi bir takım ölçütlerle elenmesi yoluna gidilmiştir. Ancak, bu değerleri çok eleme yapacak seviyelerde tutmak bazı önemli kuralların kaybedilmesi riskini de beraberinde getirmektedir. Bu sebeple, kurallar önce yaklaşık 3'te iki oranında azaltlmış geri kalanları ise firma ile birlikte değerlendirilmiştir.

Sonuç olarak, birliktelik analizlerinin elektronik devre kartı üreten VESTEL firmasında olduğu gibi düzenli ve ürün bazlı veri toplayan sektörlerde bir çok aşamada faydalı olacağı düşünülmektedir. Veri toplama sisteminin geliştirilmesi elde edilecek faydayı artıracaktır. Üretim hattına entegre edilmiş etkili bir veri toplama sistemi, bu verileri otomatik olarak analiz edecek ve yeni veriler eklenmesi ile kural kümelerini güncelleyecek birliktelik analizi algoritmaları, üretimde kaliteyi artırıcı faaliyetleri olumlu yönde etkileyecek, sebep sonuç ilişkilerinin daha hızlı ortaya konmasını ve çözümlerin hızlı bir şekilde uygulanmasını, dolayısı ile de maliyetlerin düşürülmesi ve verimliliğin artırılmasını sağlayacaktır.

2.2.2.4 Tahmin etme

Döküm verisi üzerinde y2 tipi hata yüzdesini tahmin etme modelleri altı farklı yaklaşım kullanılarak kurulmuştur. Bunlar, karar ağaçları, yapay sinir ağları, MARS, çoklu doğrusal regresyon, robust regresyon ve bulanık regresyondur. Her yaklaşımın kendi içinde mevcut farklı algoritmaları sınanarak, Kısımda 2.2.2.1'de açıklanan çapraz doğrulama düzenine göre en iyi modellere ulaşılmıştır.

2.2.2.4.1 Karar ağaçları

Karar ağaçları; ilk düğüm (root node), iç düğümler (internal nodes), bağlantılar (arcs) ve son düğümler (leaf nodes)'den oluşan ağaç benzeri basit yapılar ile temsil edilen modellerdir (Russell & Norvig, 2003). İlk düğüm tüm veri kümesini içeren düğümdür. İç düğümler bir girdi değişkeninin aldığı eşik değerlerine göre kayıtların gruplandığı düğümlerdir. Son düğümler problemin tipine göre (sınıflandırma veya tahmin etme) çıktı değişkenin tahmin edilen sınıflarını (kategorik çıktı) veya tahmin değerlerini (sürekli çıktı) gösterirler. Bağlantılar ise girdi değişkenleri üzerinde yapılan test sonuçlarını temsil ederler. Karar ağaçlarının anlaşılması ve yorumlanması basit olup kural setlerine dönüştürülmeleri kolaydır. Bunun yanı sıra, karar ağaçları birçok istatistik yöntemde olduğu gibi veri hakkında klasik varsayımlara ihtiyaç duymazlar. Bu nedenlerden ötürü karar ağaçları birçok tahmin ve sınıflandırma probleminde yaygın olarak kullanılmaktadır.

Bu bölümde, karar ağaçları tahmin etme amaçlı kullanılmıştır. İki farklı karar ağacı algoritması denenmiştir. Bunlardan birincisi CART - Classification and Regression Tree algoritmasıdır. CART en çok kabul gören ve kullanılan karar ağacı yöntemlerinden birisidir (Breiman, Friedman, Olshen, and Stone, 1984). Algoritma veriyi ilk düğümde ve her iç düğümde bir önceki düğümden daha homojen (çıktı değişkenlerinin benzer değerler aldığı) düğümler elde edecek şekilde ikiye böler. Homojenlik en küçük kareler sapması ölçütü ile değerlendirilir. Ayırma işlemi homojenlik kriteri sağlanıncaya veya başka bir durdurma kuralına (zaman, düğümdeki minimum kayıt sayısı, vb.) ulaşılıncaya kadar devam eder. İkinci olarak, Kass, (1980) tarafından geliştirilen CHAID - Chi-squared Automatic Interaction Detection algoritması kullanılmıştır. CHAID algoritması, veriyi bağımlı değişkendeki varyasyon gruplar içinde minimum gruplar arasında ise maksimum olacak şekilde farklı alt gruplara ayırmayı hedefler. Gruplamada en etkili süreç değişkenlerinin belirlenmesinde, bağımlı değişkenin sürekli olması halinde F-testi kullanılır. CART algoritmasının aksine CHAID düğümleri ikiden fazla alt gruba bölebilmektedir.

Karar ağacı modeli belirlenirken, öncelikle CART ve CHAID algoritmaları için en iyi modelleri üreten parametreler belirlenmiştir. Bunlar, katışıklık (impurity) ölçüsü, ağaç derinliği, budama kuralı, değişkenleri bölme ve birleştirmede kullanılan alfa (α) parametreleridir. İlgili parametrelerin farklı değerlerinde elde edilen modeller ortalama mutlak hata (MAE) ve korelasyon (r) ölçüleri ışığında doğrulama verisinde test edilip en iyi modellere ulaşılmıştır. İki algoritmanın belirlenen en iyi modelleri, 3 tekrarla elde edilen 3-katlı çapraz doğrulama test veri setlerinde sınanmış ve ortalamada CART modelinin daha iyi sonuçlar verdiği gözlenmiştir.

2.2.2.4.2 Yapay sinir ağları

Yapay sinir ağları, ağırlıklandırma ile birbirlerine bağlanmış bir çok işlem elemanından (nöron) oluşan matematiksel sistemlerdir (Haykin, 1994). Her bir işlem elemanı bir transfer fonksiyonunu ifade eder, diğer nöronlardan sinyaller alır, bunları birleştirir, dönüştürür ve sayısal bir sonuç ortaya çıkartır. Ağın ilk bölümünde girdi katmanı ve bu katmanda her biri farklı bağımsız değişkenleri temsil eden girdi düğümleri mevcuttur. İkinci bölümde gizli katman(lar) ve gizli düğümler, son bölümde ise çıktı katmanı ve çıktı düğümleri bulunmaktadır. Girdi düğümleri ağırlıklı bağlantılarla gizli düğümlere, gizli düğümler ise yine ağırlıklı bağlantılarla çıktı düğümlerine bağlanırlar. Çıktı katmanı bir veya daha fazla çıktı değişkenini temsil edebilirler. Yapay sinir ağları hem tahmin etme hem de sınıflandırma amaçlı kullanılabilir. İleri besleme geri yayılım ağları (feed-forward backpropagation networks – FFBN), radyal temelli fonksiyon ağları (radial-based function networks) ve Kohonen kendi düzenlenen haritalar (self-organizing maps) literatürde var olan bazı sinir ağları algoritmalarıdır. Geri yayılım, ağın gizli katmanlarına bağlı olan ağırlıklarını değiştirmek için kullanılan denetimli bir öğrenmedir. Bir gözlem için ileri besleme aşaması tamamlandığında, tahmin değeri ile gerçek değer arasındaki fark kullanılarak bir hata değeri hesaplanır. Daha sonra geri besleme ile bu hata değeri ağ yapısına gönderilir ve benzer bir örnek geldiğinde daha az hata oluşturacak şekilde ağırlıklar değiştirilir.

Bu bölümde, yapay sinir ağları tahmin etme amaçlı kullanılmıştır. Quick, Dynamic, Prune ve RBFN olmak üzere dört farklı teknik kullanılmıştır. Kullanılan tekniklerin hepsi ileri besleme geri yayılım algoritmasına dayalıdır. Quick tekniği, ağ topolojisi için genel kurallardan ve verinin özelliklerinden faydalanır. Bu yöntemde ağ parametreleri için literatürde kabul görmüş genel parametre değerleri kullanılır ve veri üzerinde bir defa öğrenme gerçekleştirilerek sinir ağı modeline ulaşılır. Dynamic yöntemi, istenilen doğruluğa erişinceye kadar, performansı iyileştirmek için network topolojisini değiştirerek öğrenme gerçeklestirir. Öncelikle bir başlangıç topolojisi oluşturur, öğrenme süresince gizli düğümler ekleyip çıkararak son modelini oluşturur. Prune yöntemi, kavramsal olarak, dynamik yönteminin tam tersi islev görür. Küçük bir ağ ile başlayıp onu geliştirmek yerine, büyük bir ağ ile başlar ve gereksiz sinir hücrelerini girdi ve gizli katmanlarından atarak asamalı olarak budar. RBFN yöntemi ise çıktı değerlerine göre veriyi bölmek için k-ortalama yöntemine benzer bir yöntem kullanan özel bir yapay sinir ağıdır.

Yapay sinir ağı modeli belirlenirken, öncelikle Quick, Dynamic, Prune ve RBFN algoritmaları için en iyi modelleri üreten parametreler belirlenmistir. Bunlar, gizli katman ve nöron sayısı, öğrenme sürecinde hiçbir değisim görünmediği durumda öğrenimin devam etme çevrim sayısını ifade eden devam etme parametresi (persistence), ağırlık değişimlerinin momentumu, ağırlıkların değisim miktarını kontrol eden öğrenme oranıdır. İlgili parametrelerin farklı değerlerinde elde edilen modeller ortalama mutlak hata (MAE) ve tahmini doğruluk değeri (estimated accuracy) ölçüleri ısığında doğrulama verisinde test edilip en iyi modellere ulaşılmıstır. İleri besleme geri yayılım algoritmasına dayalı bu dört teknikle elde edilen en iyi modeller, 3 tekrarla elde edilen 3-katlı çapraz doğrulama test veri setlerinde sınanmıs ve ortalamada Prune tekniği ile elde edilen modelinin daha iyi sonuçlar verdiği gözlenmistir.

2.2.2.4.3 Çoklu doğrusal regresyon

Çoklu doğrusal regresyon analizi bir bağımlı değişken (y) ile birçok bağımsız değişken x'ler arasında doğrusal bir ilişkinin olduğunu varsayarak istatistiksel olarak önemli bir model geliştirmeyi hedefler (Montgomery ve diğ., 2006; Kutner ve diğ., 2004). Çoklu doğrusal regresyon yönteminde bağımlı değişken sürekli değer alırken bağımsız değişkenler sürekli veya kesikli tipte olabilir. Bağımlı değişken y ile x'ler arasındaki ilişki

$$y = x'\beta + \varepsilon = f(x, \beta) + \varepsilon$$

olarak ifade edilir. Hata terimlerinin bağımsız ve sıfır ortalamalı, sabit varyanslı aynı normal dağılıştan geldiği varsayılmaktadır. Yani, ε ~ iid $N(0, \sigma^2)$. Bu nedenle $E(y) = f(x, \beta)$ olup, $f(x,\beta)$ parametre vektörü β'nın doğrusal bir fonksiyonudur. Eğer ε hata terimine ilişkin varsayımlar gerçekleşirse $\hat{\beta}$'lar en iyi, doğrusal ve yansız tahminleyicilerdir.

Çoklu regresyon uygulamasında veri kümesinin küçük olması ve çok fazla bağımsız değiskenin olması bir problem olmuştur. Değişken sayısının fazla olması ikili etkileşimli değişkenlerin analize dahil edilmesini engellemiştir. Modeller, Belli Model (Enter), Adım Adım Yöntemi (Stepwise), Geriye Doğru Eleme (Backwards) ve İleriye Doğru Seçim (Forwards) olmak üzere dört farklı teknik kullanılarak oluşturulmuş, kurulan modeller MSE, F-testin p-değeri, R2 ve Adj. R2 ölçüleri kullanılarak karşılaştırılmıştır. Sonuç olarak en iyi modelin Geriye Doğru Eleme metoduyla bulunduğu görülmüştür.

2.2.2.4.4 Robust Regresyon

İstatistiksel bir yöntem, istatistiksel model varsayımlarının doğru olmadığı koşullarda bile iyi bir performans gösterebiliyor ise bu yönteme robust veya sağlam denir. Çoklu doğrusal regresyon yaygın kullanılan bir tahmin etme yöntemi olsa da hata terimlerinin normal dağılması varsayımını sağlamak çoğu kez zordur. Kalite verilerinde sıklıkla rastlanan aykırı gözlemler, bu varsayımı doğrulamayı engelleyen en önemli nedenlerdendir. Bu tür aykırı değerlere duyarlı olmayan bazı robust regresyon yöntemleri geliştirilmiştir. Bunlardan en yaygın bilinenleri, En Küçük Medyan Kareler (LMS), En Küçük Budanmış Kareler (LTS), Huber M-regresyon, MM-metodu, En Küçük Mutlak değerler (LAD), ve LOWESS olarak sayılabilir (Avcı,2009). Avcı (2009), yüksek lisans tezi kapsamında bu yöntemleri belli aykırı değer senaryolarına göre karşılaştırmış ve bunun sonucunda Huber-M regresyon, En Küçük Medyan Kareler ve En Küçük Budanmış Kareler metotlarının en iyi başarımı gösterdiğini belirlemiştir.

Robust regresyon yaklaşımlarından Huber M-regresyon (Huber 1981, Birkes ve Dodge 1993) döküm verisine uygulanmıştır. Uygulamada S-PLUS yazılımından yararlanılmıştır. Modellemede, bağımsız değişkenlerin çok sayıda olması çoklu doğrudaşlık (*multi-collinearity*) ve hesaplama zamanında artışa yol açtığı için, bir model seçme prosedürü (adımsal regresyon) uygulanmıştır. Bağımlı değişken "yüzde" olduğu için logit (veya omega) dönüşümü uygulanmıştır. Normallik varsayımı tam olarak sağlanamamaktadır. Aykırı gözlemler belirlenmiştir. Bu durum verinin robust Huber-M metodu için uygun olduğunu göstermektedir.

3-tekrarlı, 3-katlı veri düzenine göre Huber-M metodu ile elde edilen tahmin modelleri, R^2 ve MSE gibi ölçülere göre düşük performans göstermiştir. Bunun olası nedenleri arasında değişkenler arasındaki ilişkinin karmaşık olması ve bağımlı değişkeni daha iyi açıklayabilecek bağımsız değişkenlerin veride bulunmaması düşünülebilir. Huber-M metodu aykırı değerlere karşı sağlam olsa da bu tür durumlara duyarlıdır.

2.2.2.4.5 Bulanık regresyon

Bulanık regresyon çalışmaları temel olarak Bulanık Kural Tabanlı Modeller (FRBM), Bulanık Regresyon Modelleri (FRM) ve Bulanık Fonksiyonlar (FF) olmak üzere üç bölümde incelenebilir.

FRBM ilk olarak Zadeh tarafından geliştirilmiş ve Mamdani tarafından uygulanmıştır. Daha sonra bu modeller Takagi-Sugeno ve Sugeno-Yasukawa tarafından çeşitli şekillerde iyileştirilmiştir. Takagi ve Sugeno (1985) çalışmalarında bulanık çıkarsamaların ve anlamlandırmaların kullanıldığı sistemlerde FRBM yapmayı sağlayacak matematiksel bir araç geliştirmiştir. Takagi ve Sugeno tarafından geliştirilen modellerin güçlü yönü bulanık ifadelerin modellenmesine izin vermesidir. Sözkonusu çıkarsamalar sayesinde, gündelik yaşamda kullanılan dilsel ifadeler geliştirilecek modellere eklenebilmiştir. Böylece doğrusal olmayan ilişkiler tanımlanabilmiş ve sonuçta modellerin uygulama alanı genişletilmiştir.

Diğer yandan Sugeno-Yasukawa (1993) bulanık mantığa dayanan nitel modelleme konusunda çalışmış, nitel modellemeyi bulanık modelleme ve dilsel yaklaşım olarak iki alana ayırmış, bulanık modelin belirlenmesi aşamasında Bulanık c-Ortalamalar (FCM) yönteminin kullanılmasını önermiştir.

FRM ilk olarak Tanaka v.d. (1982) tarafından geliştirilmiş olup, genellikle "Olabilirlikli (possibilistic) Regresyon" olarak isimlendirilmektedir. Temel olarak bu modelleme yöntemi tahminlenen aralıkların gözlemlenen aralıkları içermesi kısıtı altında toplam bulanıklığın en küçüklenmesini amaçlayan bir doğrusal programlama yöntemi ile bulanık regresyon katsayılarının tahminlenmesi prensibine dayanmaktadır. Bulanık regresyon yöntemi daha sonra Pedrycz ve Savic (1991), Sakawa ve Yano (1992), Peter (1994), Ozelkan ve Duckstein (2000) ve Hojati ve diğ. (2005) gibi birçok farklı araştırmacı tarafından tekrar gözden geçirilmiştir. Hojati ve diğ. (2005) girdi değişkenlerin kesin sayı (*crisp*) olması durumunda HBS1; bulanık sayı olması durumunda ise HBS2 olarak isimlendirdikleri tahminlenen bağımlı değişkenler ile gözlemlenen bağımlı değişkenler arasındaki farkın en küçüklenmesini amaçlayan bir yöntem geliştirmişlerdir. Geliştirdikleri bu yöntemi ayrıca Tanaka ve diğ. (1989), Sakawa ve Yano (1992), Peter (1994) ve Ozelkan ve Duckstein (2000)'ın geliştirdikleri yöntemler ile karşılaştırmış ve az sayıda bağımsız değişken için kendi yöntemlerinin başarımının daha iyi olduğunu göstermişlerdir. Bu yaklaşımlar dışında Celmins (1987) ve Diamond (1988) tarafından geliştirilen bulanık en küçük kareler (LS) yöntemine dayanan yaklaşımlar da bulunmaktadır. Bulanık regresyon yöntemi olarak isimlendirilebilecek diğer bir yaklaşım ise Hathaway ve Bezdek (1993) tarafından geliştirilen Bulanık c-Regresyon (FCR) yöntemidir. Bu yöntem verinin kümelere ayrılması ve değişken katsayılarının tahminlenmesi işlemlerinin birarada eşanlı olarak gerçekleştirilmesi temeline dayanmaktadır. Bu yöntemde FCM yöntemi ya da farklı bir kümeleme yöntemi kullanılarak elde edilen üyelik değerleri fonksiyonun tahminlenmesindeki ağırlık değerleri olarak kullanılmaktadır.

Bulanık fonksiyonlar (FF) adı verilen yöntem FRBM ve FRM yöntemlerine seçenek olarak Türkşen (2008) tarafından geliştirilmiştir. Genel olarak bu fonksiyonlar LS yöntemi (FF-MLR) veya destek vektör makineleri

(SVM) (FF-SVM) kullanılarak oluşturulabilmektedir. Bu yöntemde çeşitli geçerlilik indeksleri kullanılarak bulanıklık derecesi (optimum m) ve sınıf sayısı (c) değerleri ve daha sonra da FCM yöntemi kullanılarak sınıf merkezleri belirlenmektedir. Belirlenen sınıf merkezlerinden seçilen girdi uzayına ilişkin sınıf merkezleri kullanılarak, gözlemlerin aidiyetleri belirtilen kümelere ilişkin normalleştirilmiş üyelik dereceleri elde edilir. Bu üyelik dereceleri, girdi değişkenleri ve ihtiyaç duyulan düzeyde eklenebilecek üyelik dönüşümleri ile her bir sınıf için ayrı bir girdi matrisini oluşturur ve FF tahminlenmesinde kullanılır. Çıktı değerinin tahmini ise LS yöntemi ile belirlenen değerin kendisine ait üyelik değeri ile ağırlıklandırılması sonucu elde edilir. Yapılan çalışmalar FF'nin FRBM ile karşılaştırıldığında daha iyi tahminler verdiğini göstermektedir (Türkşen ve Çelikyılmaz 2006). Çelikyılmaz (2008) çalışmasında, standart FCM yönteminin iyileştirilmesi ile yeni bir bulanık sınıflandırma yöntemi (IFCM) elde edilmiştir. Elde edilen yeni sınıflandırma yöntemi, sadece veri kümesinin optimum şekilde sınıflandırılmasına katkıda bulunmakla kalmamış, bulanık fonksiyonlar ile üyelik değerlerinin çıktı değişkeni tahminleme gücünü de arttırmıştır.

Döküm verileri incelendiğinde girdi ve çıktı değişkenlerdeki belirsizliğin bulanıklık olarak yorumlanabileceği belirlenmiştir. Modelleme amacıyla, öncelikle girdi ve çıktı değişkenlerinin her ikisinin de bulanık olduğu durum için HBS2 (Hojati v.d. 2005) yöntemi kullanılmıştır. Ancak bu yöntemin tam olarak kullanılabilmesi için girdi değişken sayısının çok az sayıda olması gerekmektedir. Döküm verileri için girdi değişkenleri ikiye azaltılarak tam model, sekize azaltılarak da kısmi model kurulmuş ve regresyon parametreleri bunlara göre ayrı ayrı tahmin edilmiştir. Tahmin sonuçları her iki durum için de yeterince başarılı olmamıştır.

Döküm verisine ayrıca Türkşen ve Çelikyılmaz (2006) tarafından, FRBM'ye göre üstünlüğü gösterilmiş olan FF yöntemi, en küçük kareler doğrusal regresyon yöntemi yardımı ile uygulanmıştır. Burada da sekiz girdi değişkeni kullanılmıştır. Sonuçlar HBS2 sonuçlarından çok daha başarılı olmuştur. Bunların yanısıra, FF yönteminin iyileştirilmesi ile elde edilen IFCM (Çelikyılmaz 2008) yöntemi de aynı veri kümesine uygulanmak istenmiştir. Ancak bu uygulamada gerek duyulan regresyon modellerinin önerildiği gibi parametrik olması, her iterasyonda uygun modelin seçimi güçlüğünü ortaya çıkarmıştır. Bu nedenle ilgili iterasyonlarda parametrik olmayan MARS modellerinin kullanımı düşünülmüş ve Kısım 3.4'de açıklanan şekilde IFCM yöntemi iyileştirilerek uygulama gerçekleştirilmiştir.

Bu uygulamalara ait sonuçlar bir yüksek lisans tezi kapsamında Kılıç (2009) tarafından yayımlanmıştır.

2.2.2.4.6 MARS

Friedman (1991) tarafından geliştirilen çok değişkenli uyarlanabilir regresyon eğrileri (MARS) algoritması, varsayımlara bağlı olmayan yapısı nedeniyle daha kolay uygulanabilen bir regresyon yöntemidir. Bu yöntem ileri ve geri doğru olmak üzere iki adımdan oluşmaktadır. İleri doğru adımda temel fonksiyonlar ve/ya bu fonksiyonların çarpımları en büyük (yani, en karmaşık) modele ulaşılıncaya dek eklenir. Burada temel fonksiyonlar bağımsız değişkenleri en uygun düğüm noktalarıyla aralıklara bölen parçalı doğrusal regresyon eğrileridir. Bu temel fonksiyonlar ya tek değişkenli ya da çok değişkenli etkileşim terimlerinden oluşmaktadır. İleri doğru adım algoritmasının her aşamasında kullanılabilecek en uygun düğüm noktaları ile temel fonksiyonları belirlemek amacı ile genelleştirilmiş çapraz doğrulama (GCV) adı verilen bir ölçüden yararlanılır. Bu ölçü, hata kareler toplamının, model karmaşıklığının bir göstergesine oranıdır. MARS algoritmasının ilk adımında oluşturulan en büyük modelin yorumlanması ve kullanımı kolay olmadığından dolayı ikinci adımda en büyük model budanarak, yani önemli bağımsız değişkenler ve bu değişkenlerin etkileşimleri belirlenerek, GCV ölçüsü en küçük olan model elde edilir.

Hem tahmin etme hem de sınıflandırma uygulamalarında en iyi MARS modelinin belirlenmesi süreci aynıdır. Salford System yazılımı kullanarak döküm verileri üzerinde yapılan MARS uygulamalarında yazılımın sunduğu maksimum temel fonksiyon ve etkileşim sayısı gibi seçeneklere farklı değerler girerek çok sayıda model elde edilmiştir. Bu modellerin farklı veya aynı GCV değerleri olabilmektedir. En iyi MARS modeli, GCV ölçüsü en düşük olan model olarak belirlenmiştir. Bu model veriye uyumu en iyi, aynı zamanda karmaşıklığı en az olan modeldir. GCV ölçüsü aynı olabilen farklı modeller söz konusu olduğunda ise; R2 ve R2 düz (Adj-R2) değerleri en büyük ve kullanılan temel fonksiyon sayısı en düşük olan model seçilmiştir.

Yukarıda bahsedilen değişik tahmin etme yöntemlerinin 3-tekrarlı, 3-katlı döküm verileri üzerindeki uygulamalarından elde edilen özet sonuçlar EK'de verilmiştir. Bu sonuçların karşılaştırılması ise Kısım 2.2.3.2'de sunulmaktadır.

2.2.2.5 İkili Sınıflandırma

Döküm verisi üzerinde y2 tipi hatanın kabul edilebilir düzeyde olup olmadığını ve koltuktan memnuniyet verisi üzerinde memnuniyet düzeyinin az veya çok olduğunu tahmin etmek üzere altı farklı sınıflandırma yaklaşımı kullanılmıştır. Bunlar, karar ağaçları, yapay sinir ağları, MARS, lojistik regresyon, destek vektör

makinaları, Mahalanobis Taguchi sistemi ve bulanık sınıflandırma fonksiyonlarıdır. Her yaklaşımın kendi içinde mevcut (varsa) farklı algoritmaları sınanarak, 2.2.2.1 Kısımda açıklanan çapraz doğrulama düzenine göre her iki veri için de en iyi modellere ulaşılmıştır.

2.2.2.5.1 Karar ağaçları

Kısım 3.2.2.4'de özetlenen karar ağacı yöntemlerinden CART ve CHAID, bu kısımda sınıflandırma amacı ile kullanılmıştır. Bunlara ek olarak, sadece sınıflandırma yapabilen iki yeni algoritma, C5.0 ve QUEST, kullanılmıştır. C5.0 algoritması, her adımda, en fazla bilgiyi sağlayan değişkene göre gözlemleri alt gruplara bölerek çalışmaktadır. Ticari bir ürün olduğu için algoritma detayları verilmemektedir. QUEST - Quick, Unbiased, Efficient Statistical Tree, diger algoritmalara göre daha yeni olup CART algoritması gibi ağacı ikili bölünmelerle oluşturmaktadır. Algoritma her değişken ile çıktı değişkeni arasında bağımsızlık testi yaparak *p* değerlerini hesaplar. Kesikli değişkenler için Pearson chi-square testi veya Levene's testi, sürekli değişkenler için ise F testi kullanılır. *p* değeri en küçük olan değişken önceden belirlenmiş α değerinden küçükse bölünme için bu değişken kullanılır. Aksi durumda ilgili düğüm bölünmez. Seçilen değişkenin ayırma noktası quadratik ayırma analizi ile belirlenir.

Döküm ve koltuktan memnuniyet sınıflandırma verileri, dört karar ağacı algoritması ile modellenirken, öncelikle, tahmin etme çalısmasında olduğu gibi, dogrulama verileri ile her algoritmanın en iyi modellerini ortaya çıkaran serbest parametreler belirlenmistir. Bu aşamada, doğru sınıflandırma oranı, kesinlik ve duyarlılık değerlerine bakılmıştır. Elde edilen modeller çapraz doğrulama test verilerinde sınanmış ve en başarılı sonuçların C5.0 algoritması ile elde edildiği görülmüştür.

2.2.2.5.2 Yapay sinir ağları

Kısım 3.2.2.4'de özetlenen yapay sinir ağı teknikleri, bu bölümde sınıflandırma amaçlı kullanılmış, teknikler kendi içinde değerlendirilmiş ve en iyi modellere ulaşılmıştır. Tahmin etme modellerinde olduğu gibi, sınıflandırma modellerinde de en iyi sonuçları prune tekniği vermiştir.

2.2.2.5.3 MARS

En iyi MARS sınıflandırma modeli Kısım 2.2.2.4 te açıklandığı gibi belirlenmiştir.

2.2.2.5.4 Lojistik regresyon

Lojistik regresyon, kesikli çıktı değişkeninin modellenmesi için kullanılan bir yaklaşımdır. Çıktı değişkeni ikili veya çoklu olabilir. Lojistik regresyon, bağımsız değişkenlerin çıktı değişkeni üzerindeki etkilerini olasılıkla ifade eder. Model parametrelerinin tahmininde, en çok olabilirlik, en küçük kareler ve minimum logit ki-kare, kullanılabilecek yöntemlerdir. Lojistik fonksiyonu

$$f(z) = \frac{1}{1 + e^{-z}}$$

ile ifade edilir. Burada,

$$z = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \cdots + \beta_k x_k,$$

'dir ve β_0, β_1, ..., β_k model parametrelerini göstermektedir.

Döküm ve koltuktan memnuniyet veri kümesinde lojistik regresyon uygulaması yapmadan önce bağımsız değişkenler arasındaki korelasyona bakılmıstır. Koltuktan memnuniyet verisinde bağımsız değişkenler arasındaki korelasyonların yüksek olması sebebiyle sürekli veriler için temel bileşenler analizi sonucunda elde edilmis faktörlerin kullanılmasına karar verilmis ve daha önce Çabuk (2008) tarafından geliştirilmis modeller üzerinden çalışılmıştır. Bu faktörler uygulama avantajlarından dolayı standardize edilmiştir. Yüksek korelasyon sorunu, kendisini uygun olmayan sınıflandırma oranları, başarısız modeller, standart sapması yüksek parametre katsayıları, bir sonuca yakınsamayan modeller olarak göstermiştir. Ayrıca, küçük örneklem sayısına karşılık çok sayıda değişkenin olması yakınsama problemini ortaya koymuştur. Değişken sayının çok fazla olması sebebi ile 2'li etkileşimli değişkenler modele eklenememiştir. Modeller oluşturulurken, Belli Model (Enter), Adım Adım (Stepwise), İleri Doğru Seçim (Forwards), Geriye Doğru

Eleme (Backwards) ve Geriye Doğru Adım Adım Seçim (Backwards Stepwise) olmak üzere 5 yöntem kullanılmıştır. Oluşturulan modeller Deviance, Classification rate, likelihood ratio testlerinin *p* değerleri, modelde etkili olan değişken sayıları ve bir R2 hesabı olan Cook ve Snell istatistiği kriterlerine bakılarak karşılaştırılmıştır. Genel olarak Belli Model ve Geriye Doğru Eleme yöntemleri en iyi model sonuçlarını vermiştir.

2.2.2.5.5 Destek Vektör Makinaları

Destek Vektör Makinaları (DVM) yöntemi, sınıflandırma ve doğrusal olmayan fonksiyon yaklaşımı problemlerinin çözümü için 1992'de Vapnik tarafından önerilen eğiticili bir öğrenme algoritmasıdır. DVM yöntemi temelde doğrusal olarak ayırt edilebilen iki sınıflı problemlerin çözümü için kullanılmakla beraber, doğrusal olarak ayırt edilemeyen veya çok sınıflı (*multi-class*) problemlerin çözümünde de genelleştirilerek kullanılmaktadır.

DVM yöntemi veri kümesinin noktalarını iki sınıfa ayırmaya çalışan bir ayırıcı (*separator*) bulmaya çalışır. Çok boyutlu uzayda ayırıcı, bir hiperdüzlem *(hyperplane)* olacaktır. Ayırıcıya en yakın noktalar destek vektör *(support vector)* olarak adlandırılmaktadır. Destek vektörler ile ayırıcı arasındaki uzaklık ise *marjin* olarak adlandırılır. Sınıfları birbirinden ayıran birden fazla ayırıcı olabilir. DVM yöntemi ile bu ayırıcılardan optimal olanı bulunmaya çalışılır. Optimal olan marjini en büyük yapan ayırıcıdır (Awad, 2004).

Bizim çalışmamızda DVM yöntemi uygulamaları için MATLAB ile kodlanmış "libsvm v.2.83" adlı yazılım kullanılmıştır. Bu yazılımda sınıf sayısı (2), veriyi çok boyutlu uzaya örtüştüren kernel fonksiyonu parametreleri (2^{-15},2^{-13},2^{-11},2^{-9},2^{-7},2^{-5},2^{-3},2^{-1},2,2^3), tahmin etmede tolare edilecek hatayı sınırlandırmak için hata sabiti (2^{-5},2^{-3},2^{-1},2,2^3,2^5,2^7,2^9,2^{11},2^{13},2^{15}) ve eğitim/test verilerini ayırırken kullanılan bölme (fold) sayısı (5) yazılıma verilmiştir.

2.2.2.5.6 Mahalanobis Taguchi Sistemi (MTS)

MTS çok değişkenli sınıflandırma için geliştirilmiş bir yöntemdir (Taguchi, Chowdhury ve Yuin, 2001). Bu yöntem kalite iyileştirme özelinde, örneğin, çok sayıda süreç değişkeninin üretilen parçanın hatalı veya hatasız çıkmasını belirlediği durumlarda bu değişkenlerin hangilerinin bu sonuca ne kadar etki ettiğini gösteren modeller geliştirmede kullanılabilir.

MTS, Mahalanobis Uzaklığı (MU) ile Taguchi'nin robust tasarım yaklaşımının birleştirilmesi ile oluşturulmuştur. MTS, Mahalanobis uzaklığını, bir gözlemin tanımlanmış bir referans grubuna olan uzaklığını ölçecek bir uzaklık ölçeği olarak kullanmaktadır. Mahalanobis uzaklığı, birden fazla değişkenin olduğu durumlarda, değişkenleri homojen bir grup gibi algılar. Değişkenlerin arasındaki korelasyonu dikkate alarak uzaklığı belirler.
MTS uygulanacak veri, normal ve anormal sınıflar olmak üzere ikiye ayrılır. (Normal sınıf, döküm verileri için ilgili döküm hatası içeren sınıf, müşteri memnuniyeti verileri içinse koltuktan biraz memnun olan sınıf olarak alınmıştır.) Normal sınıf kullanılarak MU belirlenmesinin ardından sinyal gürültü oranları hesaplanarak önemli bağımsız değişkenler belirlenir. Bu değişkenler dikkate alınarak anormal sınıfın normal sınıfa Mahalanobis uzaklığı hesaplanır. Her test veya öğrenme gözlemi için bu uzaklık belli bir sınır değerden büyükse ilgili gözlem anormal sınıfa, değilse normal sınıfa atanır. Bunun için Matlab dilinde geliştirdiğimiz özel bir kod kullanılmıştır.

2.2.2.5.7 Bulanık Sınıflandırma Fonksiyonları

Bulanık sınıflandırma alanında yapılan çalışmalar temel olarak Bulanık Kural Tabanlı Modeller (FRBM) ve Bulanık Sınıflandırma Fonksiyonları (FCF) olmak üzere iki gruba ayrılabilir.

FRBM yöntemleri genel olarak bulanık kümelemeye dayalı yöntemler ile Uyarlamalı Ağ Tabanlı Bulanık Çıkarsama Sistemi'ni (ANFIS) içermektedir. Bulanık kümelemeye dayalı yöntemler için Abe ve Thawonmas (1997) tarafından geliştirilen Bulanık Sınıflandırıcı (FC) ile Setnes ve Babuşka (1999) tarafından geliştirilen Bulanık İlişkisel Sınıflandırıcı (FRC) örnek olarak verilebilir. Abe ve Thawonmas (1997) tarafından geliştirilen FC yönteminde herbir sınıf için kümeler belirlenir ve belirlenen her küme için bulanık kurallar geliştirilir. Setnes ve Babuşka (1999) tarafından önerilen FRC yöntemi ise herhangi bir kümeleme yöntemi ile küme üyelik değerlerinin elde edilmesi ve elde edilen üyelik değerleri ile sınıf üyelik değerleri arasındaki bulanık ilişkinin kurulmasına dayanan bir yöntemdir. Huang ve diğ. (2007) ANFIS'e dayalı bir sınıflandırıcı geliştirmiştir. Geliştirilen bu yöntem karakter tanıma ve ANFIS uygulaması olmak üzere iki aşamadan

oluşmaktadır. Karakter tanıma aşamasında ortogonal bir dizey kullanılarak bağımsız değişkenlerin seçimi gerçekleştirilmektedir. Daha sonra ANFIS kullanılarak bulanık modelleme yapılmaktadır.

FCF, Türkşen (2008) tarafından geliştirilen "çoklu regresyon analizi ile bulanık fonksiyonlarla bulanık sistem modelleme" yaklaşımının sınıflandırma problemine uyarlanması olarak tanımlanabilir. Çelikyılmaz ve diğ. (2007) tarafından geliştirilen bu yöntemde FCM yöntemi ile elde edilen her bir küme için bağımsız değişkenler ile küme üyelik değerleri ve çeşitli dönüşümlerinden oluşan yeni bir girdi matrisi kullanılarak ayrı bir sınıflandırma modeli geliştirilir. Lojistik Regresyon (LR) ya da SVM gibi çeşitli sınıflandırıcıların kullanılabildiği bu yöntemde her bir küme için ayrı ayrı elde edilen sınıflandırma değerleri küme üyelik değerleri ile ağırlıklandırılarak tek bir tahmin değeri elde edilir. Çelikyılmaz ve diğ. (2007) geliştirdikleri bu yöntemi LR, ANFIS, Yapay Sinir Ağları (ANN) ve SVM gibi klasik sınıflandırma yöntemleri ile karşılaştırmış ve FCF'nin daha iyi sonuçlar verdiğini göstermiştir. Çelikyılmaz (2008), FCM yöntemini tekrar gözden geçirerek FF ve FCF yöntemlerinde kullanılan üyelik değerlerinin daha güçlü tahminleyiciler haline getirilmesinin amaçlandığı "İyileştirilmiş FCM" (IFCF) yöntemini önermiştir. Bu yöntemin kümeleme aşamasında üyelik değerleri elde edilirken, yalnızca bu üyelik değerleri ve dönüşümleri kullanılarak, bir sınıflandırıcı yoluyla bağımlı değişken tahmin edilmeye çalışılmaktadır. FCM yönteminde kullanılan amaç fonksiyonuna yalnızca üyelik değerleri ve çeşitli dönüşümleri kullanılarak tahmin edilen bağımlı değişken ve gözlemlenen bağımlı değişken değerinin hata karesini içeren bir terim eklenerek elde edilecek üyelik değerlerinin tahminleme gücü artırılmaya çalışılmaktadır.

Yukarıda belirtilen yöntemlere ek olarak Bulanık Regresyon kısmında açıklanan yöntemler de kesikli bağımlı değişkenin herhangi bir dönüşüm fonksiyonu ile sürekli değişken haline getirilmesinden sonra sınıflandırma amacı ile de kullanılabilmektedir.

Projede döküm ve müşteri memnuniyeti verilerinin bulanık yöntemler kullanılarak modellenmesi yapılmıştır. Türkşen ve Çelikyılmaz (2006) ile Çelikyılmaz ve diğ. (2007) geliştirdikleri yöntemleri FRBM ile karşılaştırmış ve FF'nin daha iyi sonuç verdiğini göstermişlerdir. Ayrıca FRBM'nin üyelik fonksiyonunun belirlenmesi, uygun işlemcinin (c-norm, t-norm vb.) seçilmesi, bulanık sonucun sayısal sonuca dönüştürülmesi (*defuzzification*) gibi zorluklarının olduğuna da değinmişlerdir. Bu nedenle döküm ve müşteri memnuniyeti verilerinin modellenmesi aşamasında FRBM'ye üstünlüğü kanıtlanan FCF ile IFCF'nın kullanılması kararlaştırılmıştır. Çelikyılmaz v.d. (2007) nin önerdiği FCF-LR modeli her iki veriye de uygulanmıştır. Ancak IFCF yaklaşımının kullanımında gerek duyulan sınıflandırma modellerinin önerildiği gibi parametrik olması, her iterasyonda uygun modelin seçimi güçlüğünü ortaya çıkarmaktadır. Bu nedenle IFCF yerine FCF yöntemi, Matlab ve R yazılımları yardımıyla kullanılmıştır. IFCF yönteminin daha etkili ve kolay kullanılabilir hale getirilmesi ile ilgili yaptığımız çalışmalar ayrıca Kısım 3.4'de aktarılmıştır. Ayrıca sınıflandırma amacıyla Tanaka'nın bulanık regresyon yaklaşımına dayalı yeni bir sınıflandırma metodu da geliştrilmiş ve bu metot Kısım 3.3'te özetlenmiştir.

Yukarıda bahsedilen değişik sınıflandırma yöntemlerinin 3-tekrarlı, 3-katlı döküm ve müşteri veri kümeleri üzerindeki uygulamalarından elde edilen özet sonuçlar EK'de verilmiştir. Bu sonuçların karşılaştırılması ise Kısım 2.2.3.2 de sunulmaktadır. Sınıflandırma çalışmalarının ilk sonuçları proje personeli Berna Bakır'ın yüksek lisans tezinde (Bakır, 2007) ve eski proje personeli Fatma Aydınlık Güntürkün'ün yüksek lisans tezinde (Güntürkün, 2007) yayınlanmıştır.

2.2.2.6 Optimizasyon

Üretim süreçleri karmaşık oldukları kadar süreç değişkenlerinin değerlerine hassas bir şekilde bağımlıdır. Üretim sektöründe çalışanlar için hataya sebep olan süreç parametrelerini belirlemenin yanı sıra bu parametrelerin en uygun değerlerini veya değer aralıklarını belirlemek de zor ve önemli bir iştir. Önemli süreç değişkenlerinin en iyi değerlerinin belirlenebilmesi için çeşitli yaklaşımlar kullanılabilir. Veri madenciliği uygulamalarından en yaygın kullanılan optimizasyon yaklaşımları arasında yapay sinir ağları ve yanıt yüzeylerinin optimizasyonu bulunmaktadır. Bu projede, döküm verileri üzerinde süreç değişkenlerinin optimal değerlerinin belirlenmesi için bu iki yaklaşım kullanılmıştır.

Döküm verilerinin en önemli iki hatasının yüzdesini, y2 ve y3, en aza indirgemek için x14, x16, x22 ve x29 numaralı süreç değişkenleri seçilmiştir. Değişkenler ile ilgil bazı istatistikler Tablo 2-7'de verilmiştir. Bu değişkenlerin en iyi değerlerini belirlemek üzere sinir ağları uygulamasında kalite (çıktı) değişkenleri girdi, süreç (girdi) değişkenleri ise çıktı olarak alınarak modelleme yapılmış ve üç farklı yaklaşım ile elde edilen sonuçlar karşılaştırılarak en iyi modele ve en iyi süreç değişken değerlerine ulaşılmıştır. Yanıt yüzeyleri optimizasyonu yaklaşımını uygulamak için ise her iki kalite değişkeni için ayrı ayrı yanıt yüzey modelleri geliştirilmiştir. Daha sonra bu modellerin yardımıyla tanımlanan genel çekicilik fonksiyonu optimize edilmiştir.

Tablo 2-7 Optimizasyonda Kullanılan Değişkenlerin Tanımlayıcı İstatistikleri

İstatistik	x14	x16	x22	x29	y2	y3
Ortalama	4,92	20,69	17,22	3,32	0,02	0,06
Standart sapma	0,12	3,60	5,92	0,04	0,06	0,06
Değişim aralığı	0,5	16,8	22,58	0,20	0,29	0,36
Minimum	4,7	13,2	10,85	3,21	0	0
Maksimum	5,2	30,0	33,43	3,41	0,29	0,36

Sinir ağları ile optimizasyon

Sinir ağı yaklaşımı, SPSS Clementine® 11.1 kullanılarak veriye uygulanmıştır. Veri kümesinin (92 kayıt) % 70'i öğrenme ve % 30'u ise test amaçlı olarak ikiye ayrılmıştır. Modellemede yazılımın *dynamic* ve *multiple* yaklaşımlarının sunduğu çeşitli ağ topolojileri kullanılmıştır. Öncelikle tüm yöntemler için 10 adet rassal sayı kökü üretilmiş ve tüm karşılaştırmalarda kullanılmak üzere sabitlenmiştir. Bu rassal sayı kökleri kullanılarak her yöntem 10 kez tekrarlanmıştır. Ayrıca, sonuçların güvenilirliğini artırmak amacıyla aşağıdaki üç farklı yaklaşım denenmiştir. En iyi sinir ağı modeli ve yaklaşımının belirlenmesinde çeşitli performans ölçütlerinden (sinir ağı karmaşıklığı, tahmin doğruluğu, kararlılık) yararlanılmıştır.

1. *Kalite değişkenlerinin (y2 ve y3) girdi değişkeni olması:* Bu yaklaşımda veri kümesinde yer alan iki kalite değişkeni (y2 ve y3) sinir ağı modeline girdi, dört süreç değişkeni (x14, x16, x22, x29) ise modelin çıktısı olarak kullanılmıştır.
2. Kalite değişkenlerinden elde edilen çok değişken fonksiyonunun girdi değişkeni olması: Bu yaklaşımda kalite değişkenleri y2 ve y3, kayıp fonksiyonu (Pignatiello, 1993), Mahalanobis fonksiyonu (Khuri ve Conlon, 1981) ve çekicilik fonksiyonu (Derringer ve Suich, 1980) kullanılarak tek bir değişken olarak modellenmiş, daha sonra söz konusu değişken sinir ağı modeline girdi, dört süreç değişkeni ise modelin çıktısı olarak ele alınmıştır. Çok değişkenli fonksiyon değerlerinin elde edilmesinde karar vericinin tercih bilgisi de çeşitli parametreler (kalite değişkeninin hedeflenen değeri, üst ve alt sınırları, birbirlerine göre önem dereceleri) aracılığıyla modellenmiştir.
3. Kalite değişkenlerinden (y2 ve y3) elde edilen bağımsız bileşenlerin girdi değişkeni olması: Bu yaklaşımda kalite değişkenleri y2 ve y3 öncelikle Box-Cox dönüşüm yöntemi ile normalize edilmiştir. Daha sonra normalize edilen değişkenler, Temel Bileşenler Analizi ile bağımsız bileşenlere ayrılmıştır. Son olarak; söz konusu bağımsız bileşenler sinir ağı modeline girdi, dört süreç değişkeni ise modelin çıktısı olarak kullanılmıştır.

Yapılan karşılaştırmalar sonucunda, *dynamic* yöntemi ile üçüncü yaklaşım (ana bileşenlerin girdi olması), diğerlerine göre daha basit bir tahmin modeli ürettiği ve tahmin doğruluğu ve kararlılık açısından diğerlerinden daha kötü olmadığı için seçilmiştir. Ancak elde edilen tahmin modelinin başarımı yeterince yüksek olmamıştır (öğrenme korelasyonu: 0,13 (x14), 0,43 (x16), 0,68 (x22), 0,06 (x29); test korelasyonu: 0,01 (x14), 0,35 (x16), 0,60 (x22), -0,04 (x29)). Bu modelin veriye uygulanması sonucunda örnek olarak Tablo 2-8'de verilen sonuçlara ulaşılmıştır. Örneğin; y2 ve y3 değişkenlerinin hata yüzdelerini temsil ettiğini hatırlarsak, modelimizin sonuçlarına göre, en iyi kaliteye ulaşabilmek için (y2=0 ve y3=0) x14=4,957, x16=21,032, x22=16,909 ve x29=3,314 değerlerini almalıdır.

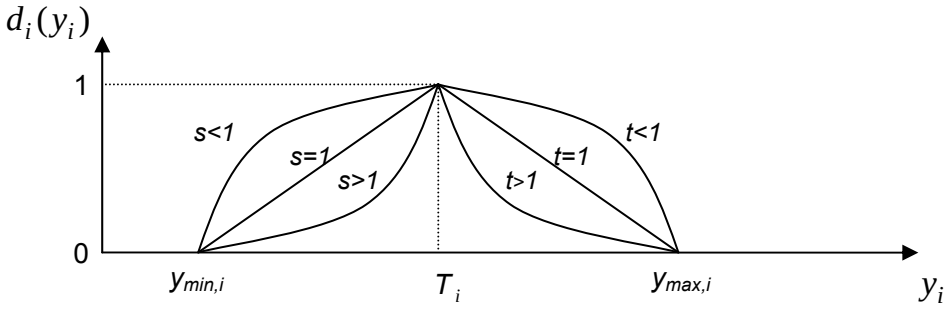
Tablo 2-8 Sinir Ağı Yaklaşımı Örnek Optimizasyon Sonuçları

Nokta	y2	y3	x14	x16	x22	x29
SA1	0	0	4,957	21,032	16,909	3,314
SA2	0	0,358	4,965	19,833	15,768	3,307
SA3	0,286	0	4,944	22,994	19,075	3,324
SA4	0,116	0,048	4,949	22,269	18,216	3,320
SA5	0,025	0,102	4,953	21,612	17,501	3,317

Yanıt yüzeylerinin çekicilik fonksiyonu ile optimizasyon

Verideki tüm kayıtlar kullanılarak y2 ve y3 için ayrı ayrı doğrusal regresyon modelleri geliştirilmiştir. Bu amaçla her iki değişkene de *arcsin(sqrt(.))* dönüşümü uygulanmış ve modellerin en yüksek R^2, düzeltilmiş R^2 (veya en düşük MSE) değerlerini verecek ve varsayımları sağlayacak şekilde geliştirilmesine çalışılmıştır. Elde edilen y2 modeli başarılı sayılabilecek iken (R^2=%67.6, düzeltilmiş R^2=%64.0), y3 için yeterince iyi bir tahmin modeli elde edilememiştir (R^2=%35.5, düzeltilmiş R^2=%27.3). Buna rağmen genel çekicilik fonksiyonu aşağıdaki şekilde tanımlanmış ve optimize edilmiştir.

$$D = \sqrt{d_1(y_2) \times d_2(y_3)}$$



Burada T_i=0.05, $y_{min,i}$=0, $y_{max,i}$=0.3, s=t=0.1, i=1,2, olarak seçilmiştir.

Minitab®15 yazılımı yardımıyla değiştirilmiş bir genelleştirilmiş indirgenmiş gradyant (*modified generalized reduced gradient*) algoritması kullanılarak aşağıdaki çözüm elde edilmiştir:

x14 = 5,05304
x16 = 25,4178
x22 = 10,85
x29 = 3,20802

Bu çözümdeki tahmini y2 ve y3 değeri 0.0025 (veya %0.25) dir. Çözümün çekicilik fonksiyon değeri 1 dir. Sinir ağı ve yanıt yüzeyi optimizasyon yaklaşımlarının sonuçları 3.2.3.4. kısımda karşılaştırılmaktadır.

2.2.3 Uygulanan Veri Madenciliği Yöntemlerinin Karşılaştırılması

2.2.3.1 Kümeleme

Daha önce Kısım 2.2.2.2'de seçtiğimiz kümeleme yöntemleri ile döküm verisinde yaptığımız uygulamaların sonuçlarından bahsetmiştik, dahili indisler yardımı ile yaptığımız geçerlilik testlerinden sonra döküm verisinde optimal küme sayısının k=2 olduğunu teğit etmiştik. Döküm verilerinde kümeleme çalışmasının tamamlanması açısından farklı kümeleme yöntemlerinde bulunan sonuçların birbirleriyle ne oranda uyuştuğu kümeleme yöntemlerinin özelliklerini ve başarımını anlamak için gereklidir, bunun için harici indisler kullanılmıştır ve kümeleme yöntemlerinin birbirleriyle ne oranda örtüştüğü ile ilgili sonuçlar sunulmuştur. Yine harici indisler kullanılarak döküm verisindeki y2 ve y3 kalite çıktı değerleri hatalı ve hatasız olarak kodlanarak, kümeleme sonuçlarımızın bunlarla ne oranda örtüştüğü bulunmuştur Kullandığımız harici indisler: Rand (Halkidi,2001), Folkes-Mallows (Halkidi,2001), Jaccard (Halkidi,2001). Tüm indis uygulamaları için MATLAB ile kodlanmış Cluster Validity Analysis Platform (CVAP) v. 3.42_7 yazılımı kullanılmıştır. Bu ücretsiz yazılım, kümeleme ve kümeleme geçerlilik indisleri ile ilgili literatürde bulduğumuz en kapsamlı ve kullanımı kolay yazılımdır.

Aşağıdaki 12 parçadan oluşan tabloda tabloların başlıklarındaki ilk yöntemden elde edilen kümeleme sonuçları ile ikinci yöntemden elde edilen sonuçların ne oranda örtüştüğüne bakılmıştır. Bu karşılaştırmada

k-Ortalamalar, MEB, SOM ve Aşamalı Tam Bağlantı Yöntemleri (H/C) sonuçları kullanılmıştır. Bulanık c-ortalamalar ve değiştirilmiş k-ortalamalar yöntemlerinin kendilerine has yazılımlarının sonuçları indisleri hesaplamada kullandığımız yazılıma girdi olarak verilememektedir. Rand, Jaccard ve FM indislerinin değeri ne kadar büyükse iki bölünme arasındaki uygunluk o kadar fazladır.

Tablo 2-9 Kümeleme Sonuçlarının Uyumu

1. MEB - k-Ortalamalar				
Ölçüler	**k=2**	**k=3**	**k=4**	**k=5**
RAND	**0.49546**	0.49188	0.49355	0.49259
JACCARD	**0.37348**	0.25053	0.22001	0.2009
FM	**0.5458**	0.40845	0.37548	0.35414

2. MEB - H/C				
Ölçüler	**k=2**	**k=3**	**k=4**	**k=5**
RAND	**0.52341**	0.45724	0.45748	0.45079
JACCARD	**0.43484**	0.27366	0.24577	0.2241
FM	**0.60614**	0.44791	0.42323	0.4011

3. MEB - SOM				
Ölçüler	**k=2**	**k=3**	**k=4**	**k=5**
RAND	-	**0.35141**	0.29766	0.27186
JACCARD	-	**0.34102**	0.28571	0.25803
FM	-	**0.58061**	0.53117	0.50568

4. k-Ortalamalar - MEB				
Ölçüler	**k=2**	**k=3**	**k=4**	**k=5**
RAND	**0.49666**	0.46966	0.46441	0.44386
JACCARD	**0.37589**	0.28132	0.24841	0.18258
FM	**0.5482**	0.45589	0.42458	0.35448

5. k-Ortalamalar - H/C				
Ölçüler	**k=2**	**k=3**	**k=4**	**k=5**
RAND	**0.49546**	0.46369	0.45843	0.43789
JACCARD	**0.37882**	0.28114	0.24859	0.18355
FM	**0.55171**	0.45672	0.42611	0.35756

6. k-Ortalamalar - SOM				
Ölçüler	**k=2**	**k=3**	**k=4**	**k=5**
RAND	-	**0.35643**	0.30053	0.21978
JACCARD	-	**0.3458**	0.28898	0.20516
FM	-	**0.58499**	0.53394	0.45

7. H/C - k-Ortalmalar				
Ölçüler	**k=2**	**k=3**	**k=4**	**k=5**
RAND	**0.49666**	0.49666	0.4957	0.49068
JACCARD	**0.37847**	0.33111	0.32166	0.20239
FM	**0.55151**	0.49752	0.48697	0.35516

8. H/C - MEB				
Ölçüler	**k=2**	**k=3**	**k=4**	**k=5**
RAND	**0.53488**	0.51959	0.51672	0.45581
JACCARD	**0.44467**	0.39024	0.38021	0.2257
FM	**0.61562**	0.56396	0.55461	0.40044

9. H/C - SOM				
Ölçüler	**k=2**	**k=3**	**k=4**	**k=5**
RAND	-	**0.50621**	0.48901	0.27425
JACCARD	-	**0.49806**	0.48007	0.26155
FM	-	**0.70378**	0.69116	0.50813

10. SOM - k-Ortalamalar				
Ölçüler	**k=2**	**k=3**	**k=4**	**k=5**
RAND	-	**0.49307**	0.48973	0.49474
JACCARD	-	**0.25044**	0.20417	0.17704
FM	-	**0.40859**	0.35684	0.32838

11. SOM - MEB				
Ölçüler	**k=2**	**k=3**	**k=4**	**k=5**
RAND	-	**0.48113**	0.45485	0.45748
JACCARD	-	**0.28834**	0.22723	0.20148
FM	-	**0.46587**	0.40138	0.3799

12. SOM - H/C				
Ölçüler	**k=2**	**k=3**	**k=4**	**k=5**
RAND	-	**0.47516**	0.44744	0.44864
JACCARD	-	**0.28808**	0.22642	0.19972
FM	-	**0.46665**	0.40154	0.37866

SOM algoritması k=2 için kümeleme yapamamaktadır. Geçen dönem raporunda ayrıntılı olarak sunduğumuz gibi SOM yönteminin Erkunt verisinde bulduğu optimal küme sayısı 3'tür. Tüm tablolarda k=2 (SOM'un bulunduğu tablolarda k=3) en büyük harici indisleri sağlamaktadır. Sonuçlar k=2 durumunda MEB ile H/C (bkz Tablo 2-9, 2. kısım) ve k=3 durumunda H/C ile SOM (bkz Tablo 2-9, 9. kısım) metodlarının birbirleriyle örtüşmekte diğerlerinden daha başarılı olduğunu göstermektedir. Tablolar veri kümelemede iki kümeye bağlı kalmanın anlamlı olduğunu göstermiştir.

Kümeleme Sonuçlarının y2 ve y3 kalite çıktı değerlerine uygunluğu

Döküm verisindeki y2 ve y3 kalite çıktı değerleri hatalı ve hatasız olarak kodlanarak, k-ortalamalar, medoidler etrafında bölümleme (MEB), kendi kendini düzenleyen haritalar (SOM) ve hierarşik aşamalı tam bağlantı (H/C) yöntemi ve bu yöntemlere ek olarak bulanık c-ortalamalar (FCM) yönteminin veriyi hatalı ve hatasız sınıflara ne oranda uygun kümelediği tespit edilmiştir. Bunun için y2 ve y3 hata tipleri ile ilgili hatalı

ürün yüzdesini gösteren y2(p): Gaz boşluğu (N gazı) (%) ve y3(p): Gaz boşluğu (H gazı) (%) çıktı değişkeni değerleri, sırasıyla hatasız ve hatalıyı gösterecek şekilde 1 ve 2 olarak kodlanarak kategorik değişkene çevrilmiştir. Burada y(p)>0.01 olduğunda hatalı yani 2, diğer durumda hatasız yani 1 olarak kodlanmıştır.

Tablo 2-10 Kümeleme sonuçlarının y2 hata tipi sınıflarına uygunluğu

y2					
Ölçüler/Yöntemler	**MEB**	**k-Ortalamalar**	**SOM**	**H/C**	**FCM**
RAND	0.50621	0.49666	0.50287	**0.50979**	0.49474
JACCARD	0.40122	0.35684	0.28439	**0.40796**	0.36677
FM	0.57309	0.52656	0.45682	**0.58014**	0.53679

Tablo 2-10'a bakarak her üç indise göre y2 hata tipine en uygun kümelemeyi Aşamalı Tam Bağlantı (H/C) yöntemi bulmuştur. Tüm yöntemlerin bulduğu kümelerin y2'ye uygunluğuna göre bir sıralama yapacak olur isek Rand indisine gore H/C, MEB, SOM, k-Ortalamalar, FCM; Jaccard indisine göre H/C, MEB, FCM, k-Ortalamalar, SOM ve FM indisine gore sıralalmanın yine H/C, MEB, FCM, k-Ortalamalar, SOM şeklinde olduğunu görmekteyiz.

Tablo 2-11 Kümeleme sonuçlarının y3 hata tipi sınıflarına uygunluğu

y3					
Ölçüler/Yöntemler	**PAM**	**k-Ortalamalar**	**SOM**	**H/C**	**FCM**
RAND	0.58075	0.53488	0.45652	**0.59006**	0.56355
JACCARD	0.50702	0.43728	0.30407	**0.51825**	0.47197
FM	0.67426	0.61524	0.49568	**0.68375**	0.646

Tablo 2-11'e bakarak her üç indise göre y3 hata tipine en uygun kümelemeyi yine Aşamalı Tam Bağlantı (H/C) Yöntemi bulmuştur. Bu yöntemlerin bulduğu kümelerin y3'e uygunluğuna göre bir sıralama yapacak olur isek her üç indise göre de sıralamanın H/C, PAM, FCM, k-Ortalamalar, SOM şeklinde olduğunu görmekteyiz.

Sonuç: Kümeleme yöntemlerinin döküm verisine uygulanmasında tüm yöntemler birbirine destekler şekilde veri kümesinde 2 küme bulmuştur. SOM yöntemi k=2 için kümeleme yapmaktadır, bizim veri kümemizde 3 grup bulmuştur. Kümeleme sonuçlarında PAM ile H/C (k=2 için) ve H/C ile SOM (k=3 için) metotlarının birbirleriyle diğer yöntemlere kıyasla daha fazla örtüstüğü görülmüstür. Hata tiplerine uygunluk bakımından en iyi kümelemeyi her iki hata içinde Aşamalı Tam Bağlantı yöntemi bulmuştur.

2.2.3.2 Tahmin etme ve sınıflandırma

Projemizin amaçlarından birisi de veri madenciliğinde sıkça kullanılan sınıflandırma ve tahmin etme metotlarını, kalite iyileştirme amaçlarına yönelik başarımına göre değerlendirmek ve buna göre uygulayıcılara tavsiyelerde bulunmaktır. Bunun için ilgili metotları uygun performans ölçütlerine göre karşılaştırmak ve öncelik sırasına dizmek gerekir. Çalışmamızda, önce kalite iyileştirme için önemli olan performans ölçüt ve ölçüleri belirlenmiş daha sonra bunları kullanarak önceliklendirme yapmak üzere Tohumcu ve Karasakal (2007) tarafından önerilen Analitik Ağ Süreci (ANP) ile bütünleşik PROMETHEE yaklaşımı kullanılmıştır. Bu çalışma hem kalite iyileştirmeye özgü bir önceliklendirmeyi ilk defa yapıyor olması, hem de bu tip bir önceliklendirmede ilk defa böyle kapsamlı bir yöntem kullanıyor olması bakımından özgündür.

Performans Ölçülerinin Belirlenmesi
Yaptığımız literatür ve saha çalışmalarına göre tahmin etme ve sınıflandırma kapsamında kalite iyileştirme amaçları genel olarak aşağıdaki şekilde ifade edilmektedir:

- kalite çıktılarını etkileyen ürün, süreç ve diğer değişkenleri ve bunlara göre çıktı değerini doğru belirleme

- bu belirlemeyi kolay ve hızlı yapma
- sonuçları kolay ve doğru yorumlama ve kullanabilme.

VM metotlarının bu kalite iyileştirme amaçlarına yönelik performansını değerlendirmek için kullanılabilecek performans ölçütleri ve ölçüleri literatürde VM metotlarının karşılaştırılması için kullanılanlarla hemen hemen aynıdır. Farklılık sözkonusu ölçülerin önem ağırlıklarındadır. Bu nedenle ilgili ölçüler benzer VM metot karşılaştırma kaynaklarından derlenmiştir (Fielding ve Bell, 1997, West v.d. 1997, Dhar ve Stein, 1997, Manel, 1999, Han ve Kamber, 2001, Ye, 2003, Munoz ve Felicisimo, 2004, Brumen v.d. 2004, Rokach ve Maimon, 2006, Hyndman ve Koehler, 2006). Bu derlemeden elde edilen kapsamlı listelerden yakınlık diyagramı yöntemiyle ölçüt grupları ve bunlara ait detaylı ölçüler aşağıdaki gibi oluşturulmuştur.

Sınıflandırma metotlarının başarımı için kullanılan ölçüler:

Doğruluk
- Yanlış Sınıflandırma Oranı (MCR) ya da Doğru Sınıflandırma Oranı (PCC)
- Kappa,
- $F_{0.5}$,
- F_1,
- F_2,
- Dengelilik (Stability of PCC),
- Kesinlik (Precision),
- Duyarlılık (Recall),
- Özgüllük (Specificity),
- Güven Aralığı (Confidence Interval (CI)),
- AUROC (Area under ROC).

Sağlamlık
- Kategorik ve sürekli değişkenlere karşı sağlamlık
- Karmaşıklığa karşı sağlamlık
- Veri gürültüsüne karşı sağlamlık
- İlgisiz değişkenlere karşı sağlamlık
- Kayıp verilere karşı sağlamlık

Modelleme kolaylığı
- Hesaplamada kullanılacak yazılım ve donanım kaynakları
- Uzmandan bağımsızlık
- Ölçeklenebilirlik
- Esneklik

Modelin kullanım kolaylığı
- Yorumlanabilirlik Yoğunluk
- Var olan sisteme yerleştirilebilirlik

Hız
- Öğrenme eğrisi ihtiyacı
- Gelişme Hızı
- Cevap Verme Hızı

Tahmin etme metotlarının başarımı için kullanılan ölçüler:

Doğruluk
- MSE (Mean Square Error)
- MAE (Mean Absolute Error)
- MAPE
- RMSE
- R
- R^2
- Düzeltilmiş R^2 (Adjusted R^2)
- Dengelilik (MSE açısından) (Stability of MSE)
- Dengelilik (RMSE açısından) (Stability of RMSE)
- PWI1 (Proportion Of Plots Within Some User-Specified Range)
- PWI2 (Proportion Of Plots Within Some User-Specified Range)

Sağlamlık
- Kategorik ve sürekli değişkenlere karşı sağlamlık
- Karmaşıklığa karşı sağlamlık
- Veri gürültüsüne karşı sağlamlık
- İlgisiz değişkenlere karşı sağlamlık
- Kayıp verilere karşı sağlamlık

Modelleme kolaylığı

- Hesaplamada kullanılacak yazılım ve donanım kaynakları (computing resource)
- Uzmandan bağımsızlık (Independence from experts)
- Ölçeklenebilirlik (Scalability)
- Esneklik (Flexibility)

Modelin kullanım kolaylığı

- Yorumlanabilirlik (Interpretability)
- Yoğunluk (Compactness)
- Var olan sisteme yerleştirilebilirlik (Embeddability)

Hız

- Öğrenme eğrisi ihtiyacı (Learning curve requirement)
- Gelişme Hızı (Development Speed)
- Cevap Verme Hızı (Response Speed)

Bu ölçüler birbirinden tamamen bağımsız değildir. Örneğin, bir metodun sağlamlık (*robustness*) düzeyi arttıkça elde edilen sonuçların doğruluk düzeyinin de bir miktar artması beklenebilir. Benzer şekilde, bir tahmin etme metodunun MSE performansı ile Düzeltilmiş R^2 performansı arasında güçlü bir korelasyon vardır. Her ne kadar seçtiğimiz önceliklendirme yöntemi, ANP, bu tür bağımlılık ilişkilerini modelleyebiliyorsa da ölçü sayısının çok olması söz konusu ilişkileri hassas bir şekilde belirlemeyi zorlaştırmaktadır. Bu nedenle, ölçü sayısının azaltılması gerekmiştir. Söz konusu azaltma için, VM metotlarının döküm ve müşteri memnuniyeti verileri üzerindeki uygulama sonuçlarının faktör analizi gerçekleştirilmiş ve buna göre doğruluk ölçülerinin oluşturduğu gruplardan uygun seçimler yapılmıştır. Bu seçimler, aşağıda belirtilmektedir:

Sınıflandırma metotlarının doğruluğu için kullanılan ölçüler:

- MCR (1-PCC)
- Kappa
- Güven Aralığı (Confidence Interval (CI))
- Dengelilik (Stability of PCC)
- Duyarlılık (Recall)
- Kesinlik (Precision)
- AUC

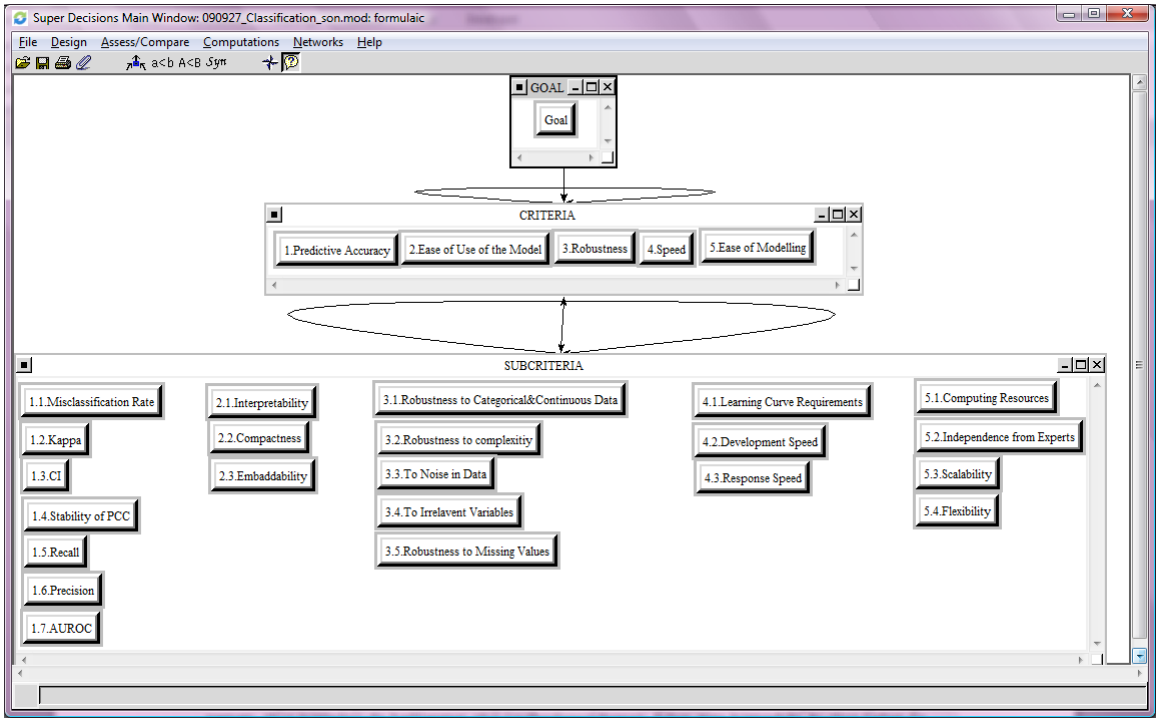
Tahmin etme metotlarının doğruluğu için kullanılan ölçüler:

- RMSE
- R2
- Dengelilik (RMSE açısından) (Stability of RMSE)

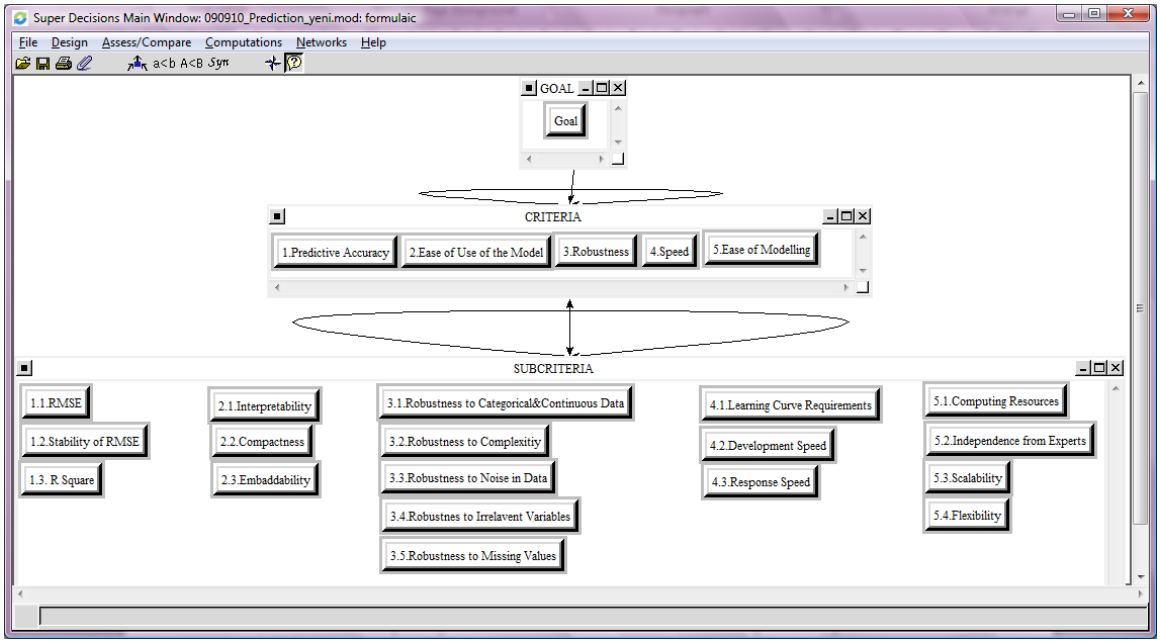
Ölçü ağırlıklarının ANP ile belirlenmesi

Çok sayıda ölçütün (ölçünün) ve alternatifin varlığında taraflı ve tarafsız ölçüleri bir arada değerlendirebilen ve ölçüler ve alternatifler arası karmaşık etkileşim yapısını değerlendirebilen geçerli bir yaklaşım Analitik Ağ Prosesi (ANP) dir (Saaty, 1996, 1999). Bu nedenle, yukarıdaki ölçüt ve ölçülerin ağırlıklarının belirlenmesi için bu yöntem kullanılmıştır.

ANP uygulamaları önceliklendirme probleminin yapısına göre farklılık gösterebilmektedir. Bu çalışmada, anlamsız ikili karşılaştırmaları önlemek, farklı ölçütlere ait ölçüleri karşılaştırabilmek ve ölçütlerin amaca olan katkısını değerlendirebilmek için Tohumcu ve Karasakal (2007) tarafından önerilen ağ yapısı kullanılmıştır. Ölçü ağırlıklarını belirlemede kullanılan ağlar Şekil 2-5 ve Şekil 2-6'da gösterilmiştir.



Şekil 2-5 Sınıflandırma ölçüleri ağ yapısı



Şekil 2-6 Tahmin etme ölçüleri ağ yapısı

Kalite iyileştirme amaçlarına en uygun sınıflandırma/tahmin etme metodunun belirlenmesi bizim ana amacımız olarak belirtilmiştir. Bu amaca göre ölçüler arasındaki mevcut etkileşimler incelenmiştir ve ölçüler arasında bağlar kurulmuştur. Farklı kümelerdeki ölçüler bağlandığı zaman, ölçü kümeleri de bağlanmış olur (*outerdependence*). Eğer aynı kümedeki iki ölçü birbiri ile etkileşim içinde ise bu küme kendi ile etkileşim (*innerdependence*) içinde demektir.

Ağ yapısı oluşturulduktan sonra ikili karşılaştırmalar yapılmıştır. Bu aşamada taraflı ölçüler için, projemizde VM metotlarını bizzat uygulayan ve yeterli bilgi ve deneyim sahibi araştırmacıların uzman görüşleri kullanılmıştır. Örneğin, doğruluk ve modelleme kolaylığı ölçütlerinin karşılaştırılarak ilişki

seviyelerinin belirlenmesi, modelleme kolaylığının tarafsız bir ölçekte ölçülememesi nedeni ile yalnızca uzman görüşüne dayalı olarak belirlenmiştir. Tarafsız ölçüler için ise uygulamalardan elde ettiğimiz istatistikler değerlendirilmiştir. Bütün karşılaştırmalar tamamlandıktan sonra *Super Decisions* yazılımının yardımıyla ANP yöntemine uygun olarak ölçüt ve ölçülerin ağırlıkları belirlenmiştir. Bu ağırlıklar Tablo 2-12 ve Tablo 2-13'de gösterilmektedir.

Tablo 2-12 Sınıflandırma ölçülerinin ağırlıkları

Sub-criteria	1.1. Yanlış Sınıflandırma Oranı	0.02419
	1.2.Kappa	0.01617
	1.3. Güven Aralığı	0.00700
	1.4. Dengelilik	0.00546
	1.5. Duyarlılık	0.03368
	1.6. Kesinlik	0.02093
	1.7.AUC	0.01283
	2.1. Yorumlanabilirlik	0.02375
	2.2 Yoğunluk	0.01162
	2.3. Var olan sisteme yerleştirilebilirlik	0.00788
	3.1. Kategorik ve sürekli değişkenlere karşı sağlamlık	0.02702
	3.2. Karmaşıklığa karşı sağlamlık	0.22656
	3.3. Veri gürültüsüne karşı sağlamlık	0.10012
	3.4. İlgisiz değişkenlere karşı sağlamlık	0.04470
	3.5. Kayıp verilere karşı sağlamlık	0.09105
	4.1. Öğrenme eğrisi ihtiyacı	0.02235
	4.2. Gelişme Hızı	0.05368
	4.3. Cevap Verme Hızı	0.07631
	5.1. Hesaplamada kullanılacak yazılım ve donanım kaynakları	0.02886
	5.2. Uzmandan bağımsızlık	0.04128
	5.3. Ölçeklenebilirlik	0.06153
	5.4. Esneklik	0.06303

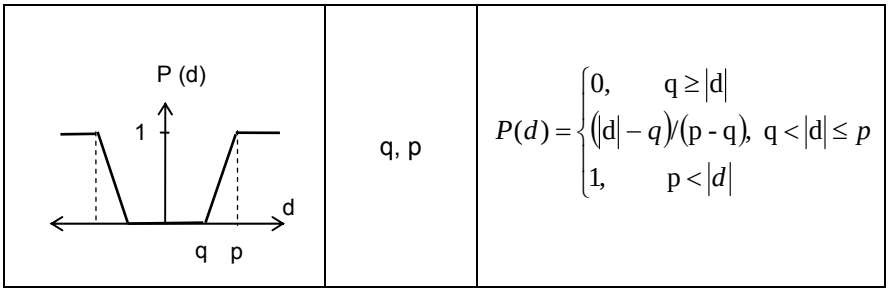
Tablo 2-13 Tahmin etme ölçülerinin ağırlıkları

Sub-criteria	1.1.RMSE	0.03496
	1.2. Dengelilik (RMSE açısından)	0.00880
	1.3. R2	0.03115
	2.1. Yorumlanabilirlik	0.02510
	2.2. Yoğunluk	0.01228
	2.3. Var olan sisteme yerleştirilebilirlik	0.00829
	3.1. Kategorik ve sürekli değişkenlere karşı sağlamlık	0.02828
	3.2. Karmaşıklığa karşı sağlamlık	0.23597
	3.3. Veri gürültüsüne karşı sağlamlık	0.10346
	3.4. İlgisiz değişkenlere karşı sağlamlık	0.04637
	3.5. Kayıp verilere karşı sağlamlık	0.09523
	4.1. Öğrenme eğrisi ihtiyacı	0.02363
	4.2. Gelişme Hızı	0.05717
	4.3. Cevap Verme Hızı	0.07865
	5.1. Hesaplamada kullanılacak yazılım ve donanım kaynakları	0.03161
	5.2. Uzmandan bağımsızlık	0.04593
	5.3. Ölçeklenebilirlik	0.06456
	5.4. Esneklik	0.06857

VM metotlarının PROMETHEE ile önceliklendirilmesi

PROMETHEE alternatiflerin çok ölçüte göre sıralanması için geliştirilmiş ve yaygın kullanılan bir yöntemdir (Brans ve Vincke, 1985). Bu yöntem ANP ye göre karşılaştırma sayısını önemli ölçüde azaltmaktadır. Ancak kullanılan ölçülerin bağımsız olduğunu varsaymaktadır. Tohumcu ve Karasakal (2007) PROMETHEE yönteminde kullanılacak ölçü ağırlıklarının ANP yöntemiyle belirlenmesi durumunda bu varsayıma olan duyarlılığın azaltılabileceğini iddia etmiştir. Bu çalışmada da benzer şekilde, VM metotlarının sıralanmasında PROMETHEE yöntemi ile birlikte ANP den elde edilen ölçü ağırlıkları kullanılmıştır.

PROMETHEE yönteminin uygulanmasında önce bütün ölçüler için tercih fonksiyonları belirlenmiştir. Bu fonksiyonların parametre değerleri uzman görüşü, literatür (Dhar ve Stein, 1997, Patel, 2003) ve uygulama sonuçlarının istatistiksel analizi yardımıyla bulunmuştur. Burada EK' de verilen tahmin etme ve sınıflandırma verileri ayrı ayrı kullanılarak her ölçü için, metotlar arasında anlamlı farklar olup olmadığı RANOVA ve LSD çoklu kıyaslama yöntemleri ile incelenmiştir. Buna göre her ölçü için, anlamlı farklılık gösteren metotlar arasındaki ortalama fark, tercih fonksiyonlarındaki belli sınır değerlerini belirlemek için kullanılmıştır. Şekil 2-7'de tarafsız ölçüler için kullanılan bir tercih fonksiyonu görülmektedir. Örneğin, AUC ölçüsü için p=0.01 , q=0.05 olarak belirlenmiştir. Burada d, seçilen iki metot arasındaki farkı göstermektedir.



Şekil 2-7 Tarafsız ölçüler için kullanılan tercih fonksiyonu

VM metotlarının her çifti için, tüm ölçüler üzerinden bu çiftin tercih fonksiyonlarında aldıkları değerlerin ağırlıklı ortalaması bulunmuştur. Bir metodun başka bir metoda tercih edilmesinin büyüklüğünü gösteren bu değerler kullanılarak her metot için pozitif ve negatif üstünlük derecesi hesaplanmıştır. Pozitif üstünlük, φ+, o metodun diğerlerine tercih edilmesinin büyüklüklerinin toplamı iken, negatif üstünlük, φ-, diğer metotların o metoda tercih edilmesinin büyüklüklerinin toplamıdır. Net üstünlük, φ, ise pozitif üstünlük ile negatif üstünlük arasındaki farktır. Metotlar net üstünlüklerine göre öncelik sırasına dizilmiştir. Sonuçta, Tablo 2-14 ve Tblo 2-15'de gösterilen sıralama elde edilmiştir.

Tablo 2-14 Sınıflandırma metotlarının üstünlük değerleri

Sınıflandırma Metodu	**φ+**	**φ-**	**φ**
MARS	1,619836	0,568826	1,05101
Destek vektör makinaları	1,386868	0,840558	0,54631
Yapay sinir ağları	1,378201	1,063973	0,314228
Bulanık sınıflandırma fonksiyonları	1,110211	1,068264	0,041947
Lojistik regresyon	1,033309	1,04559	-0,01228
Karar ağaçları	0,908688	1,056931	-0,14824
Mahalanobis Taguchi Sistemi	0,406206	0,568826	-0,16262

Tablo 2-15 Tahmin etme metotlarının üstünlük değerleri

Tahmin Etme Metodu	φ+	φ-	φ
MARS	1,04925	0,427506	0,621744
Robust regresyon	0,53237	0,427506	0,104864
Yapay sinir ağları	0,892457	0,811772	0,080685
Bulanık fonksiyonlar	0,736704	0,759342	-0,02264
Çoklu doğrusal regresyon	0,855377	0,902011	-0,04663
Karar ağaçları	0,529906	1,02483	-0,49492

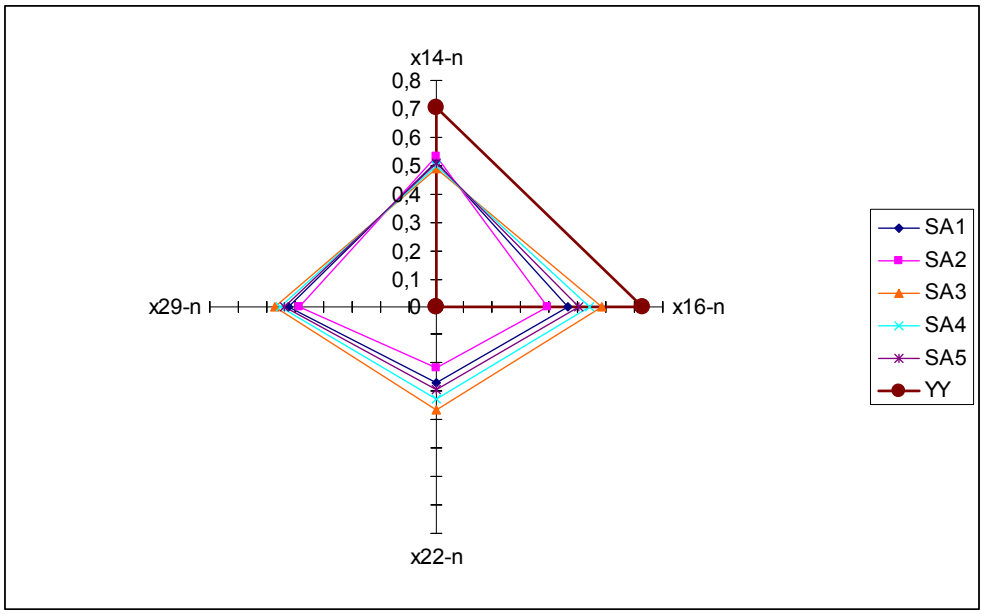
Sonuç

Veri madenciliğinde sıkça kullanılan sınıflandırma ve tahmin etme metotları, kalite iyileştirme amaçlarına yönelik başarımına göre değerlendirilmiş ve hem sınıflandırmada hem de tahmin etmede MARS metodunun diğerlerine göre öncelikle tercih edilebileceği belirlenmiştir. Sınıflandırma için MARS' tan sonra destek vektör makinaları ve yapay sinir ağlarının kullanımı önerilebilir. Tahmin etme için ise robust regresyon veya yapay sinir ağları ikinci tercih olarak önerilebilir. Bu sonuç, özellikle VM yaklaşımlarına yabancı olan imalat kalite sorumlularına bu alandaki eğitim planlama ve uygulamalarında yol gösterecektir.

Çalışmanın detayları ve elde edilen önceliklerin belli sıralama parameterlerine duyarlılığı konusunda yorumlar, Anaklı (2009) tarafından yüksek lisans tezi olarak yayımlanmak üzeredir. Çalışma ile ilgili bir uluslararası makale hazırlığı devam etmektedir.

2.2.3.3 Optimizasyon

Sinir ağları ve yanıt yüzeylerinin optimizasyonu yaklaşımları ve bunlar ile elde edilen en iyi döküm süreç değişken değerleri 2.2.2.6. Kısımda açıklanmıştır. Şekil 2-8'de söz konusu değerler veri kümesinde gözlenen minimum ve maksimum değerlere göre (0,1) ölçeğine çevrilmiş ve bu ölçek üzerinde gösterilmiştir. Şekilden de görüldüğü gibi en iyi değerler optimizasyon yöntemlerine göre farklılık göstermektedir.



Şekil 2-8 Sinir ağları ve yanıt yüzeyleri yaklaşımlarından elde edilen en iyi çözümler

Bu farklılık önemli ölçüde optimizasyonda kullanılan modellerden kaynaklanmaktadır. Sinir ağları yaklaşımı veriyi doğrudan girdi değişkenlerini tahmin edecek şekilde modellerken, yanıt yüzeyi optimizasyonu

yaklaşımında önce verinin girdileri ile çıktıları arasındaki ilişkiler modellenmekte, sonra bu modeller yardımıyla oluşturulan üçüncü bir fonksiyonun (çekicilik fonksiyonunun) en iyi değerleri bulunmaktadır. Yanıt yüzeyi modelleri veriyi ne kadar temsil ederse çekicilik fonksiyonu da o kadar anlamlı olmaktadır. Bu vaka özelinde yanıt yüzeyi modelleri (özellikle y3 kalite değişkeni için) yeterli başarımı gösterememiştir. Bu da çekicilik fonksiyonlarının optimizasyonu ile elde edilen sonuçların geçerliliği konusunda şüphe uyandırmaktadır. Öte yandan, sinir ağları modellerinin başarımı da yüksek olmamıştır. Bu nedenle, hangi yaklaşımla elde edilen çözümün daha iyi olduğunu söylemek mümkün değildir.
Optimizasyonda, kullanılan modellerin yanısıra, en iyi çözümün bulunması için kullanılan arama algoritması da önemlidir. Sinir ağlarında bu, girdi değişkenlerinin çıktıya bağlı olarak modellenmesinde kullanılan ağ topolojisi seçimi ve hata minimizasyonu yöntemlerini içerir. Sinir ağları genellikle verideki karmaşık ilişkileri modellemede başarılı bulunmaktadır (Haykin, 1994). Çekicilik fonksiyonlarının optimizasyonu ise bu fonksiyonların sürekli türevlenemeyen (pürüzlü) noktalar içermesi bakımından sorunlu olabilmektedir. Geliştirilen çeşitli yaklaştırma yöntemleri ile bu sorun belli ölçüde aşılabilmektedir. Bu konuda projemizde geliştirmekte olduğumuz alternatif pürüzlü optimizasyon yaklaşımları (bkz. Kısım 4.8) da yakın gelecekte kullanılabilecektir.

Elde edilen en iyi çözümlerin hangilerinin geçerli olduğu, ancak gerçek üretim ortamında denenerek belirlenebilir. Verinin sahibi döküm firmasının bu deneyi yapması mümkün olmamıştır. İlgili süreç, verinin sağlandığı zamandan bu yana önemli ölçüde değiştirilmiş, mevcut durumda farklı değişkenlerin süreç üzerinde daha önemli etkileri olduğu anlaşılmıştır. Ayrıca, bu çalışmada kullanılan veri, ilgili süreç değişkenlerinin ayar değerleri yerine gözlenen değerlerini içerdiğinden belirlenen en iyi süreç değişken değerlerinin kullanılabilmesi için bu gözlemleri verecek ayarlamalar konusunda ayrı bir çalışma yapılması gereklidir.

Bu veya benzer bir imalatçı, bu raporda özetlenen optimizasyon yaklaşımlarını izleyebilir. Yeni bir uygulamada veri toplama ve ön işleme aşamasında dikkat edilmesi gereken önemli bir nokta optimizasyonda kullanılan süreç değişkenlerinin ayar değerlerinin imalatçı tarafından kontrol edilebilir olmasıdır. Kullanılan veri, bu ayar değerlerini içermelidir. Optimizasyona uygun veri toplamada Taguchi yöntemleri (Phadke, M. 1989) kullanılabilir. Literatürde Taguchi yöntemleri ile sinir ağları veya yanıt yüzeyleri yaklaşımlarının birlikte kullanımı yaygındır (Koksal v.d., 2009c).

2.3 Literatür ve Saha Çalışması Sonuçlarının Değerlendirmesi

Yukarıda aktarılan literatür taraması, imalat kuruluşlarında yapılan inceleme ve toplanan veriler üzerinde yapılan uygulama ve analizler göstermiştir ki imalatta kalite iyileştirme amacıyla kullanılabilecek çok sayıda VM yaklaşımı mevcuttur. Ancak bunların başarılı olabilmesi için veri hazırlama ve önişleme aşamalarında analize uygun verinin toplanması ve/veya analize uygun hale getirilmesi büyük önem taşır. Bu konuda imalat kuruluşlarının veri yönetim düzenlerini gözden geçirmeleri ve iyileştirmeleri gerekmektedir.

Bu kuruluşlarda VM uygulamalarını gerçekleştirecek olan personelin ilgili metotlar konusundaki farkındalığı ve bilgi birikimi azdır. Yapılacak uygulamaların başarılı olabilmesi için yazılım desteği, sağlam ve kullanımı ile öğrenmesi kolay metotların seçimi ve personelin eğitimi gereklidir. Kalitenin tanımlanması ve veri önişlemede yararlı olabilecek kümeleme çalışmaları için, yaygın kullanılan k-ortalamalar metodundan ziyade H/C metodunun tercih edilebileceği gösterilmiştir. Özellikle tahmin etme ve sınıflandırma işlevlerini gören VM metotları arasında yaptığımız kapsamlı karşılaştırmalar MARS metodunun öncelikle tercih edilebileceğini bununla birlikte sınıflandırma için destek vektör makinaları ve yapay sinir ağlarının, tahmin etme için ise robust regresyon ve yapay sinir ağlarının da kullanımının önerilebileceğini göstermiştir. Kalite iyileştirmeye yönelik ürün ve süreç parametrelerinin optimizasyonu için ise yapay sinir ağları ve yanıt yüzeylerinin çekicilik fonksiyonlarının optimizasyonu metotları çalışılmıştır. Bunların her ikisinin de başarımı verinin nasıl toplandığına önemli ölçüde bağlıdır. Kullanıcı hangisini daha iyi kullanıyorsa onu tercih edebilir.

Bazı VM metotlarının uygulamalarında karşılaşılan güçlükler veya bu metotların zayıf yanları nedeniyle bunlar üzerinde iyileştirme çalışmaları yapılmış veya yeni yöntemler geliştirilmiştir. Bunlar 3. Kısımda açıklanmaktadır.

3. Kalite İyileştirme için Geliştirilen Veri Madenciliği Yöntemleri

3.1 Küçük ve Dengesiz Veriler ile İkili Sınıflandırma İçin Bir Yeniden Örnekleme Yaklaşımı

Kalite ve üretim verilerinin en belirgin özelliği, hedef değişkenlerinin (örneğin; hatalı ürün-hatasız ürün) dağılımı bakımından dengesiz olmalarıdır. Bir veri kümesinde, sınıflar hemen hemen aynı sıklıkta temsil edilmediğinde (minor/majör), o veri kümesi dengesiz kabul edilir (Chawla vd. 2002). Modelleme yöntemlerinin çoğu için bu durum bir dezavantajdır. Yöntemler genellikle, öğrenme sırasında bu dağılımı dikkate almamakta ve modelleme başarısını sınıflar bazında artırmak yerine toplamda artırmayı hedeflemektedir. Bu durum, azınlıkta olan sınıfa ait bir karar bölgesi oluşturulmasını engeller ya da güvenilirliği düşük kararlar oluşturulur. Oysa ki, üretimde hatalı ürünleri oluşturan koşulların belirlenebilmesi, hatalı ürünler ile istenen kalitedeki ürünlerin gerçekleştiği koşulların birbirlerinden ayrılabilmesi önemlidir. Bir ürünün hatalı olması az rastlanan bir durum olmakla beraber, gerçekleştiğinde meydana gelen maddi kayıp ciddi olabilmektedir. Kalite verisinde sıkça karşılaşılan bir diğer problem gözlem sayısının süreç değişken sayısına oranla az olmasıdır. Küçük veri setleri, oluşturulan modellerin geçerliliğini ve güvenilirliğini etkilemenin yanı sıra MTS gibi kimi modelleme yaklaşımlarının kullanımını da, yöntemin veri sayısına ilişkin kısıtları sebebi ile engellemektedir.

Üretim ve bazı diğer uygulama alanlarında (sağlık, sigorta, vs.) karşılaşılan dengesiz ve küçük veri problemi için literatürde bir takım veri örnekleme yaklaşımları önerilmektedir. Veri örnekleme (çoğaltma/azaltma) arkasında yatan düşünce basittir ve bu işlem üç şekilde yapılabilir (Estabrooks vd. 2004); bunlar, (1) öğrenme setinde yer alan gözlemlerin şansa bağlı olarak seçilerek minor ve major sınıfsal denge kurulana kadar çoğaltılması, (2) öğrenme setinde yer alan gözlemlerin şansa bağlı olarak seçilerek minor ve major sınıfsal denge kurulana kadar azaltılması, (3) her iki metodun kombinasyonudur. Bazı çoğaltma metodları, verinin ezberlenmesine (overfitting) yol açarken, bazı azaltma metodları ise gözlem azaltılması sırasında önemli bilgilerin kaybolmasına neden olabilir (Liu, 2004). Tüm bu yaklaşımlar için, model dengeli test veri setlerinde test edilir ve performans değerlerine bakılır. Veri çoğaltma oranı kurulacak sisteme iliskin bir parametredir (Estabrooks vd., 2004, Weiss et. al. 2003). Literatürde, bu oranın çalışılan veriye bağlı olarak değişeceği söylenmektedir (Estabrooks vd., 2004). Weiss vd. (2003) çalışmasında orjinal örnek dağılımına ait test verisi kullanmıstır. Bu şekilde, şansa bağlı 25% minor sınıf gözlemleri ve 25% de major sınıf gözlemlerinden test verisi oluşturulmuştur. Kalan gözlemleri öğrenme seti yaparak önceki ve sonraki sınıf dağılımı olmak üzere karar ağacı ile modellemişlerdir. Bu şekilde % 10.6 oranında hata azalması sağlanmıstır. Estabrooks vd. (2004) ise, öğrenme setinde minor ve majör sınıfların en fazla aynı oranda olmasına kadar, veri çoğaltmak için gereken oranı tespit etmeye çalışmıştır. Bu şekilde yaparak, %20 oranında bir veri çoğaltmasıyla optimal sonuçlara ulaşmıştır. Chawla vd. (2002) çalısmasında, daha dengeli veri kümesi oluşturmak için, minör sınıfında kalan gözlemleri kullanarak bu sınıfa ait yeni örnekler yaratmıştır. SMOTE olarak adlandırılan bu yöntemde, azınlık sınıfındaki her bir verinin en yakın k komsusu bulunmakta ve bu komşulardan bir tanesi rasgele seçilmektedir. Daha sonra, her bir değişken için bu iki noktayı birleştiren doğru üzerinde rastgele yeni (sentetik) bir değer üretilir. Tüm değişkenler için aynı işlem yapıldığında yeni bir sentetik gözlem vektörü elde edilmiş olur. Naive Bayes, C4.5 Karar Ağacı algoritması, Ripper gibi sınıflandırma yöntemleri kullanılarak yapılan karşılaştırmalarda SMOTE yönteminin C4.5 ve Ripper gibi sınıflandırma yöntemleri ile kullanılmasının, ROC eğrilerinde ve sonucunda hesaplanan AUC değerlerinde artış sağladığı görülmüştür. Özetle, veri çoğaltma/azaltma yöntemleri ile dengeli hale getirilen veri setleri ile kurulan modellerin, orijinal veri kümesi ile kurulan modellere göre daha iyi sonuçlar verdiği ortaya konmuştur.

Literatürdeki çalışmalarda, genellikle, deneme-yanılma yaklaşımının izlendiği görülmüştür. Ayrıca, hesaplama maliyeti sebebi ile veri artırma oranının sınırlı sayıda farklı değeri denenmiştir. Dolayısı ile en iyi değerler sınırlı sayıda sonuç içinden seçilmiş, olası başka çözümlerin varlığı irdelenmemiştir. Chawla vd. (2002) kullandığı SMOTE algoritmasında çoğaltma işlemini minör grubun tam katlarına kısıtlamıştır. Kullanılan komşuluk sayısının sonuçlara etkisi olup olmadığına bakılmamış ve 5 tam kata kadar çoğaltma yapılabilmesi için 5 olarak belirlenmiştir. Bunlara ek olarak, örneklem büyüklüğünün model performansı üzerindeki etkisi de incelenmemiştir. Çalışmalarda dengesiz veri kümelerinin modellenmesindeki zorluk dile getirilirken, her çalışma bir veya bir kaç spesifik probleme odaklanmış ve çözümleri de yine çalışılan spesifik problemlere yönelik olmuştur.

Bu çalısmada, SMOTE yöntemi temel alınarak ve yöntem bazı bakımlardan geliştirilerek, sürekli girdi değişkenleri ve ikili çıktı değişkenine sahip dengesiz veri setleri için Matlab programı ile bir yeniden örnekleme algoritması geliştirilmiştir. Çalışma iki ana amaca yöneliktir. Birincisi, literatürde ortaya konan, veri çoğaltma sistemine ait parametrelerin problem ve veriye özgü olduğu bulgusunu geniş kapsamlı bir

örnekleme kümesi ile sınamak ve olası kuralları aramak, ikincisi ise belirtilen tipte dengesiz veri kümeleri için en iyi veri çoğaltma parametrelerini otomatik olarak önerebilmektir. Algoritma verilen dengesiz bir veri kümesi için, azınlık grubun verideki oranı (r), modellemeye giren toplam veri sayısı (n) ve SMOTE yönteminde veri üretmek için kullanılan komşu nokta sayısı (k) olmak üzere 3 farklı yeniden örnekleme parametresinin en iyi değerlerini sistematik bir biçimde aramaktadır. Bu parametrelerin alabileceği maksimum değerler ve iterasyonlardaki adım büyüklüğü kullanıcı tarafından tanımlanabilmektedir. Öncelikle orijinal veri kümesini 3 tekrarlı 3-katlı çapraz doğrulama yaklaşımı ile bölmektedir. Yeniden örnekleme işlemi elde edilen 9 öğrenme kümesine uygulanmakta, 9 test kümesi sonuçların test edilmesi için orijinal hali ile bırakılmaktadır. Kullanıcının tanımladığı değerlere göre, 3 parametrenin olası tüm kombinasyonlarına karşılık yeniden örnekleme yapılmaktadır. Her iterasyonda parametrelerden birinin değeri değişmekte ve bu senaryoya karşılık olarak ya SMOTE ile veri çoğaltma ya da rasgele seçim ile veri azaltma işlemi yapılmaktadır. Azınlık grubun verideki oranı sabit iken modellemeye giren veri sayısının farklı değerleri için de yeniden örnekleme yapıldığından, sadece minör grup değil majör grup da SMOTE algoritması ile çoğaltılabilmektedir. Aynı durum farklı sayıda komşuluk kullanılarak tekrarlanmaktadır. Oluşturulan her veri kümesi, MTS ve KA (CART) yaklaşımları ile modellenmekte, elde edilen modellerin orijinal oranda bırakılan test kümelerinde performansı ölçülmektedir. Algoritma, performans değerlenirmede doğru sınıflandırma oranı, özgüllük, duyarlılık, F-ölçütü, özgüllük ve duyarlılığın geometrik ortalaması ve ROC eğrisi altında kalan alan ölçütlerini kullanmaktadır. 3 yeniden örnekleme parametresinin her kombinasyonunda bu ölçümlerin değerleri bir dosyaya yazılmaktadır.

Algoritma, Pima Indian Diabetes, Magic Gamma Telescope ve Wisconsin Breast Cancer olmak üzere literatürde sıkça kullanılan 3 veri kümesinden elde edilen farklı büyüklük ve dengesizlikte 9 veri kümesi ile çalıştırılmıştır. Veri büyüklüğü için 70, 200 ve 500, minör grubun verideki oranı için 0.1, 0.2 ve 0.3 olmak üzere 3 farklı seviyenin kombinasyonlarından 9 veri kümesi elde edilmiştir (örneğin; oluşturulan veri kümelerinden birinin örneklem büyüklüğü 200, minör grubun verideki oranı ise 0.2'dir). Tüm veri kümeleri için, 45915'i MTS ve 43842'si KA yaklaşımı ile olmak üzere toplam 89457 model üretilmiştir. Üretilen bu performans verileri ANOVA ve KA yaklaşımları ile analiz edilmiştir. ANOVA sonuçları, orijinal verinin büyüklüğünün ve dengesizlik düzeyinin (70, 200, 500, 0.1, 0.2 ve 0.3 değerleri) model performans değerleri üzerinde önemli bir etkisinin olmadığını göstermiştir. Bununla birlikte, *r, n* ve *k* değerleri ile kullanılan veri kümesinin (Indian, Telescope, Cancer) model performans değerleri üzerindeki etkisi önemli çıkmış ancak bu etkinin büyüklüğü sadece kullanılan veri kümesi değişkeni için anlamlı olmuştur (partial eta square=0.491). Özetle, ANOVA sonuçları kullanılan veri kümesinin sonuçlar üzerinde etkili olduğunu göstermiş, bu da literatürdeki bulguları desteklemistir. KA yaklaşımında ise anlamlı kurallar elde edilememiştir. Sonuç olarak, yeniden örnekleme parametrelerinin model performansı üzerindeki etkisi çok büyük bir örneklem ile sınanmış ve genel bir kural olmadığı görülmüştür. Farklı problemler ve veri yapıları için parametrelerin optimum değer ya da değer aralıkları farklılık gösterebilmektedir. Bu sebeple, her yeni problem ve veri yapısı için en iyi değerlerin yeniden taranması uygun olacaktır. Bu islem elle yapıldığında zaman açısından maliyeti çok yüksektir. Örneğin, 5 farklı r değeri için, 10 farklı n değeri ve 5 farklı k değeri denemek 250 tane model kurmayı gerektirecektir. 3-katlı çapraz bölünme olması halinde ise bu sayı 750'ye çıkmaktadır. Bu durumda, parametre uzayının en iyi değerler için otomatik olarak taranabilmesi hesaplama maliyetini çok ciddi ölçüde düşürecektir. Örneğin, çalışılan veri kümelerinde orijinal veri ve en iyi performansı gösteren yeniden örneklenmiş veri kullanılarak elde edilen modellerin performan sonuçları özgüllük ve duyarlılığın geometrik ortalaması ölçütüne göre Tablo 3-1 ve Tablo 3-2'de karşılaştırılmıştır. Tablo 3-1'deki modeller MTS yaklaşımı ile Tablo 3-2'dekiler ise KA yaklaşımı ile kurulmuştur. Sonuçlar incelendiğinde, önerilen en iyi parametre değerlerinin model performanslarını önemli ölçüde artırdığı görülmüştür. Özellikle, KA yöntemleri, dengesizliği fazla olan verilerde çok başarısızken, algoritma tarfından önerilen parametre değerlerinde performansda büyük artışlar olduğu gözlenmiştir. Örneğin, 3 numaralı veri kümesinde, minör grubun 0.2 oranında yer aldığı 70 gözlemli orijinal veri kümesinde kurulan modelin geometrik ortalama değeri 0 olup minör grup hiç sınıflandırılamazken, veri kümesini 5 komşuluk kullanarak tam dengeli hale getirecek ve toplam veri sayısını 120'ye çıkaracak şekilde yeniden örnekleme yapıldığında, bu değer 0.823'e çıkmaktadır.

Tablo 3-1 MTS Modellerinin sonuçlarına göre önerilen en iyi yeniden örnekleme parametreleri ve orijinal veri ile performans karşılaştırmaları

Veri Kümesi	Başlangıç N	Başlangıç R	r	n	k	Gmean (orijinal veri)	Standart Sapma (orijinal veri)	Gmean (yeniden örneklenen veri)	Standart Sapma (yeniden örneklenen veri)
1	70	0.1	0.3	444	1	0.875	0.089	0.967	0.030
1	70	0.2	0.5	310	6	0.802	0.105	0.941	0.054
1	200	0.1	0.1	175	7	0.930	0.042	0.946	0.036
1	200	0.2	0.3	337	7	0.960	0.021	0.978	0.020
2	70	0.1	0.2	124	1	0.385	0.304	0.576	0.082
2	70	0.2	0.3	32	1	0.652	0.090	0.660	0.116
2	200	0.1	0.5	126	7	0.687	0.072	0.728	0.057
2	200	0.2	0.3	285	1	0.642	0.053	0.690	0.033
3	70	0.1	0.2	74	3	0.588	0.247	0.654	0.139
3	70	0.2	0.2	198	5	0.644	0.109	0.754	0.086
3	200	0.1	0.1	277	2	0.704	0.110	0.756	0.059
3	200	0.2	0.4	164	1	0.666	0.064	0.701	0.058

Tablo 3-2 KA Modellerinin sonuçlarına göre önerilen en iyi yeniden örnekleme parametreleri ve orijinal veri ile performans karşılaştırmaları

Veri Kümesi	Başlangıç N	Başlangıç R	r	n	k	Gmean (orijinal veri)	Standart Sapma (orijinal veri)	Gmean (yeniden örneklenen veri)	Standart Sapma (yeniden örneklenen veri)
1	70	0.1	0.3	16	1	0.246	0.382	0.894	0.132
1	70	0.2	0.3	132	5	0.839	0.107	0.941	0.090
1	200	0.1	0.2	118	2	0.821	0.097	0.861	0.067
1	200	0.2	0.4	317	6	0.857	0.051	0.904	0.060
2	70	0.1	0.3	116	2	0.061	0.183	0.416	0.317
2	70	0.2	0.5	168	4	0.120	0.238	0.758	0.178
2	200	0.1	0.5	78	2	0.000	0.000	0.687	0.080
2	200	0.2	0.5	304	6	0.247	0.238	0.653	0.048
3	70	0.1	0.5	110	3	0.000	0.000	0.554	0.325
3	**70**	**0.2**	**0.5**	**120**	**5**	**0.000**	**0.000**	**0.823**	**0.089**
3	200	0.1	0.4	335	2	0.351	0.218	0.754	0.070
3	200	0.2	0.5	154	2	0.421	0.325	0.778	0.032

Yapılan bir diğer çalışma, bir modelleme metodu ile bulunan en iyi yeniden örnekleme değerlerinin başka bir modelleme yönteminde nasıl sonuç verdiğinin irdelenmesidir. Bu amaçla, MTS modellerinin önerdiği en iyi yeniden örnekleme değerleri ile çoğaltılmış veriler hem MTS hem de KA yaklaşımı ile modellenmiş, aynı şekilde, KA modellerinin önerdiği en iyi yeniden örnekleme değerleri ile çoğaltılmış veriler de yine hem MTS hem de KA yaklaşımı ile modellenmiştir. Genel olarak, her iki yöntem de farklı bir modelleme yaklaşımı ile bulunmuş en iyi yeniden örnekleme değerlerinde başarılı sonuçlar üretebilmiştir. Ancak küçük veri setlerinde üretilen en iyi değerlerde kullanılan yaklaşım değiştiğinde performansta bir miktar azalma görülmüştür. Örneğin, Tablo 3-3'de görüldüğü gibi, 2 nolu veri kümesinden elde edilen N=70, R=0.1 dengesiz veri kümesinde, MTS yaklaşımı ile bulunan en iyi parametre değerleri kullanılarak kurulan MTS modelinde ROC eğrisi altında kalan alan 0.636 iken, aynı veri için KA yaklaşımının önerdiği en iyi parametre değerleri kullanılarak kurulan MTS modelinde bu değer 0.585 olmuştur. Aynı veri kümesi için KA modellerindeki değişim ise KA ve MTS kullanıldığında sırası ile 0.670 ve 0.599 olmuştur. Verideki dengesizlik azaldıkça bu farkın kapandığı görülmektedir. Genel olarak, en iyi parametrelerin bulunduğu yaklaşım ile modelleme yapılacak yaklaşım farklı olsa da, önerilen yeniden örnekleme parametreleri ile oluşturulan modellerde görülen performans artışı bir çok denemede gözlenmiştir.

Tablo 3-3 Farklı modelleme yöntemlerinin kullanımının sonuçlara etkisi

							AUC	
Yeniden Örnekleme Metodu	**Data**	**Başlangıç N**	**Başlangıç R**	**r**	**N**	**k**	**MTS**	**KA**
1	1	70	0.1	0.3	444	1	**0.978**	0.817
2	1	70	0.1	0.3	16	1	NaN	**0.816**
1	1	70	0.2	0.5	310	6	**0.985**	0.884
2	1	70	0.2	0.3	132	5	**0.972**	0.851
1	1	200	0.1	0.1	175	7	**0.990**	0.863
2	1	200	0.1	0.2	118	2	**0.984**	0.836
1	1	200	0.2	0.3	337	7	**0.998**	0.881
2	1	200	0.2	0.4	317	6	**0.997**	0.861
1	2	70	0.1	0.2	124	1	**0.636**	0.599
2	2	70	0.1	0.3	116	2	0.585	**0.670**
1	2	70	0.2	0.3	32	1	**0.774**	0.548
2	2	70	0.2	0.5	168	4	0.696	**0.778**
1	2	200	0.1	0.5	126	7	**0.814**	0.724
2	2	200	0.1	0.5	78	2	**0.777**	0.680
1	2	200	0.2	0.3	285	1	**0.770**	0.679
2	2	200	0.2	0.5	304	6	**0.774**	0.695
1	3	70	0.1	0.2	74	3	**0.683**	0.646
2	3	70	0.1	0.5	110	3	0.624	**0.691**
1	3	70	0.2	0.2	198	5	**0.827**	0.808
2	3	70	0.2	0.5	120	5	**0.849**	0.790
1	3	200	0.1	0.1	277	2	**0.778**	0.705
2	3	200	0.1	0.4	335	2	0.789	**0.794**
1	3	200	0.2	0.4	164	1	0.785	**0.796**
2	3	200	0.2	0.5	154	2	0.782	**0.795**

Yeniden Örnekleme Metodu: 1: MTS, 2: KA

Sonuç olarak, geliştirilen algoritma tüm işlemleri (veri bölme, veri çoğaltma, modelleme, performans hesaplama) kullanıcı müdahalesi olmadan kendi kendine yapabilmektedir. Kullanıcının en basta veri dosyasını seçmesi ve parametrelerin hangi detayda taranacağı bilgilerini girmesi yeterlidir. Bu anlamda, veri çoğaltma işlemi yapacak bir kullanıcı, geliştirilen algoritma ile en iyi veri kümesini oluşturan değerleri belirledikten sonra elde ettiği veri kümesini farklı yazılımlar kullanarak da modelleyebilecektir. Baska bir değişle, algoritma optimum değerleri belirlemenin yanı sıra bir veri hazırlama aracı olarak da kullanılabilir.
Algoritmanın geliştirme çalışmalarına, en iyi kümeyi bulmada modelleme aracı olarak MTS ve DT yaklaşımları dışında farklı yöntemler de kullanabilmesi için devam edilmesi planlanmaktadır. Ayrıca kod, açık kaynak olarak kullanıma sunulacaktır. Bu çalışmanın daha fazla veri kümesi ve modelleme yaklaşımı ile sınanması yapılarak uluslararası bir makale hazırlanması yönünde çalışmalar devam etmektedir.

3.2 Tahmin ve Sınıflandırma için Yeni Bir Yöntem: CMARS - Sürekli Optimizasyon Tarafından Desteklenen Çok Değişkenli Uyarlanabilir Regresyon Eğrileri ile Parametrik Olmayan Regresyona Yeni Bir Katkı

Doğrusal regresyon (LRM) regresyon modelleme yöntemleri içerisinde en yaygın kullanılanıdır. Ancak bu tür parametrik modellerin dayandığı dağılım varsayımları geçersiz olduğunda başarısız olabilmektedir. Parametrik olmayan regresyon modelleri, örneğin, çok değişkenli uyarlanabilir regresyon eğrileri (MARS) bu tür dezavantajların üstesinden gelmek amacıyla geliştirilmiştir. MARS algoritması genel olarak iki aşamadan oluşmaktadır. İleriye doğru adım algoritmasında temel fonkisyonlar ve/ya bu fonksiyonların çarpımları ile en büyük modele ulaşılır. İleri doğru adım algoritmasının her aşamasında kullanılabilecek en uygun düğüm noktaları ve temel fonksiyonları belirlemek amacı ile genelleştirilmiş çapraz doğrulama (GCV) ölçüsünden yararlanılmaktadır. MARS algoritmasının ilk adımında oluşturulan bu en büyük (karmaşık) modelin yorumlanması ve kullanımı kolay olmadığından ikinci adımda bu model budanarak, yani önemli bağımsız değişkenler ve bu değişkenlerin etkileşimleri belirlenerek, GCV ölçüsü en küçük olan model elde edilir (Friedman, 1991).

Bu çalışmada MARS algoritmasının ikinci aşaması için yeni bir yaklaşım geliştirilmiştir. Bu yaklaşımda cezalı hata kareler toplamı kullanılarak Tikhonov düzenlemesi şeklini alan MARS modeli sürekli optimizasyon tekniklerinden biri olan ikinci dereceden konik karesel programlama (CQP) problemine dönüştürülmektedir. Bu yeni yaklaşıma CMARS adı verilmiştir (Yerlikaya, 2008).

Bu çalışmada CMARS yönteminin başarımı MARS ve LRM yöntemlerinin başarımı ile farklı özelliklere sahip iki veri kümesinde çeşitli ölçümler yardımı ile karşılaştırılmıştır. Kullanılan iki veri kümesinden biri gerçek yaşam verisi olup metal döküm sanayiden edinilmiştir. 'Metal döküm' verisi olarak adlandırılan bu veri kümesi 34 ürün ve/ya işlem değişkeni, 92 gözlem ve hatalı ürün yüzdesini gösteren yanıt değişkeninden oluşmaktadır. 'Uniform örneklem' verisi olarak adlandırılan diğer veri kümesi ise katı roket motoru üzerinde benzetim yöntemi ile gerçekleştirilen 'uniform örneklem' deney tasarımından elde edilmiştir. Bu veri kümesi ise yedi bağımsız değişken, toplam tepkiyi ifade eden yanıt değişkeni ve 100 gözlemden oluşmaktadır. Bu çalışmada üç-katlı üç tekrarlı çapraz doğrulama yöntemi ile dokuz eğitim ve dokuz test verisi kullanılarak modellerin başarımları karşılaştırılmıştır.

Çalışmada LRM modelleri adım adım regresyon yöntemi ile geliştirilmiştir. Bu modeller geliştirilirken en küçük kareler tahminine (LSE) ilişkili tüm varsayımlar test edilmiştir. Bu varsayımlardan herhangi biri geçersiz olduğu durumda yanıt ve/ya bağımsız değişkenlerin dönüştürülmesi vb. gibi düzeltici önlemler alınmıştır. Fakat bu önlemlerin herzaman başarılı olması mümkün olamayabilmektedir. Örneğin, 'metal döküm' verisi üzerinde LRM modeli geliştirilirken böyle bir problemle karşılaşılmıştır. Diğer yandan MARS modelleri oluşturulurken Salford System yazılımı kullanılmış; CMARS modelleri için ise MATLAB'da çalışan bir program yazılmıştır.

Yukarıda belirtildiği şekilde dokuz eğitim ve dokuz test veri kümesi için oluşturulan modeller için daha sonra çeşitli karşılaştırma ölçümleri hesaplanmış ve her bir ölçümün dokuz değerinin ortalamaları bulunmuştur. Tablo 3-4 ve Tablo 3-5 sırasıyla 'metal döküm' ve 'uniform örneklem' veri kümeleri için CMARS, MARS ve LRM modellerinin çeşitli ölçülere göre başarım ortalamalarını göstermektedir. Benzer şekilde Tablo 3-6'da her iki veri kümesi için de başarım ölçümlerinin kararlılıklarına yer verilmektedir.

Tablo 3-4 Metal döküm verisi için model başarımlarının ortalaması

	Eğitim			**Test**		
Ölçümler	**LRM**	**MARS**	**CMARS**	**LRM**	**MARS**	**CMARS**
MAE	0,2754	0,2444	0,2145	0,2999	0,8039	0,7362
MSE	0,8668	0,1627	0,1644	1,5345	4,1461	5,6863
R2	0,5078	0,8558	0,8880	0,0728	0,1598	0,1921
PWI	0,8673	0,9819	0,9819	0,8742	0,9748	0,9711
PRESS	88,9907	10,9591	10,1713	709,2968	672,5287	685,3657

Tablo 3-5 Uniform örneklem verisi için model başarımlarının ortalaması

Ölçümler	**Eğitim**	**Test**

	LRM	MARS	CMARS	LRM	MARS	CMARS
MAE	0,2582	0,0352	0,0741	0,2782	0,0604	0,0937
MSE	0,2500	0,0027	0,0242	0,3717	0,0111	0,0551
R2	0,9852	0,9976	0,9847	0,9790	0,9933	0,9817
PWI	0,9950	0,9917	0,9951	0,9900	0,9831	0,9868
PRESS	23,3219	0,2001	1,4729	88,8165	0,4372	1,4246

Tablo 3-6 Her iki veri kümesi için geliştirilmiş modellerin başarım ölçümlerinin kararlılıkları

	Eğitim			Test		
Ölçümler	**LRM**	**MARS**	**CMARS**	**LRM**	**MARS**	**CMARS**
MAE	-0,3537	-0,508	-0,533	-0,1914	-0,2603	-0,1973
MSE	-0,7028	-0,827	-0,896	-0,5089	-0,5871	-0,559
R2	0,7758	0,7017	0,666	0,0031	0,0022	0,0015
PWI	-0,0039	0,0037	0,0055	0,0026	0,0044	0,0042
PRESS	-0,8057	-0,7709	-0,7972	-0,5981	-0,3644	-0,2534

MAE: Ortalama mutlak hata; *MSE:* Hata kareler toplamı; R^2: Çoklu belirleme katsayısı*; PWI:* Kullanıcının tanım aralığındaki yanıt yüzdesi; *PRESS:* Tahmin hata kareler toplamı.

Geliştirilen modellerin istatistiksel olarak farklı olup olmadıklarına karar verebilmek amacı ile SPSS yazılımı kullanılarak tekrarlı varyans analizi (RANOVA) yöntemi uygulanmıştır. RANOVA'ya ilişkin sıfır hipotezi α = 0.05 önem seviyesinde red edildiğinde, hangi modeller arasında farklar olduğunu belirlemek için çoklu karşılaştırma testi Fisher'in en küçük önemli fark (LSD) testi kullanılmıştır.

Bu sonuçlarına göre 'metal döküm' eğitim ve test verilerinde modeller arasında R^2 ölçümleri bakımından, 'metal döküm' kararlılık verisinde ise MAE ve MSE ölçümleri açısından α=0.05 düzeyinde istatistiksel olarak anlamlı bir fark bulunmuştur. Diğer yandan 'uniform örneklem' e ilişkin yalnızca kararlılık verisinde R^2 ölçümüne göre modeller arasında aynı önem düzeyinde istatistiksel olarak anlamlı farklılık bulunmuştur.

RANOVA ve LSD test sonuçlarına dayanarak 'metal döküm' verisi için LRM'nin başarımının MARS ve CMARS'a göre daha kötü olduğu söylenebilir. Bunun nedeni bu veri kümesinin yüksek boyutlu olması ve değişkenleri arasında doğrusal olmayan bir ilişkinin var olması olabilir. R^2 ölçümüne göre CMARS ve MARS modellerinin veriye iyi bir uyum sağladığı söylenebilir. Hatta CMARS'ın yine aynı ölçüye göre MARS'tan daha başarılı olduğu görülmektedir (p-değeri=0.021). Diğer taraftan CMARS, MAE ve MSE ölçümlerinin kararlılıkları açısından LRM ve MARS a göre daha başarılı değildir.

'Uniform örneklem' veri kümesi için ise modellerin başarımları arasında herhangi bir fark tespit edilememiştir. Büyük olasılıkla bunun nedeni verinin yapısından kaynaklanmaktadır. Çünkü araştırmalı veri analizi sonuçlarına dayanarak bu veri kümesinin bağımlı ve bağımsız değişkenleri arasında doğrusal bir ilişki bulunduğu söylenebilir. Bu nedenle çalışılan tüm modeller, yani LRM, MARS ve CMARS, bu veri kümesine iyi bir uyum sağlamaktadır.

Sonuçta veride doğrusal olmayan bir yapı söz konusu olduğunda CMARS ve MARS'a göre LRM modelleri başarısız olurken, CMARS, MARS a göre veriye biraz daha iyi bir uyum sağlamaktadır. Eğer veride doğrusal bir yapı var ise MARS ve CMARS yöntemleri LRM kadar iyi bir başarım sergilediği söylenebilir (Weber vd. ,2009).
Bu çalışma *"**Journal of Computational and Applied Mathematics"*** dergisine 10. Eylül. 2009 tarihinde ***"CMARS: A New Contribution to Nonparametric Regression with Multivariate Adaptive Regression Splines Supported by Continuous Optimisation"*** adlı makale olarak gönderilmiştir.

3.3 Bulanık Sınıflandırma için Tanaka Bulanık Regresyon Yaklaşımına Dayalı Modelleme

Kalite çıktıları için sınıflandırma modellerinin geliştirilmesinde bulanık sınıflandırma yöntemlerine duyulan gereksinim 2.2.2.4.5 Kısımda belirtilmiştir. Bu tür sınıflandırma yaklaşımlarının sayısının ve başarım düzeylerinin kısıtlı olması nedeniyle, projemizde alternatif bulanık sınıflandırma yöntemlerinin geliştirilmesi üzerinde çalışılmıştır. Bununla ilgili olarak bulanık tahmin problemleri için yaygın olarak kullanılan ve Tanaka (1982) tarafından geliştirilen Bulanık Doğrusal Regresyon (FLR) yönteminin sınıflandırma amacıyla

kullanımına olanak tanıyan bir yaklaşım geliştirilmiştir. Bu yaklaşım müşteri memnuniyeti verileri kullanılarak gösterilmiştir (Özer 2009, Özer ve diğ. 2009a).

Bulanık yöntemler üzerine yapılan kapsamlı literatür çalışması, veri içerisindeki belirsizliklerin bir çok farklı şekilde yorumlanıp modele yansıtılabileceğini göstermiştir. Koltuktan memnuniyet verisinde rassallık ve bulanıklık olmak üzere iki tip belirsizliğin var olduğu düşünülmektedir. Bulanık belirsizlik kullanıcı değerlendirmelerini yansıtan sözel terimlerden kaynaklanmakta iken rassallık, rastgele değişen ve kullanıcıların genel memnuniyet seviyelerini etkileyebileceği düşünülen faktörlerin varlığından kaynaklanmaktadır. Kamburluk veya boyun/bel fıtığı gibi iskelet problemleri olan bir insanın, araç kullanımı ve demografik özellikleri birebir aynı fakat sağlıklı iskelet yapısına sahip bir insana göre sürücü koltuğundan daha az memnuniyet duyması rassallıktan kaynaklanan bir belirsizliğe örnek olarak verilebilir. Bu nedenle, yalnızca bulanık belirsizliği ele alan Tanaka'nın yöntemi ile modelleme yapılırken rassallıktan kaynaklanan belirsizliğin etkisini de göz ardı etmemek adına olasılık teorisinin araçlarından olan frekanslardan faydalanılmıştır.

Geliştirilen bulanık sınıflandırma yaklaşımı, Tanaka'nın yönteminin (1989) sürücü koltuğundan yüksek memnuniyet duyan kullanıcı oranlarını modellemek için kullanılmasını önermektedir. Ancak koltuktan memnuniyet verisi tekrarlı ölçümleri içermediğinden yüksek memnuniyet derecesini seçen kullanıcıların oranları Lojistik Regresyon (LR) yöntemi kullanılarak tahmin edilmiştir. Elde edilen sonsal oran (olasılık) değerleri Tanaka'nın yönteminde kesin çıktı değişken değerleri olarak kullanılmıştır. Bu nedenle, bu yaklaşım **Lojistik Regresyon'a Dayalı Bulanık Sınıflandırma Modeli** olarak adlandırılmıştır. Bulanık regresyon katsayılarını tahmin etmek için kullanılan Doğrusal Program (LP) aşağıda gösterilmektedir.

Min. $J = \sum_{i=1}^{N}\sum_{j=0}^{M} c_j |x_{ij}|$

s.t.

$$\sum_{j=0}^{M} m_j x_{ij} + (1-H)\sum_{j=0}^{M} c_j |x_{ij}| \geq \hat{p}_i(y_i = 1|\mathbf{x_i}) \quad \text{for } i = 1,\dots,N$$

$$\sum_{j=0}^{M} m_j x_{ij} - (1-H)\sum_{j=0}^{M} c_j |x_{ij}| \leq \hat{p}_i(y_i = 1|\mathbf{x_i}) \quad \text{for } i = 1,\dots,N$$

$c_j \geq 0$, m_j serbest for $j = 0,\dots,M$

x_{ij} : i'inci gözlemin j'inci bağımsız değişkeninin değeri
y_i : i'inci gözlemin çıktı değeri
H : uyum derecesi
m_j : j'inci bulanık regresyon katsayısının merkezi
c_j : j'inci bulanık regresyon katsayısının genişliği
M : bağımsız değişkenlerin sayısı
N : gözlem sayısı
$\hat{p}_i(y_i = 1|\mathbf{x_i})$: i'inci müşterinin verilen $\mathbf{x_i}$ girdi vektörü için sürücü koltuğundan memnun olma olasılığı tahminidir.

Görüldüğü gibi Tanaka'nın FLR yaklaşımına dayalı bulanık sınıflandırma yaklaşımında Tanaka'nın yöntemi (1989) birebir kullanılmış yalnızca kesikli olan bağımlı değişkenin LR yönteminden faydalanılarak sürekli hale getirilmesi sağlanmıştır.

Yukarıdaki LP'nin çözümü sonucunda elde edilen bulanık regresyon modeli kullanılarak verilen müşteri özelliklerine (demografik özellikler, araç kullanım alışkanlıkları, v.s.) göre bir müşterinin sürücü koltuğundan memnun kalma olasılığı bulanık bir sayı olarak tahmin edilmektedir. Tahmin edilen bulanık aralıkların merkezleri kullanılarak müşterinin memnuniyet derecesinin hangi sınıfa ait olduğu tahmin edilmektedir. Çünkü bulanık regresyon analizinde bulanık katsayıların simetrik olması durumunda merkez değerinin bağımlı değişkeni en iyi şekilde yorumlayan değer olduğu ispatlanmıştır (Wang ve Tsaur, 2000).

İstatistiksel bir yöntem olan LR yönteminde oranlar En Çok Olabilirlik Tahmin Edicisi (MLE) yöntemi kullanılarak modellenirken bu yöntemde oranlar olabilirlik teorisine dayanan FLR yöntemi ile

modellenmektedir. Böylelikle, rassallıktan ve bulanıklıktan kaynaklanan belirsizliklerin etkisinin birlikte ele alınması mümkün olmaktadır.

Geliştirilen bu yöntemin müşteri memnuniyeti verisi üzerindeki başarımı yüksek olmuştur (PCC=0.79, AUC=0.73, kesinlik=0.68, duyarlılık=0.60).

Bu yaklaşımın diğer bulanık sınıflandırma yöntemleri ile karşılaştırılması sonuçlarını da değerlendirecek bir uluslararası bir makale hazırlığı devam etmektedir.

3.4 Bulanık Tahmin/Sınıflandırma için Parametrik Olmayan İyileştirilmiş Bulanık Tahmin/Sınıflandırıcı Fonksiyon Yaklaşımları

Çelikyılmaz (2008) tarafından geliştirilen İyileştirilmiş Bulanık Fonksiyon (IFF) ve İyileştirilmiş Bulanık Sınıflandırma Fonksiyonu (IFCF) yöntemlerinin kalite verilerine uygulanmasında karşılaşılan güçlükler nedeniyle ilgili yöntemler iyileştirilmiştir. Bu kısımda söz konusu yöntemler ve yapılan iyileştirmeler aktarılmaktadır.

Çelikyılmaz (2008), Bulanık Tahmin/Sınıflandırma Fonksiyonu (FF/FCF) yaklaşımlarında yeni bağımsız değişkenler olarak kullanılan üyelik değerlerinin daha güçlü tahmin ediciler haline getirilmesinin amaçlandığı İyileştirilmiş Bulanık Kümeleme (IFC) yöntemini geliştirmiştir. Bulanık c-Ortalamalar (FCM) kümeleme yöntemi yerine IFC yöntemi kullanılarak hesaplanan üyelik değerlerinin yeni bağımsız değişkenler olarak kullanıldığı tahmin yöntemine İyileştirilmiş Bulanık Fonksiyon (IFF), sınıflandırma yöntemine ise İyileştirilmiş Bulanık Sınıflandırma Fonksiyonu (IFCF) adını vermiştir (bkz. Kısım 2.2.2.4.5).

IFC kümeleme yaklaşımında üyelik değerleri elde edilirken, yalnızca üyelik değerleri ve dönüşümleri kullanılarak problemin çeşidine göre bağımlı değişken bir tahmin veya sınıflandırma yöntemi kullanılarak tahmin edilmeye çalışılmaktadır. Tahmin problemlerinde En Küçük Kareler (LS) ile Destek Vektör Makineleri (SVM), Sınıflandırma problemlerinde ise LR, SVM ve Sinir Ağları (NN) yöntemlerinin kullanılması önerilmiştir (Çelikyılmaz, 2008). Bu yöntemde, Bulanık c-Ortalamalar yönteminde kullanılan amaç fonksiyonuna yalnızca üyelik değerleri ve çeşitli dönüşümleri kullanılarak tahmin edilen bağımlı değişken ve gözlemlenen bağımlı değişken değerinin hata karesini içeren bir terim eklenerek elde edilecek üyelik değerlerinin tahminleme gücü artırılmaya çalışılmaktadır. Böylece, üyelik değerleri hesaplanırken FCM'de olduğu gibi yalnızca girdilerin birbirine olan uzaklığı değil, çıktı değişken ile ilişkisi de dikkate alınmaktadır.

$$J_m^{IFC} = \underbrace{\sum_{i=1}^{n}\sum_{k=1}^{N} \mu_{ik}^m d_{ik}^2}_{\text{FCM}} + \sum_{i=1}^{n}\sum_{k=1}^{N} \mu_{ik}^m (y_k - f(\tau_{ik}))^2$$

μ_{ik} : k'inci gözlemin i'inci kümeye üyelik değeri,
d_{ik} : k'inci gözlemin i'inci küme merkezine uzaklığı,
y_k : k'inci gözlemin gözlemlenen değerin,
$f(\tau_{ik})$: verilen τ_{ik} girdi vektörü için k'inci gözlemin i'inci küme için tahmini değeri
τ_{ik} :elemanları k'inci gözlemin i'inci kümeye üyelik değerleri ve seçilen dönüşümlerinden oluşan satır vektörü,
m : bulanıklık derecesi
n : küme sayısı
N : gözlem sayısı

Bir algoritma kullanılarak yukarıda verilen amaç fonksiyonu aşağıdaki kısıtlar altında en küçüklenerek üyelik değerleri hesaplanmaktadır.

$$\sum_{i=1}^{n} \mu_{ik} = 1$$

$$0 \ll \sum_{i=1}^{N} \mu_{ik} \ll N$$
$$\mu_{ik} \in [0, 1]$$

Bu algoritmaya göre IFC yöntemi ile veri kümesinin kümelenebilmesi için bir döngü içerisinde her iterasyonda bir önceki iterasyonda hesaplanan üyelik değerleri ve dönüşümleri ile çıktı değişkeninin tahmin edilmeye çalışıldığı bir model kurularak yeni üyelik değerlerinin elde edilebilmesi gerekmektedir. Herhangi bir iterasyonda model kurulamadığı durumda algoritma sonlandırılmakta ve kümeleme işlemi gerçekleştirilememektedir. Özellikle, gerçek bağımsız değişkenler kullanılmayarak çıktı değişkenin yalnızca hesaplanan üyelik değerleri ve onların dönüşümleri kullanılarak tahmin edilmeye çalışıldığı bu yöntemde, parametrik yöntemler kullanarak model kurmak her zaman mümkün olamamaktadır. Ayrıca parametrik yöntemler modelin geçerliliğinin testi için birçok varsayımın sağlanıp sağlanmadığının kontrolünü gerektirmektedir. Ancak, bir döngü içerisinde yakınsama sağlanana kadar her iterasyonda bu kontrolleri gerçekleştirmek aşırı zaman kaybı ve yöntemin kullanışsız hale gelmesine neden olacaktır. Bu varsayımların kontrollerinin yapılmaması ise geçersiz olan modellerin kullanılmasına neden olabilecektir. Dolayısıyla, yöntem doğası gereği parametrik olmayan yöntemlerin kullanımına daha elverişlidir. Bu nedenle, IFC yönteminde parametrik bir yöntem kullanmak yerine parametrik olmayan bir yöntem olan Çok Değişkenli Uyarlanabilir Regresyon Eğrilerini (MARS)'ın kullanılmasına karar verilmiş ve bu yaklaşım Parametrik Olmayan İyileştirilmiş Bulanık Kümeleme (NIFC) olarak adlandırılmıştır (Özer ve diğ., 2009b). Kümeleme aşamasında MARS'ın kullanılması ile hem sınıflandırma hem de tahmin problemleri için model kurulması, optimum değişken dönüşümleri ve etkileşimlerinin seçimi otomatik hale getirilmiştir. Böylece, döngü içerisindeki her iterasyonda çıktı değişkeni tahminleyen bir model kurulması sağlanarak algoritmanın başarıyla sonuca ulaşması ve kümeleme işleminin gerçekleştirilmesi sağlanmıştır.

NIFC yöntemi ile kümeleme işlemi gerçekleştirildikten sonraki aşamada FF ve FCF yönteminde olduğu gibi hesaplanan üyelik değerlerinin gerçek bağımsız değişkenlere ilave yeni bağımsız değişkenler olarak kullanıldığı bulanık tahmin ve sınıflandırma fonksiyonlarının oluşturulması gerekmektedir. Bu fonksiyonlar oluşturulurken sınıflandırma yöntemi olarak LR veya SVM ve tahmin yöntemi olarak ise LS veya SVM yöntemlerinin kullanılması önerilmiştir (Çelikyılmaz, 2008). Kümeleme aşamasında IFC yerine NIFC yönteminin kullanılmasını öneren bulanık tahmin ve sınıflandırma yöntemleri ayrı ayrı Parametrik Olmayan İyileştirilmiş Bulanık Fonksiyon (NIFF) (Kılıç, 2009) ve Parametrik Olmayan İyileştirilmiş Bulanık Sınıflandırma Fonksiyonu (NIFCF) olarak adlandırılmıştır (Özer, 2009).

Döküm hatası Sınıflandırma ve leri kullanılarak modellenmişti. Bu raporda ise geliştirilen NIFF ve NIFCF yöntemleri Döküm Hatası Tahmin/Sınıflandırma ve Koltuktan Memnuniyet sınıflandırma verilerine uygulanmıştır. NIFF yönteminde modelleme yöntemi olarak LS, NIFCF yönteminde ise modelleme yöntemi olarak LR kullanılmıştır. Ayrıca karşılaştırma amacıyla, tahmin problemleri için kullanılan FF yöntemi döküm hata oranı tahmin verisine, FCF yöntemi ise koltuktan memnuniyet verisine uygulanmıştır. Elde edilen çapraz doğrulama sonuçlarının ortalamaları incelendiğinde NIFCF ve NIFF yöntemlerinin diğer bulanık yöntemler olan FCF ve FF'den daha iyi sonuçlar verdiği görülmüştür.

Bu iyileştirmeler ile ilgili bir ulusalararası makale hazırlığı devam etmektedir.

3.5 Mahalanobis Taguchi Sistemi ile Sınıflandırmada Sınır Değerin Belirlenmesi

Mahalanobis Taguchi Sistemi ile sınıflandırma yapılırken, normal ve anormal sınıfların ayrımını yapacak sınır değerinin belirlenmesi büyük önem taşımaktadır. MTS yönteminde sınır değeri olarak sabit bir değer belirlenememektedir ve her problem için farklı bir sınır değerinin belirlenmesi zorunluluğu bulunmaktadır. Bu çalışmada önerilen yöntem, sınır değerinin önem verilen sınıflandırma ölçüsünün en iyi değerini alacak şekilde seçilmesidir. Önerilen yöntemde Tam Kapsamlı Arama Yöntemi kullanılarak olası sınır değerleri denenmekte ve her bir sınır değeri için performans ölçüleri hesaplanmaktadır. Seçilen performans ölçüsüne göre sınır değeri değişiklik gösterebilir. Örnek olarak, hatalı örneklerin hatasız olarak yanlış sınıflandırılmasının maliyetinin yüksek olduğu bir üretim ortamında, hatalı örneklerin tahmin oranını eniyileyecek sınır değeri belirlenmek istenilebilir.

Önerilen yöntemde veri grubu tabakalı örneklem yöntemi kullanılarak 3 gruba ayrılmış, sınır değeri eğitim grubu üzerinden elde edilerek test grubu üzerinde denenmiş ve Doğru Sınıflandırma Oranı, $F_{0.5}$, F_1, F_2 ve G-ortalama gibi performans ölçüleri ölçülmüştür.

Kalite iyileştirme çalışmalarında veri gruplarının dengesiz olması sık karşılaşılan bir durumdur. Bu gibi durumlarda Doğru Sınıflandırma Oranı gibi performans ölçülerü yanıltıcı olabilir. Yüksek oranda dengesiz veri gruplarında azınlık grubunun yanlış tahmin edilmesi Doğru Sınıflandırma Oranınında büyük değişikliklere yol açmamaktadır. Bunun yerine, sınıfların tahmin oranlarını ayrı ayrı değerlendirmek daha doğru olmaktadır. G-ortalama, dengesiz veri gruplarında sıklıkla kullanılan bir performans ölçüsüdür. İki grubun doğru sınıflandırma oranlarının geometrik ortalaması olan g-ortalama, iki grubun da sınıflandırılma oranlarına eşit oranda önem verdiği için önerilmektedir.

Sınır değeri belirlenmesi için geliştirdiğimiz yöntem aşağıdaki adımlardan oluşur:

1. Verinin test ve eğitim kümelerine ayrılması.

Veriler n-katlı tabakalı örneklem yöntemine göre test ve eğitim kümelerine ayrılır. Kat sayısı n, veri grubunun büyüklüğüne göre değişmektedir. Örneklerin gruplara rastgele dağılılıyor olmasının etkilerinin ortadan kaldırılması için de birden fazla yineleme yapılması önerillir.

2. MTS modelinin kurulması

MTS modeli eğitim seti kullanılarak kurulur. Kurulan modelin çıktısı olarak örneklerin Mahalanobis Uzaklıkları elde edilir.

3. Enbüyüklenecek performans ölçüsünün belirlenmesi

Elde edilmek istenilen sınıflandırma özelliğine göre bir performans ölçüsü seçilmelidir. Örnek olarak, tüm örneklerin doğru sınıflandırma oranı önemli ise, Doğru Sınıflandırma Oranı, hatalı örneklerin doğru sınıflandırılması önemli ise anormal sınıfın Doğru Sınıflandırma Oranı önemli performans ölçüsü olarak seçilebilir.

4. Olası sınır değerlerinin hesaplanmış Mahalanobis Uzaklıkları üzerinde denenmesi ve performans ölçülerinin hesaplanması.

Olası aralıktaki sınır değerleri adım adım artırılarak denenerek her bir sınır değeri için seçilen performans ölçüsü hesaplanır.

5. Seçilen performans ölçüsünü enbüyükleyecek sınır değerinin belirlenmesi

Olası aralıktaki her sınır değeri için performans ölçüsü hesaplandıktan sonra, her bir eğitim kümesinde en yüksek performans ölçüsünü veren sınır değeri seçilir. Eğitim kümeleri için bulunan sınır değerlerinin ortalaması alınarak kullanılacak sınır değeri bulunur.

Geliştirilen bu yöntem çeşitli veri kümeleri üzerinde, farklı performans ölçüleri kullanılarak uygulanmıştır (Yenidünya, 2009). Yapılan çalışma sonucunda kalite verilerinde olduğu gibi dengesiz veri kümelerinde g-ortalama performans ölçüsü baz alınarak seçilen sınır değerlerinin test sınıfında da iyi sınıflandırma performansı gösterdiği görülmüştür.

3.6 Çoklu Sınıflandırma için Mahalanobis Taguchi Sistemi Tabanlı Yaklaşımlar

Kalite verilerinin çıktı değişkenleri tipik olarak iki sınıflı (belli bir hatayı içeren veya içermeyen, hatalı veya hatasız) olmakla birlikte sınıf sayısı ikiden fazla da olabilmektedir (müşteri memnuniyetinin derecesi gibi). İkili sınıflandırma problemlerine nazaran çoklu sınıflandırma problemleri daha az çalışılmıştır (Hsu ve Lin, 2002). Bu projede, sayısı ve performansı kısıtlı olan bu yöntemlere alternatif olarak ikili sınıflandırmada başarılı bulduğumuz Mahalanobis Taguchi Sisteminin çoklu sınıflandırma için çeşitli uyarlamaları geliştirilmiştir.

Literatürde, çoklu sınıflandırma problemleri için farklı yaklaşımlar bulunmaktadır (Su ve Hsiao, 2009). İlk yaklaşım, çoklu sınıflandırma problemlerini çözmek için kullanılacak ikili sınıflandırma yönteminde herhangi bir değişiklik gerektirmez. Bu yaklaşımla problemi çözmek için modeli bir kez çalıştırmak yeterli olduğundan problemi çözmek hızlıdır. Ancak, bu yaklaşımı kullanan yöntemlerle (örneğin, Mahalanobis uzaklığı sınıflandırıcısı, karar ağacı) pek sıklıkla karşılaşılmaz. İkinci yaklaşım ise, iki sınıflı bir algoritma üzerinde değişiklik yapılarak çoklu sınıflandırma probleminin çözülmesidir (örneğin bazı Destek Vektör Makina (SVM) algoritmaları). Son yaklaşım ise çoklu sınıflandırma probleminin iki sınıflı problemler haline dönüştürülerek problemin çözülmesidir (Ou v.d., 2004): (1) “bire-bir” yaklaşımı tüm sınıfları çiftler halinde ele alır. Ancak,

problemi bu şekilde çözmek için, algoritmayı L(L–1)/2 kez (burada L sınıf sayısını ifade eder) çalıştırmak gerekir (Friedman, 1996). (2) "bire-bütün" yaklaşımında ise, her sınıf üzerinde ayrı bir model oluşturulur. Buna göre, seçilen sınıf dışındaki sınıflar bir bütün halinde düşünülerek, L tane sınıf için, L tane model oluşturulur (Ding et al., 2001). Probleme ait veri bir bütün halinde ele alınmasına rağmen, bire-bir yaklaşıma göre daha fazla hesaplama zamanına ihtiyaç duyulur. Bu yaklaşımların detaylı karşılaştırmaları Chin (1998) çalışmasında sunulmuştur (Hsu ve Lin, 2004).

MTS yaklaşımı esas olarak ikili sınıflandırma için tasarlanmıştır (Mahalanobis, 1936). MTS'nin çoklu sınıflandırma problemlerine nasıl uyarlanacağı ise devam eden bir araştırma konusudur. İlk olarak, Su ve Hsiao (2009) değişken ağırlıklı bir yöntem sunmuşlardır. Önerilen çoklu sınıflandırma yönteminde, uzaklık hesaplaması için, Gram-Schmidt Mahalanobis uzaklığı (GS-MD) kullanılmıştır. Ayrıca, sınıfsal doğruluk oranları dikkate alınarak önerilen çok sınıflı MTS yaklaşımı Mahalanobis uzaklığı sınıflayıcısı (MDC) ile karşılaştırılmıştır.

İkinci olarak, Ayhan (2009) "bire-bütün" yaklaşımı ile MTS modelini birlikte kullanılarak çoklu sınıflandırma problemlerini çözen yeni yöntemler geliştirmiştir: (1) Çok Sınıflı MTS (MMTS) isimli yaklaşım, basit anlamda, her sınıf için ayrı bir MTS modelinin geliştirilmesini ve yeni bir gözlemin hangi sınıfa daha yakın ise ona atanmasını öngörür. (2) Değişken Ağırlıklı Çoklu MTS-I (FWMMTS-I) isimli diğer bir yöntemde ise MD'nin değişkenlere dayalı eşit ağırlıklı toplam alma özelliği hafifletilmiştir. Bu şekilde, gürültü değişkenlerinin sıfıra yakın ağırlıklarla temsil edilmesi sağlanarak, MD hesaplamasında gürültü değişkenlerinin diğer değişkenleri gizlemesi engellenmiştir. (3) Su ve Hsiao (2009) çalışmasında önerilen çok sınıflı MTS yönteminde GS-MD hesaplaması yerine MD kullanılarak diğer bir değişken ağırlıklı çoklu sınıflandırma yöntemi (FWMMTS-II) geliştirilmiştir. Ayhan (2009) çalışmasında tüm yöntemler tabakalı çapraz doğrulama yaklaşımı kullanılarak sekiz farklı çok sınıflı veri kümesinde sınanmıştır. Su ve Hsiao (2009) çalışmasında geçen, sınama verilerini sınıflara atamadaki ortalama başarı oranı (BCA) ve toplam doğruluk oranı (PCC) iki performans ölçütü olarak ele alınmıştır. ANOVA ve Bonferonni çoklu kıyaslama yöntemleri ile yapılan karşılaştırmaların sonuçlarına göre, en yüksek başarımı (ortalama BCA=%68 ve ortalama PCC=%71) MMTS ve FWMMTS-I ve göstermiştir Ayhan (2009).

Bu çalışmanın daha fazla veri kümesi üzerinde sınanması ve MTS dışındaki bazı çoklu sınıflayıcılarla da karşılaştırılması yapılarak uluslararası bir makale hazırlanması yönünde çalışmalar devam etmektedir.

3.7 İkili Çıktı Değişkeni ile Parametre Optimizasyonu için Mahalanobis Taguchi Sistemi Tabanlı Bir Yaklaşım

Nitel sonuçlu ürün ve/veya süreç kalitesini eniyileme çalışmalarında kullanılabilecek mevcut yöntemler önemli eleştiriler almıştır. Huang (2005) bu yöntemlerden bazılarını kıyaslamış ve bir örnek üzerinde zayıf yanlarını ortaya koymuştur. Öte yandan Erdural (2006), nitel bir çıktı değişkeni ile parametre optimizasyonu için geliştirdiği Lojistik Regresyon temelli yaklaşımı, projemizde çalıştığımız döküm sürecine benzer bir süreç üzerinde sınamış ve başarılı olduğunu göstermiştir. Ancak lojistik regresyon modellerinin geliştirilmesi belli ölçüde uzmanlık gerektirdiğinden bu yaklaşımın endüstride kullanımı kısıtlı olabilir. Bu nedenle, projemizde, uygulaması daha kolay olabilecek parametre optimizasyonu yöntemlerinin geliştirilmesi için çalışmalar yapılmıştır. Bu kapsamda, uygulaması kolay olan ve sınıflandırma amacına yönelik olarak geliştirilmiş olan MTS'in, ilk defa, parametre optimizasyonu amacıyla kullanımına yönelik bir yöntem geliştirilmiştir.

Geliştirilen bu yeni yaklaşıma göre ürün ve/veya süreç değişkenlerinin en iyi sonucu veren değerlerinin bulunması için önce MTS ile değişkenlerin sonuç üzerindeki etkilerini tanımlayan bir sınıflandırma modeli kurulur. Bunun için, istenen sonucu (örneğin hatasız sonuç) içeren sınıf referans sınıf olarak belirlenir. Belirlenen bu sınıfa göre MTS modeli, yani *MD* fonksiyonu elde edilir. Bu fonksiyonun en küçük değerini veren ürün ve/veya süreç değişken değerleri, uygun bir matematiksel programlama yöntemi yardımıyla bulunur.

MD eşitliğinden anlaşıldığı üzere, en küçük Mahalanobis uzaklığı, değişkenlerin, referans grubun değişkenlerinin ortalama değerlerini aldığı durumda (**x=u**) elde edilir. Bu durumda *MD* sıfıra eşit olur. Ancak, değişkenler üzerinde olabilecek kısıtlamalar yüzünden değişkenlerin referans grup ortalamalarını almaları mümkün olmayabilir. Örneğin, değişkenler sürekli değil de ikili, tamsayılı, ya da sıralı kategorik olabilir. Başka bir örnekte, referans grubun ortalama değerleri değişkenlerin işletim sınırlarının dışında olabilir. Değişkenlerin referans grup ortalamalarını almasının mümkün olmadığı durumlar için mevcut kısıtları hesaba katabilen eniyileme algoritmalarına gerek vardır.

Mahalanobis uzaklığının amaç fonksiyonu olduğu ve değişkenlerle ilgili kısıtların tanımlandığı basit bir matematiksel programlama modeli aşağıdaki şekilde tanımlanabilir:

Enküçükle $MD = \frac{1}{p}(\mathbf{x}-\mathbf{u})'\sum^{-1}(\mathbf{x}-\mathbf{u})$

Kısıtlar:

$$a_i \le x_i \le b_i, \qquad i = 1,2,...,p$$

Burada, a_i ve b_i, sırasıyla, *i*'inci değişkenin alt ve üst sınırlarını göstermektedir.

Modelin amaç fonksiyonu olarak kullanılan *MD* eşitliği ikinci dereceden bir fonksiyondur. Amaç fonksiyonunun ikinci dereceden ancak kısıtların doğrusal olduğu koşulda, eniyileme problemi doğrusal olmayan programlamanın özel bir sınıfı (karesel programlama) haline gelmektedir. Bu tür problemlerin çözümü için başarılı algoritmalar mevcuttur. Bu problem çözüldüğünde , seçilmiş olan referans sınıfın ağırlık merkezine en yakın ve kısıtları sağlayan en iyi değişken değerleri bulunabilir.

Geliştirilen bu yöntem çeşitli veri kümeleri üzerinde uygulanmış ve amaç fonksiyonu yerine uygun bir lojistik regresyon modelinin seçildiği yöntem ile karşılaştırılmıştır (Yenidünya, 2009). Yapılan çalışmada, MTS yönteminin LR yöntemi ile karşılaştırılabilir sonuçlar verdiği görülmüştür.

Bu yöntem ile ilgili uluslararası bir makale hazırlığı devam etmektedir.

3.8 Çekicilik Fonksiyonlarının Optimizasyonu İçin Pürüzlü Optimizasyon Yaklaşımları

Çekicilik fonksiyonlarının pürüzlü yapısının yumuşatılarak optimize edilmesine dayanan değiştirilmiş çekicilik fonksiyonu yöntemi ile ilgili sıkıntılar (Castillo v.d., 1996) makalesinde belirtilmiştir. Bu yaklaşıma alternatif olması amacıyla, çekicilik fonksiyonun optimizasyonunu zorlaştıran ve yeni yaklaşımlar gerektiren pürüzlü yapısıyla başedebilen iki yaklaşım izlenmiştir:

3.8.1 Mevcut pürüzlü optimizasyon yaklaşımlarının kullanılması

Pürüzlü optimizasyon metodları denilince akla ilk olarak **altgradyan yöntemi** (Shor, 1985) **demet yöntemi** (Vlcek, 1997), **ayrık gradyan yöntemi** (Bagirov,1999) ve **değiştirilmiş altgradyan yöntemi** (Burachik,2006) gelmektedir.

Tablo 3-7 Pürüzlü optimizasyon yöntemleri ve özellikleri

Yöntem	**Kullandığı türev bilgisi**	**Amaç ve kısıt fonksiyonları için gereklilikler**
Altgradyan, Demet	Altdifferensiyal, Clarke Altdifferensiyal	İç Bükey, Lipschitz Sürekli
Ayrık Gradyan	Yarıalt diferensiyal	Yönlü Türevlenebilir

Çekicilik fonksiyonu optimizasyon problemi, altgradyan, demet ve ayrık gradyan yöntemlerinin varsayımlarına uymamaktadır (Bkz. Tablo 3-7). Bu yöntemler içbükeylik gibi bizim amaç fonksiyonumuzda bulunmayan gereklilikler istemektedir, değiştirilmiş altgradyan yöntemi (MSG) sadece karar değişkenleri kümesinin kompakt ve tüm fonksiyonların sürekli olması gerekliliğini istememektedir. MSG yöntemini problemimize uygulamak için amaç fonksiyonumuzda bir düzenleme yapılmıştır. MSG yöntemi çekicilik fonksiyonu uygulamalarında yaygın olarak kullanılan 6 yanıtlı bir yarı-iletken üretim sürecine ait çok amaçlı optimizasyon probleminde (bkz. Castillo v.d.,1996) test edilmiştir. MSG bu optimizasyon probleminin içbükey dualinin yazılarak, bu dualin çözümüne dayanmaktadır. Toplam çekicilik fonksiyonun optimizasyon problemi şu şekilde yazılmalıdır:

$$\begin{aligned} &\max \ D(x,z) \\ &s.t. \quad x \in X \\ &\qquad Y_j(x) \in \left[l_j, u_j\right] \\ &\qquad h_j(z) = 0 \quad (j = 1,2,\ldots,6) \end{aligned} \qquad (1)$$

Faktör kümesi *X* kompakt ve eşitlik kısıtı $h_j(z) := z_j - z_j^2$ (*j=1,2,...,6*) olarak tanımlanmıştır, $h := (h_1, h_2, \ldots, h_6)^T$. Toplam çekilik fonksiyonu D(x,z), $x \in X$ ve $z = (z_1, z_2, \ldots, z_6)^T$ olmak üzere 6 yanıtın tekil çekilik fonksiyonları için

$$D(x,z) = \left[\prod_{j=1}^{6} d_j(Y_j(x)) \right]^{\frac{1}{6}}$$

şeklindedir. Burada MSG algoritmasını kullanabilmek için $\lfloor l_j, u_j \rfloor$ aralığında tanımlı iki taraflı tekil çekilik fonksiyonu şu şekilde yazılmalıdır: $d_j(Y_j(x))$ (*j=1,2,...,6*) $d_j(Y_j(x)) = z_j d_{j1}(Y_j(x)) + (1 - z_j) d_{j2}(Y_j(x))$. Burada $d_{j1}(Y_j(x))$ tekil çekilik fonksiyonu $d_j(Y_j(x))$'nin $[\, l_j, t_j \,)$ aralığındaki parçası, $d_{j2}(Y_j(x))$ ise $[\, t_j, u_j \,]$ aralığındaki parçasıdır, t_j çekicilik fonksiyonun hedef değerini aldığı noktadır ve $l_j < t_j < u_j$ şeklindedir.

Yukarıdaki (1) probleminin eşiti olan minimizasyon probleminin keskin artırılmış Lagrangian fonksiyonunu şu şekilde olmaktadır: $L(x,z,u,c) := -D(x,z) + c\left\|h(x,z)\right\|_2 - u^T h(x,z)$.

Burada $D : \Re^3 x \Re^6 \to \Re$ ve $h : \Re^3 x \Re^6 \to \Re^6$ şeklinde tanımlanmış fonksiyonlar, vektör $u \in \Re^6$ ve positive skalar $c > 0$ problemimizin dual parametreleridir. Bu Lagrangian fonksiyonuna dayanan dual fonksiyon $H(u,c)$, Lagrangian'in minimumu olarak tanımlanmıştır: Dual problemimiz dual fonksiyonunun maksimizasyonu olmaktadır.

6 yanıtlı yarı iletken süreci optimizasyon problemimizin bu şekilde yazılmış dual problemi Core 2 Duo T7300 2.00 GHz 2.00 GB (64-bit) bilgisayar üzerinde GAMS v.23.0.2'da kodlanmış, BARON v.8.1.5 çözücüsü ile çözülmüştür. Program çözümü fesable çözümler arasında önişlem yapma esnasında bulmuştur. Bulunan optimal x=(x(1), x(2), x(3)) ve z=(z(1), z(2), z(3), z(4), z(5), z(6)) değerleri sırasıyla Tablo 3-8 ve Tablo 3-9'daki gibidir.

Tablo 3-8 x karar değişkeni değerleri

x(1)	x(2)	x(3)
0.113	0.5475	0.3009

Tablo 3-9 z karar değişkeni değerleri

z(1)	z(2)	z(3	z(4)	z(5)	z(6
0.308	0.247	0.247	0.312	0.247	0.247

Bulunan bu optimal değerler kullanılarak toplam çekicilik fonksiyonu hesaplandığında sonuç 0. 2356 olmaktadır. Bu sonucu (Castillo v.d., 1996)'nin makalesinde verilen yumuşatılmış çekilik fonksiyonun Genelleştirilmiş Azaltılmış Gradyan (GRG) ve Hooke Jeeves (HJ) yöntemi ile karşılaştırılmasında elde edilen sonuçlar şöyledir.

Tablo 3-10 GRG, HJ ve MSG yöntemlerinin çekicilik fonksiyonu optimizasyonu probleminde bulduğu optimal değerler

	x(1)	x(2)	x(3)	Toplam Çekicilik
GRG	0.1039	1.0	0.7987	0.3061
HJ	0.1078	1.0	0.7973	0.3076
MSG	0.113	0.5475	0.3009	0.2356

Mevcut durumda elde edilen sonuç yumuşatılmış çekicilik fonksiyonları yönteminden daha kötü görünmüştür. Burada MSG yöntemi bir lokal optimale yakalanmış, global optimale gidememiştir. Bu sonucun izlenen yöntemden mi yoksa optimizasyon algoritmasından mı kaynaklandığı çalışılmaktadır. Bizim yaklaşımımız çekicilik fonksiyonunun yumuşatılmasını gerektirmemektedir, fonksiyonların formülasyonu yukarıda önerdiğimiz şekilde *z* tamsayı katsayıları ile yazıldıktan sonra MSG algoritması uygulanabilmektedir. Bu çalışmalar tamamlanınca MSG yönteminin düzenlenmiş çekicilik fonksiyonları üzerindeki başarımı kararlaştırılacaktır. Bu kısım ile ilgili uluslar arası makale hazırlığı devam etmektedir.

3.8.2 Yeni pürüzlü optimizasyon yöntemleri geliştirilmesi

Bu kısımda çekicilik fonksiyonlarının pürüzlü yapısının, pürüzlü ve sürekli optimizasyon yaklaşımları ile analizi çalışmalarımız yeralmaktadır. Burada geliştirdiğimiz yaklaşımlar iki ana başlıkta toplanabilir: analitik ve topolojik yaklaşım ve çekicilik fonksiyonlarının bileşke fonksiyonları olarak incelenmesine dayanan yaklaşım. Analitik ve topolojik yaklaşımda sunudumuz iki aşamalı optimizasyon formülasyonu çekicilik fonksiyonu optimizasyonu problemine önemli bir yenilik getirmektedir. İki aşamalı problemin alt seviye aşaması için bir metoloji önerilmiştir. Üst seviye problem bir temsil problemi olmaktadır ve bu problemin çözülmesi için yürütülen çalışmalar proje personeli Başak Öztürk'ün doktora tezi kapsamında devam etmektedir (Öztürk, 2010). Burada sunduğumuz yaklaşım, 5. proje gelişme raporunda sunulan ve yapısı çekılik fonksiyonlarına benzeyen parçalı-pürüzlü fonksiyonlar için kullanılabilecek Genelleştirilmiş Çekicilik Fonksiyonu (GÇF) yaklaşımının çekicilik fonksiyonu optimizasyonu problemi için özelleştirilmiş halidir.

3.8.2.1 Analitik ve Topolojik Yaklaşım

Çekicilik fonksiyonları ile çözülen çok cevaplı optimizasyon problemi

$$\max_{x \in X \subseteq R^n} D(x) \qquad (1)$$

olarak yazılabilir. Burada $D(x) = \left[\prod_{j=1}^{m} w_j d_j(Y_j(x)) \right]^{1/m}$ şeklindedir.

Optimizasyon problemini yaygın olarak yapıldığı gibi minimizasyon problemine çevirip, gerekli düzenlemeler yapıldıktan sonra elde edilen amaç fonksiyonuyla yazılan optimizasyon problem bir vektör optimizasyon problemi olmaktadır:

$$\begin{cases} \min & w_1 f_1(y_1) + w_2 f_2(y_2) + \ldots + w_m f_m(y_m) \\ s.t & y_j \in [l_j, u_j] \subseteq \Re \\ & f_j(y_j) \in (0,1] \\ & j = 1,2,\ldots,m \end{cases}$$

Amaç fonksiyonumuz maksimum tipi fonksiyonların ağırlıklı toplamı olan eklemeli bir fonksiyondur. Optimizasyon problemi her bir j $(j = 1,2,\ldots,m)$ için maksimum fonksiyonların minimizasyonu şeklindedir ve şu şekilde ifade edilebilir:

$$\left\{ \min_{y_j \in [l_j, u_j]} \max_{k \in \{1,2\}} f_{j,k}(y_j) \right.$$

Her bir j $(j = 1,2,\ldots,m)$ için bu problemin çözümü $\bar{y}_j$ olsun, böylece eklemeli amaçlı vektör optimizasyonu problemimizin minimumu $\bar{y} := (\bar{y}_1, \bar{y}_2, \ldots, \bar{y}_m)$ olacaktır. Altseviye problemin çözümü olan bu minimumu bulmak için aşağıdaki metodoloji geliştirilmiştir:

Önerilen Metodoloji:

Yukarıda (1)'de verilen optimizasyon problemindeki toplam çekicilik fonksiyonunda x boyutunu dikkate almadan F_i^{κ} = $\log D(y_i \mid y_1, y_{2,} \ldots, y_{i-1}, y_{i+1}, \ldots, y_m)$ şeklinde yazabiliriz. Burada $\log D$, 2 aralıkta tanımlı olsun. Her aralık κ , $(\kappa = 1,2)$ için F_i^{κ} fonksiyonu $\log D$ ile aynı tanıma sahiptir ve bu aralıkta türevlenebilir. Her bir $\kappa = 1,2$ aralığı için optimizasyon problemimiz:

$\max_{y} F^{\kappa}$ (y)

şeklinde olacaktır. Bu problemin çözümü ($\bar{y}$)$^{\kappa}$ =($(\bar{y})_1^{\kappa}, (\bar{y})_2^{\kappa}, \ldots, (\bar{y})_m^{\kappa}$) olsun. Bütün aralıklar için çözümler F ($\bar{y}^{(1)}$) $\leq$ F ($\bar{y}^{(2)}$) $\leq$ olacak biçimde sıralanabilir.

Algoritma: Alt Seviye Problem

En iyi y çözümü seçilir. Başlangıçta bu $\bar{y}^{(1)}$' dir.

Adım 1.

$\bar{y}^{(1)}$ noktasını veren *x* bulunur:

$\min_{x}$ 1 s.t. Y(x) = $\bar{y}^{(1)}$.

Çözüm var ise durulur. Eğer çözüm yok ise gevşetilmiş problem çözülür:

$\max_{x}$ $F^{(1)}$ s.t. $k_l^{(1)}$ < Y(x) < $k_u^{(1)}$.

Burada $k_l^{(1)}$ ve $k_u^{(1)}$ ilgili aralığın sınırlarıdır. Gevşetilmiş problemin çözümü $\overline{X}^{(1)}$ olsun. Çözüm yok ise Adım 2'ye geçilir.

Adım 2.

Sonraki en iyi y çözümü $\bar{y}^{(2)}$ seçilir ve Adım 1 tekrarlanır.

Eğer $F^{(1)}$(Y($\overline{X}^{(1)}$)) $\geq$ $F^{(2)}$(Y($\overline{X}^{(2)}$)) ve ikinci çözüm gevşetilmemiş problemden elde edilmemiş ise durulur. Değilse $\bar{y}^{(1)} \leftarrow \bar{y}^{(2)}$ eşlemesi yapılarak adımlar tekrar edilir. Bütün çözümler gevşetilmiş problemlerden elde edilmiş ise en büyük *F* değerini veren X seçilir.

Alt seviye problemimizin çözüm seti $\mathrm{Y}(x)$ ve üst seviye problemin çözüm seti X olmak üzere, herhangi bir *y* çözümünü veren $x \in \mathrm{X}$'i bulma üst seviye problemimizdir ve bu problem bir temsil problemidir.

$$\begin{cases} \text{"min"} & w_1 f_1(Y_1(x)) + w_2 f_2(Y_2(x)) + \ldots + w_m f_m(Y_m(x)) \\ s.t & Y(x) \in \mathrm{Y}(x) \\ & x \in X \end{cases}$$

Burada tırnak işaretleri, üst seviye problemde sadece x'e göre minimizasyon yaptığımızdan, alt seviye problem x'in tüm değerleri için uniquely determined optimal çüzüme sahip olmadığında, üst seviye problemin iyi tanımlı olmamasındandır. □

3.8.2.2 Bileşke Fonksiyonlar Olarak Çekicilik Fonksiyonlarının Çözümlenmesi

Çekicilik fonksiyonlarına daha yakından bakacak olursak tekil çekicilik fonksiyonları $d_j(Y_j(x))$'lerin her bir j ($j = 1,2,\ldots,m$) için bileşke fonksiyonlar olduklarını görmekteyiz. Faktör kümesi $X \subset \Re^n$ açık bir küme olmak üzere, bileşke fonksiyonu olarak yazdığımız toplam çekicilik fonksiyonun optimizasyon problemi bir vektör optimizasyon problemi olmaktadır:

$$\begin{cases} \min & w_1 f_1(Y_1(x)) + w_2 f_2(Y_2(x)) + ... + w_m f_m(Y_m(x)) \\ s.t. & x \in X, \\ & Y_j(x) \in [l_j, u_j] \subseteq \Re \qquad (j = 1,2,\ldots,m). \end{cases}$$

Amaç fonksiyonunun ($f(x) := \sum_{j=1}^{m} (f_j \circ Y_j)(x)$) zincir ve toplam kuralları uygulanarak herhangi bir $x \in X \subset \Re^n$ için Clarke subdifferensiyali ve yönlü türevi bulunmaktadır (Burke, 1985):

a. Clarke subdifferensiyali: co içbükey örtü işlemi olmak üzere,

$$\partial f(x) = \partial\left(\sum_{j=1}^{m} w_j (f_j \circ Y_j)(x)\right) \subset co\left\{\sum_{j=1}^{m} w_j \xi_j \mid \xi_j = \vartheta_j \nabla Y_j(x),\ \vartheta_j \in \partial f_j(Y_j(x))\ (j = 1,2,\ldots,m)\right\}.$$

b. Clarke yönlü türevi: $\hat{f}(y) := \sum_{j=1}^{m} f_j(y_j)$ olmak üzere $Y(x)$'e dayalı olarak,

$$\hat{f}(Y(x); D(x)) \le \sum_{j=1}^{m} w_j \left[\hat{f}_j(Y_j(x) + D_j(x)) - \hat{f}_j(Y_j(x))\right]$$

Yukarıdaki (1) nolu eşitliğin bir olası çözüm x^*'da, $\mu_j^l, \mu_j^u, \mu_j^X \ge 0$ ve λ_i^X Lagrange çarpanları olmak üzere gerekli optimallik koşulunu yazarsak:

$$0 \in \sum_{j=1}^{m} w_j \partial f_j(Y_j(x^*)) \nabla^T Y_j(x^*) + \sum_{j \in J_0^l(x^*)} \mu_j^l \nabla^T Y_j(x^*) - \sum_{j \in J_0^u(x^*)} \mu_j^u \nabla^T Y_j(x^*) - \sum_{i \in I^X} \lambda_i^X \nabla h_i(x^*) - \sum_{j \in J^X} \mu_j^X \nabla g_j(x^*). \qquad (2)$$

Burada $J_0^l(x^*)$ aktif indeks kümesi $Y_j(x^*) = l_j$ kısıtının $\nabla(l_j - Y_j(x))\big|_{x^*} = -\nabla Y_j(x^*)$ eşitliğindeki indekslerden, $J_0^u(x^*)$ aktif indeks kümesi $Y_j(x^*) = u_j$ kısıtının $\nabla(Y_j(x) - u_j)\big|_{x^*} = \nabla Y_j(x^*)$ eşitliğindeki indekslerden oluşur.

Yukarıdaki (2) nolu formülle verilen optimallik koşulu, pürüzlü bir yapısı olan çekicilik fonksiyonları için geliştirdiğimiz genelleştirilmiş türev bilgisi ile birlikte pürüzlü optimizasyon yöntemlerinden olan demet metodu (Hiriart-Urruty v.d.,1993)'nun aktif küme stratejisi (Schittkowski, 1992) birleştirilerek çekicilik fonksiyonlarının optimizasyonunda kullanılması hedeflenmektedir. Burada anlattığımız çalışmamız (Akteke-Öztürk, 2009b) bildirisinde sunulmuştur. Ayrıca çekicilik fonksiyonların pürüzlü yapısını analiti/topolojik yaklaşım ve bileşke fonksiyonları yaklaşımları ile analiz ettiğimiz bu çalışmamız ile ilgili uluslararası makale çalışması devam etmektedir.

3.9 Birliktelik Kurallarının Gruplandırılması ve Budanması ile İlgili İyileştirmeler

Veri madenciliğinde önemli bir yer tutan Birliktelik Analizi, veri tabanlarında saklı örüntülerin bulunarak ilginç ilişkilerin ortaya çıkarılmasında önemli rol oynayan bir tekniktir. Bir çok uygulama alanı olan birliktelik analizi özellikle endüstriyel üretim verilerinin incelenmesiyle üretimde karşılaşılan kusurların kök nedenlerinin keşfedilmesinde önemli rol oynayabilmektedir. Bir çok fayda sağlayabilme potansiyeline rağmen, birliktelik analizi uygulamalarında çeşitli problemler de gözlemlenebilmektedir. Uygulamalarda karşılaşılan önemli bir problem keşfedilen çok sayıda birliktelik kuralından ilginç olanların ayıklanmasıdır. Keşfedilen kuralların büyük bir kısmı önemsiz veya başka kuralların tekrarı niteliğinde olabilmektedir. Dolayısıyla, çok sayıda

kuralın incelenip, ilginç kuralların bulunması zaman alıcı olabilmekte ve uygulamada etkinlik azalabilmektedir. Literatürde bu problemin çözümü için çeşitli yöntemler geliştirilmiş olup, kuralların bazı kriterler çerçevesinde düzenlenmesi önem kazanmıştır. Kuralların düzenlenmesi için gruplama ve budama yaklaşımları mevcuttur. Gruplama tekniği birliktelik kurallarını kümeleyerek bunları özetlemeye çalışırken, budalama tekniği ile önemli olmayacağı düşünülen kurallar bulunup sayı azaltılmaya çalışılır. Her iki yöntem de, kuralları faydalı ve anlaşılabilir bir şekle dönüştürmeye çalışır.

Literatürde Berrado ve Runger (2007) tarafından kuralların düzenlenmesi amacıyla "metakurallar" yaklaşımı ortaya atılmıştır. Buna göre, elde edilen kurallara tekrar birliktelik analizi yapılarak kurallar arası ilişkiler, yani metakurallar belirlenir. Metakurallar arasından güven değeri yüksek olanlar seçilir ve belli bir yönteme göre gruplanır. Son olarak gruplar içindeki eşdeğer kurallar arasında destek değeri düşük olanlar elenir. Bu yaklaşım özellikle üretim veri tiplerine yönelik olarak düşünülmüş, çeşitli uygulamalarla yararları gösterilmiştir. Ancak yaklaşımının çeşitli zayıf yönleri bulunmaktadır. Yöntem özellikle seyrek (*sparse*) veri kümeleri için önerilmektedir. Eğer hedef veri tabanı seyrek değilse kurallar arası ilişkilerin sayısı gerektiğinden daha büyük veya küçük olabilmektedir. Büyük veri kümelerinde kuralların gruplanması çok vakit alabilmektedir. Metakurallar yaklaşımı gruplama işleminde tutucu bir yöntem kullanmakta ve dolayısıyla elde edilen kural kümeleri yeterince büyük olmayıp bir çok ilişkili kural kümelenmemiş halde kalabilmektedir. Bunun yanı sıra, kuralların budama aşamasında sadece destek (*support*) değeri hesaba katılmakta ve güven (*confidence*) değeri içerilmemektedir.

Metakurallar yaklaşımının iyileştirilmesi ve farklı veri tabanlarında da etkin kullanılabilmesi için algoritmada bazı değişikliklerin yapılması düşünülmüştür. Bu bağlamda, veri tabanının yoğun/seyrek olmasına bağlı kalınmadan kurallar arası ilişkilerin doğru hesaplanması için bazı çalışmalar yapılmıştır. Oluşturulan kural kümelerine güven değerleri dahil edilerek süper kuralların budanması ya da budanmamasına ait karar süreci kolaylaştırılmıştır. Örnek bir çalışmayla önerilen metodun oluşturduğu kural kümelerinin metakurallar yaklaşımıyla elde edilenlerden daha büyük olduğu ve kuralların daha iyi özetlendiği gözlemlenmiştir.

Metakurallar yaklaşımı üzerinde aşağıda özetlenen üç temel iyileştirme yapılmıştır:

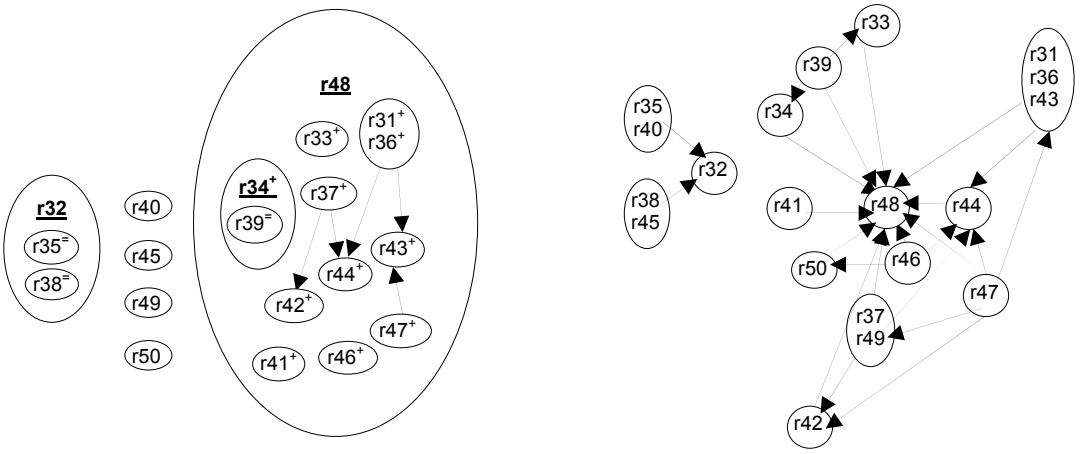
1. Metakuralların güven değerlerinin hesabı iyileştirilmiştir. Önerilen hesaba göre güven, tüm veri üzerinden değil sadece ortak sonucu (*consequent*) sağlayan veriler üzerinden bulunmalıdır. Bu değişiklik ile sadece seyrek değil yoğun veriler için etkili bir güven değeri belirlenmesini sağlamıştır.
2. Metakural yaklaşımı Apriori metodu (Agrawal ve Srikant, 1994) ile ilde edilen kurallara tekrar aynı algoritmanın uygulanmasını gerektirir. Bu nedenle metakuralların çıkarılması, veri kümesi büyüdükçe daha çok zaman almaktadır. Oysa güven değeri %100 olan süper kurallar ve %0 olan geçersiz kurallar için tüm veri kümesini taramadan da eleme yapmak mümkündür. Bu çalışmamızda söz konusu durumlarda nasıl eleme yapılacağı belirlenmiştir. Buna göre, örnek bir veri kümesi üzerinde yapılan tarama sayısında %80 civarında azaltma mümkün olmuştur.
3. Metakurallar yaklaşımı, bulunan metakuralları eşdeğer veya geçersiz olma durumlarına göre yeniden düzenlemektedir. Buna rağmen elde edilen kural grupları çok sayıda ve yorumlanması güç olabilmektedir. Bu çalışmamızda kuralları gruplandırırken daha az sayıda ve büyük gruplar oluşturacak şekilde bir düzenleme geliştirilmiştir. Buna göre, en azından bir kural diğer birisi ile ilişkili ise bu kural o gruba dahil edilir. Bir grup içindeki kurallardan en genel olanı, bunun destek değerinin yanısıra güven değeri de dikkate alınarak belirlenir. En genel kuraldan daha yüksek veya düşük güven değerinde olan kurallar, sırasıyla, + ve – işaretleri ile gösterilir. Böylece kural grupları içinde ilginç olanları bulmak için karar vericiye daha fazla bilgi sağlanmış olur.

Önerdiğimiz yaklaşım UCI Machine Learning Repository'den (http://archive.ics.uci.edu/ml/) elde edilen *Iris* veri kümesi üzerinde denemiş, elde edilen sonuçlar metakurallar yaklaşımının sonuçları ile karşılaştırılmıştır. Veri 150 kayıt, önkoşul olarak kullanılan 4 sürekli değişken ve sonuç olarak kullanılan 3 sınıflı bir kategorik değişkenden oluşmaktadır. Apriori algoritması ile birliktelik kuralları belirlenmiştir. Bundan evvel algoritmanın gerektirdiği üzere sürekli değişkenler kesikli hale getirilmiştir. Destek ve güven sınır değerleri, sırasıyla, %5 ve % 80 olarak seçilmiştir. Toplam 72 kural elde edilmiştir. Her iki yaklaşım da kullanılarak kurallar yeniden düzenlenmiştir. Sonuçlar Tablo 3-11 de özetlenmiştir.

Tablo 3-11 Iris verilerinden elde edilen kurallara ait özellikler

Özellik	Metakurallar yaklaşımı	Önerilen yaklaşım
Kurallar arası ilişki sayısı	284	118
Oluşturulan grup sayısı	13	9
Gruplanan kural sayısı	30	61
Gruplanamayan kural sayısı	42	11
Eşdeğer kural sayısı	21	9

Ayrıca “versicolor” sınıfını sonuç olarak gösteren kuralların her iki yaklaşıma göre nasıl düzenlendiği Şekil 3-1’de gösterilmiştir.



Şekil 3-1 a. Iris verilerinin kural grupları (önerilen yaklaşım) b. Iris verilerinin kural grupları (metakurallar yaklaşımı)

Şekil 3-1.a ve Şekil 3-1.b’de görüldüğü gibi önerdiğimiz yaklaşım metakurallar yaklaşımına göre çok daha az sayıda kurallar arası ilişki ve kural grubu üretmektedir. Bu yaklaşım, ayrıca projemizde topladığımız elektronik kart verilerine de uygulanmış, önceden belirlediğimiz kurallar arasında firmanın ilginç ve önemli bulduğu kuralların, belirlenen kural grupları arasında yer aldığı belirlenmiştir.

Bu ve benzeri yaklaşımların en önemli zayıflığı elenen kurallar arasında karar vericiye ilginç gelebilecek kuralların bulunmasının mümkün olmasıdır.

Sonuç

Bu çalışmada Berrado and Runger (2007) tarafından geliştirilen metakurallar yaklaşımı daha az sayıda ve etkili birliktelik kuralı elde edecek şekilde iyileştirilmiştir. Önerdiğimiz yaklaşım ile özellikle yüksek hacimde ve sık değişen ürün tipi ile üretim yapanların hızlı ve etkili bir şekilde hataların kaynaklarını belirlemeleri mümkün olabilecektir. Verilerin yoğun veya seyrek olması bir kısıt olmaktan çıkarılmıştır. Ayrıca, kural ilişkilerinin aranmasında veri kümesinin büyüklüğüne olan duyarlılık da azaltılmıştır. Öte yandan, karar vericilerin az sayıda kural grubuna odaklanması mümkün hale getirilmiş ve bu gruplar içinde yapılacak elemenin etkililiği artırılmıştır. İleri bir çalışma olarak, elenen kurallar arasında kalmış olabilecek ilginç kuralların bulunmasına yardımcı olacak bir yöntem üzerinde çalışmalarımız devam etmektedir.

Önerdiğimiz yaklaşım ve elde edilen sonuçlar kısmen Jabarnejad ve Testik (2009) tarafından sunulmuştur. Başka veri tabanları kullanılarak yapılacak karşılaştırmalar sonucunda uluslararası bir makalenin hazırlanması çalışmaları devam etmektedir.

4. Projenin Genel Değerlendirmesi ve Sonuç

Bu projede sanayi kuruluşlarında ürün ve süreçlerin kalitesini iyileştirmeye yönelik VM yaklaşımlarını belirlemek ve daha etkili yaklaşımlar geliştirmek üzere çalışmalar yapılmıştır. Literatür ve saha çalışmaları sonucunda VM yaklaşımlarının imalat sanayinde kalite iyileştirme etkinliklerinin başarısını artırabileceği belirlenmiştir. Ancak ilgili kuruluşların öncelikle veri yönetimi düzenlerini VM gereksinimlerine uygun şekilde iyileştirmeleri gereklidir. Bununla birlikte, yazılım desteği, sağlam ve kullanımı ile öğrenmesi kolay metotların seçimi ve bu kuruluşlarda VM uygulamalarını gerçekleştirecek olan personelin eğitimi gereklidir.

Bu projede belli kalite iyileştirme amaçlarına yönelik VM işlevleri için hangi metotların tercih edilebileceği belirlenmiştir. Buna göre, kalitenin tanımlanması ve veri önişlemede yararlı olabilecek kümeleme çalışmaları için, H/C metodu daha başarılı bulunmuştur. MARS metodunun hem tahmin etme (regresyon) hem de sıralama için öncelikle tercih edilebileceği, bununla birlikte sınıflandırma için destek vektör makinaları ve yapay sinir ağlarının, tahmin etme için ise robust regresyon ve yapay sinir ağlarının da kullanımının önerilebileceği gösterilmiştir. Kalite iyileştirmeye yönelik ürün ve süreç parametrelerinin optimizasyonu için denenen yapay sinir ağları ve yanıt yüzeylerinin çekicilik fonksiyonlarının optimizasyonu metotları arasında kesin bir tercih yapılamamıştır. Bunların her ikisinin de başarımı verinin nasıl toplandığına önemli ölçüde bağlıdır. Bütün bu öneriler, proje kapsamında sanayiden toplanan kısıtlı verilere dayandığı için kesin ve genelgeçer sonuçlardan ziyade dikkate değer sonuçlar olarak değerlendirilmelidir. Bu sonuçlar, VM uygulamalarına başlayacak olan firmalara altyapı ve eğitim gereksinimleri için yol gösterebilir.

Bu çalışmada mevcut metotların uygulamalarında VM madenciliğine özel ticari bir yazılım olan SPSS Clementine®, MARS uygulamaları için Salford Systems®, diğer bir çok uygulama için Matlab®kullanılmış olmakla beraber kuruluşlar çeşitli açık kaynak yazılımlardan yararlanabilir. Bunlar ile ilgili bir envanter http://www.the-data-mine.com/bin/view/Software/AllDataMiningSoftware adresinde verilmektedir.

VM metotlarının uygulanması aşamasında karşılaşılan bazı problemlerin giderilmesi ve mevcut yöntemlerin kullanım kolaylığı ve/veya etkililiğinin artırılması için projede geliştirilen yöntemler kalite iyileştirme çalışmalarına olumlu katkı sağlayabilecektir. Bunlardan kalite verilerinin yeniden örneklenmesi için geliştirilen yöntem, dengesiz kalite verilerinin önişleme aşamasında dengeli hale getirilmesine yardımcı olacaktır. Tahmin etme ve sınıflandırma için tercih edilmesini önerdiğimiz MARS yöntemine alternatif olarak geliştirdiğimiz CMARS yöntemi ile veriye uyum ve modelin karmaşıklığını dengelemede, deneyimli kullanıcıya esneklik sağlamak mümkün olmuştur. Bunların yanısıra, ikili sınıflandırmada kullanımı kolay olan MTS metodunun çok sınıf ve ayrıca parametre optimizasyonu için kullanımı da mümkün hale getirilmiştir. Öte yandan, kalite verilerindeki bulanıklığı modellemek için geliştirdiğimiz uygun alternatif yaklaşımlar (bulanık regresyona dayalı modeller) ve parametrik olmayan bulanık tahmin etme ve sınıflandırma fonksiyonları ile modellemenin kolaylığı ve etkililiği artırılmıştır. Benzer şekilde birliktelik analizinden çıkan çok sayıda kuralın seçimine yönelik etkili bir yaklaşım geliştirilmiştir. Son olarak parametre optimizasyonunda sıklıkla kullanılan çekicilik fonksiyonlarının optimizasyonu için mevcut metotlardan daha başarılı olabilecek alternatifler üzerinde çalışılmıştır. Bu çalışmalar pürüzlü optimizasyon alanıında yeni bir metodolojinin temellerini atmıştır. Bu doğrultuda proje sonrasında da devam edecek araştırmalar ile mevcut yöntemlerden daha başarılı yeni yöntemlerin geliştirilmesi beklenmektedir.

Projede geliştirilen yöntemlerin ileride açık kaynak yazılımlara dönüştürülmesi ile proje sonuçlarının hem ülkemizde hem de dünyada özellikle imalat sanayinde yaygın kullanımı mümkün olabilecektir. Her ne kadar bu projede imalat sanayine odaklanıldı ise de geliştirilen yöntemler imalat sanayi dışında inşaat, madencilik, bankacılık, perakendecilik gibi alanlarda da bir çok kalite iyileştirme ve bilgi keşfetme gereksinimine yanıt verebilecektir.

Akademik çalışmalar özellikle kalite verilerinin az sayıda, dengesiz ve karışık tipte olduğu durumlar için daha etkili analiz ve çözüm yöntemleri geliştirme, elde edilen sonuçların yorumlanması ve değerlendirilmesini kolaylaştırma doğrultusunda ilerleyebilir.

İmalat sektörü dışındaki kuruluşlarda da kalite iyileştirme ve kontrol amaçlı benzer ya da farklı yeni uygulamalar olması kaçınılmaz gözükmektedir. Örneğin, müşteri servis verilerinin daha etkili analizi yoluyla müşteri memnuniyetinin artırılmasında VM'nin yeni uygulamalarından Metin Madenciliği, e-iş ortamında kalitenin iyileştirilmesinde Ağ Madenciliği gibi yaklaşımlar daha yaygın olarak kullanılabilecektir.

Proje bulgularına dayalı bir ulusal makalenin yayımlanması kabul edilmiştir; iki uluslararası makale ise değerlendirme sürecindedir. Toplam beş uluslararası bildiri ve dört ulusal bildiri yayımlanmış, onbir uluslararası ve dokuz ulusal konferans sunuşu gerçekleştirilmiştir. Proje kapsamında sekiz yüksek lisans tezi tamamlanmış, iki yüksek lisans tezinin ise tamamlanması beklenmektedir. Geliştirilen yöntemler ile ilgili

uluslararası makale hazırlıkları ile bir doktora tezinin çalışmaları devam etmektedir. Projede geliştirilen yöntemlere ait kodların uygulayıcıların kullanabileceği açık kaynak yazılımlar halinde sunumu için çalışmalar devam etmektedir.

KISALTMALAR

FCF:	Bulanık Sınıflandırma Fonksiyonları
FCM:	Bulanık c-Ortalamalar
FF:	Bulanık Fonksiyonlar
FLR:	Bulanık Doğrusal Regresyon
IFC:	İyileştirilmiş Bulanık Sınıflandırma
IFCF:	İyileştirilmiş FCF
IFF:	İyileştirilmiş FF
LP:	Doğrusal Program
LR:	Lojistik Regresyon
LS:	En küçük kareler
MARS:	Çok Değişkenli Uyarlanabilir Regresyon Eğrilerini
MLE:	En Çok Olabilirlik Tahmin Edicisi
NIFC:	Parametrik Olmayan IFC
NIFCF:	Parametrik Olmayan IFCF
NIFF:	Parametrik Olmayan IFF
NN:	Yapay Sinir Ağları
DVM:	Destek vektör makineleri
MTS:	Mahalanobis Taghuci Systemi
AUC:	ROC eğrisi altında kalan alan
CMARS:	Sürekli Optimizasyon Tarafından Desteklenen Çok Değişkenli Uyarlanabilir Regresyon Eğrileri
ÇDR:	Çoklu doğrusal regresyon
ANOVA:	Varyans analizi,
KA:	Karar ağaçları
YSA:	Yapay sinir ağları,
YKT:	Yaklaşık küme teorisi,
GA:	Genetik algoritmalar ,
TBA:	Temel bileşenler analizi,
DD:	Dalgacık dönüşümü,
ÖD:	Öznel değerlendirme,
DA:	Diskriminant analizi,
BS:	Bileşik sınıflandırıcılar,
NA:	Nesne ayrıştırma,
MÇB:	Medoidler çevresinde bölme,
KDH:	Kendini düzenleyen haritalar,
OLAP:	Çevrimiçi analitik işleme
GDM:	Genel doğrusal modeller,
NBS:	Naif Bayes sınıflandırıcı,
RBF:	Radyal baz fonksiyon,
VKÖ:	Vektör kuantalamalı öğrenme ,
BKT:	Bulanık küme teorisi,
R:	Regresyon,
DOR:	Doğrusal olmayan regresyon,
ZSA:	Zaman serileri analizi,
YYM:	Yanıt yüzeyi metodu,
ÇKP:	Çok katmanlı perseptronlar,
GYA:	Geri yayılım algoritması,
LM:	Levenberg Marquart ,
BA:	Bayes ağları,
VDD:	Vakaya dayalı düşünce,
ANFIS:	Ağ yapılı bulanık sonuç çıkarım sistemi,
TM:	Taguchi metodu,
USA:	Uyarlanabilir sinir ağları,
AP:	Ardışık programlama,
BT:	Benzetilmiş tavlama,
BM:	Bulanık mantık

Kaynaklar

Agrawal, R., Imielinski, T., Swami, A., Mining association rules between sets of items in large relational databases, *Proceedings of ACM SIGMOD international conference on management of data*, 207-216, (1993).

Agrawal, R., Srikant, R., Fast algorithms for mining association rules, *Proceedings of the 20th International Conference on Very Large Data Bases*, 487-499, 1994.

Akteke-Öztürk, B., Weber, G.W., Kayalıgil, S., Kalite İyileştirmede Veri Kümeleme: Döküm Endüstrisinde Bir Uygulama, Yöneylem Araştırması ve Endüstri Mühendisliği 27. Ulusal Kongresi (YA/EM 2007) Bildiriler Kitabı, p.1207-1212, İzmir, Türkiye, Temmuz 02-04, (2007).

Akteke-Öztürk, B., Weber, G.W., Köksal, G., Çekicilik Fonksiyonlarının Bileşke Fonksiyonlar Olarak Çözümlenmesi, Yöneylem Araştırması ve Endüstri Mühendisliği 29. Ulusal Kongresi (YA/EM 2009) Bildiriler Kitabı (CD), Ankara, Türkiye, Haziran 22-24, (2009).

Anaklı, Zeynep, *A Comparison of DM Algorithms for Prediction and Classification of Quality Data*. Yüksek lisans tezi. Orta Doğu Teknik Üniversitesi, Endüstri Mühendisliği Bölümü, Ankara, Aralık 2009. (yayımlanması beklenmektedir)

Aster, A., Borchers, B., and Thurber, C., *Parameter Estimation and Inverse Problems,* Akademik Baskı, 2004.

Ayhan, D., *Multi-Class Classification Methods Utilizing Mahalanobis Taguchi System and a Re-Sampling Approach For Imbalanced Data Sets*. Yüksek lisans tezi, Orta Doğu Teknik Üniversitesi, Endüstri Mühendisliği Bölümü, Ankara, (2009).

Avcı, Ezgi, *A Comparison of Robust Regression Methods for Outliers.* Yüksek lisans tezi. Orta Doğu Teknik Üniversitesi, Endüstri Mühendisliği Bölümü, Ankara, Eylül 2009.

Baek, D.H., Jeong, I.J. ve Han, C.H., Application of data mining for improving yield in water fabrication system. *In*: Gervasi, O. et al. (eds.), *Computational Science and Its Applications – ICCSA 2005*, 9-12 Mayıs 2005 Singapore. Berlin: Springer-Verlag, 3483, 222-231, (2005).

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

Bagirov,A.M., Minimization methods for one class of nonsmooth functions and calculation of semi-equilibrium prices. In Eberhard,A. *et al*. (eds.) *Progress in Optimization: Contribution from Australasia*. Kluwer Academic Publishers, Dordrecht, pp. 147–175, (1999).

Bagirov, A.M., Rubinov, A.M., Soukhoroukova, N.V., Yearwood, J., Unsupervised and supervised data classification via nonsmooth and global optimization, *TOP*, 11, 1, 1-93, (2003).

Bakır, B., Defect Cause Modeling with Decision Tree and Regression Analysis: A Case Study in Casting Industry. Yüksek lisans tezi. ODTÜ, Enformatik Enstitüsü, Ankara, 2007.

Bakır, B., Batmaz, I., Güntürkün, F. A., Ipekci, İ. A., Köksal, G. ve Özdemirel, N. E., Defect cause modeling with decision tree and regression analysis. *In*: *Proceedings of XVII. International Conference on Computer and Information Science and Engineering*, 8-10 December 2006 Cairo. Cairo: World Enformatica Society, 16, 266-269, (2006)

BARON v.8.1.5 (2009)

Batmaz, I., Data mining applications on manufacturing data: a casting quality improvement case.

Ayhan, H.O. ve Batmaz, I. (editörler), *Recent Advances in Statistics*. Ankara: TUIK, 197-206, (2007)

Berrado, A., Runger, G.C., Using metarules to organize and group discovered association rules. *Data Mining and Knowledge Discovery,* 14, 409-431, (2007).

Bertino, E., Catania, B. ve Caglio, E., Applying data mining techniques to wafer manufacturing. *Principles of Data Mining and Knowledge Discovery*, 1704, 41-50, (1999).

Birkes, D., and Dodge, Y., *Alternative Methods of Regression.* New York, John Wiley & Sons. Inc., (1993).

Brans, J.P., and Vincke, Ph., A Preference Ranking Organisation Method: (The PROMETHEE Method for Multiple Criteria Decision-Making), *Management Science*, 31, 6, 647-656, (1985)

Breiman, L., J. H. Friedman, R. A. Olshen, and C. J. Stone. *Classification and Regression Trees*. New York: Chapman & Hall/CRC, (1984).

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

Brumen, B., Golob, I., Jaakkola, H., Welzer, T. and Rozman, I.: Early Assessment of Classification Performance, *Australian CS Week Frontiers*, 91–96, (2004).

Burachik, R.S., Gasimov, R.N., Ismayilova, N.A., and Kaya C. Y., On a Modified Subgradient Algorithm for Dual Problems via Sharp Augmented Lagrangian, *Journal of Global Optimization*, 34, 1, 55 – 78, (2006).

CAAT, Cluster Accuracy Analysis Tool Documentation, 2005, http://caat.bioinfo.cipf.es/CAAT_docs.pdf

Calinski, T., Harabasz, J., A dendrite method for cluster analysis, *Communications in statistics*, 3, 1, 1--27, (1974).

Castillo, E. D., Montgomery, D. C., McCarville, D. R., Modified Desirability Functions for Multiple Response Optimization, *Journal of Quality Technology*, 28, 3 337–345, (1996).

Ch'ng, C. K., Quah, S. H., and Low, H. C., A New Approach for Multiple-Response Optimization, *Quality Engineering*, 17, 621–626, (2005).

Chawla, N.V., Bowyer, K.W., Kegelmeyer W.P. and Hall, L.O., SMOTE: Synthetic Minority Oversampling Technique, *Journal of Artificial Intelligence Research*.16, 341-378 (2002).

Chen, W. C., Lee, A. H. I., Deng, W. J. ve Liu, K. Y., The implementation of neural network for semiconductor PECVD process, *Expert Systems with Applications*, 32, 4, 1148–1153, (2007).

Chiang, T.L., Su, C.T., Li, T.S., ve Huang, R.C.C.,, Improvement of process capability through neural networks and robust design: A case study. *Quality Engineering*, 14, 2, 313-318, (2002).

Chien, C., Li H., ve Jeang, A., Data mining for improving the solder bumping process in the semiconductor packaging industry. *Intelligent Systems in Accounting, Finance and Management,* 14, 1-2, 43-57, (2006).

Chien, C.F.,Wang, W.C., Cheng, J.C., Data mining for yield enhancement in semiconductor manufacturing and an empirical study. *Expert Systems with Applications*, 33,1, 192-198, (2007).

Chin, K. K., 1998, *Support Vector Machines Applied to Speech Pattern Classification*. Yüksek

lisans tezi, Cambridge Üniversitesi, Cambridge, U.K.

Clarke F., *Optimization and Nonsmooth Analysis,* SIAM's Classics in Applied Mathematics Series, (1983)

Clementine 11.0 Algorithms Guide., USA: Integral Solutions Limited (2007).

http://www.spss.com/clementine/

Cook, D.F., Ragsdale, C.T., Major, R.L., Combining a Neural Network with a Genetic Algorithm for Process Parameter Optimization, *Engineering Applications of Artificial Intelligence*, 13, 391-396, (2000).

Cool, T., Bhadeshia, H.K.D.H., MacKay, D.J.C., The Yield and Ultimate Tensile Strength of Steel Welds, *Materials Science and Engineering*, A223, 186-200, (1997).

Cser, L., Gulyas, J., Szucs, L., Horvath, A., Arvai, L. ve Baross, B., Different kinds of neural networks in control and monitoring of hot rolling mill. Monostori, L., Vancza, J. ve Ali, M. (editörler), *Proceedings of the 14th International conference on industrial and engineering applications of artificial intelligence and expert systems: engineering of intelligent systems, IEA/AIE 2001, LectureNotes In Computer Science*, 4-7 June 2001 Budapest Hungary. London: Springer-Verlag, 2070, 791 – 796, (2001).

Cus, F., Balic, J., Optimization of Cutting Process by GA Approach, *Robotics and Computer Integrated Manufacturing*, 19, 113–121, (2003).

Çabuk, V., 2008, Modeling and Analysis of Customer Requirements From a Drivers Seat, Yuksek lisans tezi, Orta Doğu Teknik Üniversitesi, Endustru Muh. Bilimi, Ankara (2008).

Çelikyılmaz, A., *Modeling Uncertainty with Evolutionary Improved Fuzzy Functions*, Doktora Tezi, Toronto Üniversitesi Makine ve Endüstri Mühendisliği Bölümü, (2008).

De Abajo, N., Diez, A. B., Lobato, V. ve Cuesta, S. R., ANN quality diagnostic models for packaging manufacturing: an industrial data mining case study. *In*: Kohavi, R. et al. (eds.), *KDD-2004: proceedings of the tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining,* 22-25 August 2004 Seattle Washington. New York: ACM Press, 799-804, (2004).

Deng, B. ve Liu, X. (2002). Data mining in quality improvement. SUGI27: *Proceedings of the Twenty-Seventh Annual SAS® Users Group International Conference*, 14-17 April 2002 Orlando, Florida [online]http://www2.sas.com/proceedings/sugi27/Proceed27.pdf [23 Nisan 2008 tarihinde erişilmiştir].

Derringer, G., and Suich, R., Simultaneous Optimization of Several Response Variables, *Journal of Quality Technology*, 12, 4, 214–219, (1980).

Dhar, V., and Stein, R., *Seven Methods for transforming Corporate data Into Business Inteligence*, Prentice Hall, 1-29, 1997.

Dunham, M. H. (2003). *Data mining introductory and advanced topics*, New Jersey, Prentice Hall/Pearson Education.

Dunn, J. C., A Fuzzy Relative of the ISODATA Process and Its Use in Detecting Compact Well-Separated Clusters, *Journal of Cybernetics* 3, 32-57, (1973).

Dutta, J., Generalized Derivatives and Nonsmooth Optimization, a Finite Dimensional Tour, *TOP*,

13, 2,185-314, (2005).

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

Estabrooks, A., Jo T. ve Japkowcz N., A multiple resampling method for learning from imbalanced data sets, *Computational Intelligence*, 20,1, (2004).

Fielding, A. H., Bell J. F., A review of methods for the assessment of prediction errors in conservation presence/absence models, *Environmental Conservation,* 24, 1, 38–49 (1997).

Friedman, J., *Another Approach to Polychotomous Classification. Technical Report*, Stanford Üniversitesi, İstatistik Bölümü, Stanford, CA, (1996).

Friedman, J.H., Multivariate adaptive regression splines, *The Annals of Statistics* 9, 1, 1-141, (1991).

GAMS v.20.0.2 (2009)

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

Georgilakis, P., Hatziargyriou, N., On the Application of Artificial Intelligence Techniques to the Quality Improvement of Industrial Processes, Vlahavas, I.P. ve Spyropoulos, C.D. (eds.), *Proceedings of the Second Hellenic Conference on AI: Methods and Applications of Artificial Intelligence, Lecture Notes In Computer Science*, 11-12, Nisan, Thessaloniki, Greece**.** London: Springer-Verlag, 2308**,** 473 – 484, (2002).

Guessasma, S., Salhi, Z., Montavon, G., Gougeon, P., Coddet, C., Artificial Intelligence Implementation in the APS Process Diagnostic, *Materials Science and Engineering B*, 110, 285–295, (2004).

Güntürkün, F., A Comprehensive Review of Data Mining Applications in Quality Improvement and A Case Study. Yüksek Lisans Tezi. ODTÜ, İstatistik Bölümü, Ankara, 2007.

Halkidi, M., Batistakis, Vazirgiannis, M., On Clustering Validation Techniques, *Journal of Intelligent Information Systems,* 17, 2/3, 107–145, 2001.

Hamedi, M., Shariatpanahi, M. ve Mansourzadeh, A., Optimizing spot welding parameters in a sheet metal assembly by neural networks and genetic algorithm. *Proceedings of the Institution of Mechanical Engineers Part B-Journal of Engineering Manufacture*, 221, 7, 1175-1184, (2007).

Han, J. and Kamber, M, *Data Mining: Concepts and Techniques*, Morgan Kaufmann Publishers,2001

Harding, J.A., Shahbaz, M., Srinivas, S. ve Kusiak, A., Data mining in manufacturing: A review. *Journal of Manufacturing Science and Engineering-Transactions of ASME*, 128, 4, 969-976, (2006).

Harrington, E. C., JR., The Desirability Function, *Industrial Quality Control,* 21, 494-498, (1965).

Haykin, S. *Neural Networks: A Comprehensive Foundation*, New York: Macmillan, (1994).

Holena, M., Baerns, M., Feedforward neural networks in catalysis - A Tool for the Approximation of the Dependency of Yield on Catalyst Composition, and for Knowledge Extraction, *Catalysis Today*, 81, 485–494, (2003).

Hsieh, K., Tong, L., Optimization of Multiple Quality Responses Involving Qualitative and Quantitative Characteristics in IC Manufacturing Using Neural Networks, *Computers in Industry*, 46, 1-12, (2001).

Hsu, C. W. ve Lin, C. J., A comparison of methods for multi-class support vector machines. *IEEE Transactions on Neural Networks*, 13, 2, 415-425, (2002).

Hsu, S.C. ve Chien, C.F., Hybrid data mining approach for pattern extraction from wafer bin map to improve yield in semiconductor manufacturing. *International Journal of Production Economics*, 107, 1, 88-103, (2007).

Hu, C., Su, S., Hierarchical Clustering Methods for Semiconductor Manufacturing Data, *Proceedings of the IEEE International Conference on Networking*, 21-23 Mart 2004 Taipei, Taiwan. IEEE, 2, 1063 – 1068, (2004).

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

Huang, H., Wu, D., 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, (2005).

Huber, P. J., *Robust Statistics*., NY: Wiley and Sons. (2003).

Hung, Y. H., Optimal process parameters design for a wire bonding of ultra-thin CSP package based on hybrid methods of artificial intelligence. *Microelectronics International*, 24, 3, 3-10, (2007).

Hyndman R. J., Koehler A. B., Another look at measures of forecast accuracy, *International Journal of Forecasting*, 22, 679– 688, (2006).

Jabarnejad, Masood, A Method for Grouping and Pruning of Association Rules. Yüksek lisans tezi. Orta Doğu Teknik Üniversitesi, Endüstri Mühendisliği Bölümü, Ankara, Aralık 2009. (yayımlanması beklenmektedir)

Jain A.K., and Dubes R.C., *Algorithms for Clustering Data*, Englewood Cliffs, NJ, Prentice Hall, 1988.

Jain A.K., Murty M.N., and Flynn P.J., Data clustering: A review, *ACM Computing Surveys* 31, 264-323, (1999).

Jain A.K., Topchy A., Law M.H.C, and Buhmann J.M., Landscape of clustering algorithms, to appear in: *Proc. IAPR International Conference on Pattern Recognition*, Cambridge, UK, 2004.

Jiao, Y., Lei, S., Pei, Z.J., Lee, E.S., Fuzzy Adaptive Networks in Machining Process Modeling: Surface Roughness Prediction for Turning Operations, *International Journal of Machine Tools & Manufacture*, 44, 1643–1651, (2004).

Jongen, H. Th., Pallaschke, D., On linearization and continuous selections of functions, *Optimization* 19, 3 343–353, (1988).

Jongen,H.Th., Jonker,P., Twilt, F., *Nonlinear Optimization in Finite Dimensions*. Nonconvex Optimization and its Applications, Vol. 47, Kluwer (2000).

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

Karim, M.A., Halgamuge, S., Smith, A. J.R. ve Hsu, A.L., Manufacturing yield improvement by clustering. *In*: King, I. et al. (eds.), *Neural Information Processing: 13th International Conference Proceedings Part,* (2006).

Kass, G. An exploratory technique for investigating large quantities of categorical data. *Applied Statistics*, 29, 2, 119–127, (1980).

Kaufman, L., Rousseeuw, P., *Finding Groups in Data: An Introduction to Cluster Analysis*. John Wiley & Sons, New York, (1990).

Kılıç, T., *Fuzzy Regression Modeling of Defect Rate in a Metal Casting Process*, (Yüksek Lisans Tezi), Çankaya Üniversitesi Endüstri Mühendisliği Bölümü, Ankara, (2009).

Kim, E., Oh, C., Lee, S., Lee, B., Yun, I., Modeling and Optimization of Process Parameters for GaAs/AlGaAs Multiple Quantum Well Avalanche Photodiodes Using Genetic Algorithms, *Microelectronics Journal*, 32, 563-567, (2001).

Kim, I., Son, J., Yarlagadda, P. K.D.V. A Study on the Quality Improvement of Robotic GMA Welding Process, *Robotics and Computer Integrated Manufacturing*, 19, 567–572, (2003).

Kim, S., Lee, C. M., Nonlinear Prediction of Manufacturing Systems Through Explicit and Implicit Data Mining, *Computers and Industrial Engineering*, 33, 461-464, (1997).

Kohonen, T., *Self-Organizing Maps*, Springer Series in Information Sciences, Vol. 30, Springer, Berlin, Heidelberg, New York, (1995).

Köksal, G., Anaklı, Z., Batmaz, İ., Yerlikaya Özkurt, F., Testik, M.C., Kartal, E., Kayalıgil, S., Bakır, B., Karasakal, E., Comparison of Data Mining Algorithms for Classification and Prediction in Quality Improvement, *23rd European Conference on Operational Research*, Bonn, 39, Temmuz (2009a).

Köksal, G., Ayhan, D. ve Yenidünya, B., Kasım 2009b, Kalite Sınıflandırma ve Eniyileme için Mahalanobis Taguchi Sistemi Yaklaşımları, *18. Kalite Kongresi Bildirileri*'nde yayımlanacaktır.

Köksal, G., Batmaz, İ. ve Kartal, E., Developing a Classification Model For Customer Satisfaction with a Driver's Seat: A Comparative Case Study, Proceedings of 6th International Symposium on Intelligent and Manufacturing Systems, Sakarya, 520-530 (2008a).

Köksal, G., Batmaz, İ. ve Testik, M.C., Data mining processes and a review of their applications for product and process quality improvement in manufacturing industry, Teknik Rapor No: 08-03, Orta Doğu Teknik Üniversitesi, Endüstri Mühendisliği Bölümü, Ankara, (2008b).

Köksal, G., Batmaz, İ. ve Testik, M.C., A review of data mining applications for description, prediction, classification and optimisation of quality in manufacturing industry, *International Journal of Management Reviews* dergisinde, birinci revizyonu değerlendirilmektedir, Haziran 2009c.

Köksal, G., Batmaz, İ, Testik, M.C. ve Güntürkün, F., İmalat Sektöründe Kalite İyileştirmede Veri Madenciliği Tekniklerinin Kullanımı, *Verimlilik Dergisi*'nde yayımlanacaktır. (2008c).

Krimpenis, A., Benardos, P.G., Vosniakos, G.-C., Koukouvitaki, A., 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, (2006).

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

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

Kusiak, A. Data mining: Manufacturing and service applications. *International Journal of Production Research*, 44, 18/19, 4175-4191, (2006).

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

Kutner, M.H., C.J. Nachtsheim, J. Neter ve L. William, *Applied Linear Statistical Models*. McGraw-Hill/Irwin, (2004).

Lasdon, L.S., Waren, A. D., Jain, A., and Ratner, M., Design and testing of a generalized reduced gradient code for nonlinear programming, *ACM Transactions on Mathematical Software*, 4(1), 34-50, (1978)

Lee, S.H. ve Dornfeld, D.A., Prediction of burr formation during face milling using an artificial neural network with optimized cutting conditions. *Proceedings of the Institution of Mechanical Engineers Part B-Journal of Engineering Manufacture*, 221, 12, 1705-1714, (2007).

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

Li, T.S., Huang, C.L. ve Wu, Z.Y., Data mining using genetic programming for construction of a semiconductor manufacturing yield rate prediction system. *Journal of Intelligent Manufacturing*, 17, 355-361, (2006).

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

Lin, W. S., Wang, K. S., Modeling and Optimization of Turning Processes for Slender Parts, *International Journal of Production Research*, 38, 3, 587-606, (2000).

Liu, X., Tang, H., Fan, Z., Deng, B.,. Quality improvement modeling and practice in baosteel based on KIV-KOV Analysis, Lecture Notes in Computer Science, 3129, 720-725, 2004.

Awad, M., ve Khan, L., *Applications and Limitations of Support Vector Machines*, ed. John Wang, Encyclopedia of Data Warehousing and Mining by Information Science Publishing, (2004).

MacQueen, J., Some methods for classification and analysis of multivariate observations. volume 1 of *Proceedings of the Fifth Berkeley Symposium on Mathematical statistics and probability*, pages 281-297, Berkeley, University of California Press, (1967).

Mahalanobis, P. C. (1936). On the Generalized Distance in Statistics*, Proceedings of the National Institute of Science of India*, 49-55.

Maimon, O. ve Rokach, L.S., Data mining by attribute decomposition with semiconductor manufacturing case study. *In*: Braha, D. (ed.), *Data Mining for Design and Manufacturing*. Dordrecht: Kluwer Academic Publishers, 311-336, (2001).

Manel, S., Dias, J-M, Ormerod, S.J. Comparing Discriminant Analysis, Neural Networks And Logistic Regression For Predicting Species Distributions: A Case Study With A Himalayan River Bird, *ELSEVIER Ecological Modelling*, 120, 337–347, (1999).

MATLAB version 7.5 (R2007b)

Meng, T.K. ve Butler, C. Solving multiple response optimisation problems using adaptive neural networks. *International Journal of Advanced Manufacturing Technology*, 13, 9, 666-675, (1997).

Montgomery, D., E.A.Peck ve G.G. Vining, *Introduction to Linear Regression Analysis*. NY: Wiley and Sons., (2006).

Muñoz J. & Felicísimo Á., Comparison of statistical methods commonly used in predictive modelling, *Journal of Vegetation Science*, 15, 285-292, (2004).

Myers, R.H., Montgomery, D.C., *Response Surface Methodology: Process and Product Optimization Using Designed Experiments*,New York, Wiley, (2002).

Nemirovski, A., Lectures on modern convex optimization, Israel Institute Technology, http://iew3.technion.ac.il/Labs/Opt/opt/LN/Final.pdf, (2002).

Olabi, A.G., Casalino, G., Benyounis, K.Y., Hashmi, M.S.J., An ANN and Taguchi Algorithms Integrated Approach to the Optimization of CO2 Laser Welding, *Advances in Engineering Software*, 37, 643–648, (2006).

Ordieres Meré J.B., González Marcos A., González J.A. ve Lobato Rubio V. Estimation of mechanical properties of steel strip in hot dip galvanising lines. *Ironmaking and Steelmaking*, 31, 1, 43-50, 8, (2004).

Ou, G., Murphey, Y.L. ve Feldkamp, L., Multi-class pattern classification using neural networks. In ICPR '04: *Proceedings of the Pattern Recognition*. 4, 584-568, (2004).

Özçelik, B., Erzurumlu, T. 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, (2006).

Özer, G., *Fuzzy Classification Models Based On Tanaka's Fuzzy Linear Regression Approach And Nonparametric Improved Fuzzy Classifier Functions*, (Yüksek Lisans Tezi), Orta Doğu Teknik Üniversitesi Endüstri Mühendisliği Bölümü, Ankara, (2009).

Özer, G., Kılıç, T., Kartal, E., Batmaz, İ., Türker Bayrak, Ö., Köksal, G., Türkşen B., A Nonparametric Improved Fuzzy Classifier Function Approach for Classification Based on Customer Satisfaction Survey Data, *23rd European Conference on Operational Research*, Bonn, 63, (2009b).

Özer, G., Köksal, G., Batmaz, İ., Türker Bayrak, Ö., Kılıç, T., Kartal, E., Classification Models Based on Tanaka's Fuzzy Linear Regression Approach: The Case of Customer Satisfaction Modeling, *1st International Fuzzy Systems Symposium Proceeding*, Ankara, (2009a) (yayımlanacaktır).

Öztürk, B., Datamining for Quality Improvement: New Advances by Nonsmooth Optimization and Continuous Optimization, Doktora tezi, Orta Doğu Teknik Üniversitesi, Uygulamalı Matematik Enstitüsü, Ankara, Aralık 2010. (yayımlanması beklenmektedir)

Taylan, P., Weber, G.W., and Yerlikaya, F., Continuous optimisation applied in MARS for modern applications in finance, science and technology, *ISI Proceedings of 20th Mini-EURO Conference Continuous Optimisation and Knowledge-Based Technologies*, Neringa, Lithuania, 317-322, (2008).

Paolo, G. Applied *Data Mining: Statistical Methods for Business and Industry*, Wiley and Sons, NewYork, (2003).

Perzyk, M., Biernacki, R., Kochanski, A., Modeling of Manufacturing Processes by Learning Systems: The Naive Bayesian Classifier Versus Artificial Neural Networks, *Journal of Materials Processing Technology*, 164–165, 1430–1435, (2005).

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

Pham, D.T. ve Afify, A.A., Machine-learning techniques and their applications in manufacturing. *Proceedings of the Institution of Mechanical Engineers Part B-Journal of Engineering Manufacture*, 219 5, 395-412, (2005).

Ribeiro, B., 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, (2005).

Rockefellar, R.T., *Convex Analysis,* Princeton University Press, New Jersey (1972).

Rokach, L and Maimon, O, Data mining for improving the quality of manufacturing: a feature set decomposition approach, *J Intell Manuf* , 17, 285-299, (2006).

Rokach, L. ve Maimon, O., Data mining for improving the quality of manufacturing: a feature set decomposition approach. *Journal of Intelligent Manufacturing*, 17, 3, 285-299, (2006).

Quinlan, J. R., Induction of Decision Trees. *Machine Learning*, 1, 81, (1986).

Russell, S. J., Norvig, P., *Artificial intelligence : a modern approach*, N.J., Prentice Hall, (2003)

Saaty T., Fundamentals of the analytic network process, *ISAHP*, Kobe, Japan, August 12-14, (1999).

Saaty, T., *Decision Making with Dependence and Feedback: The Analytic Network Process*, RWS Publications, Pittsburgh, 1996.

Sadeghi, B.H.M., A BP-Neural Network Predictor Model for Plastic Injection Molding Process, *Journal of Materials Processing Technology*, 103, 411-416, (2000).

Salford Systems, Academic MARS 2.0, http://www.salfordsystems.com/mars.php

Sarimveis, H., Doganis, P., Alexandridis, A., A Classification Technique Based on Radial Basis Function Neural Networks, *Advances in Engineering Software*, 37, 218–221, (2006).

Sharma, S.C.. *Applied Multivariate Techniques.* John Wiley and Sons (1996).

Shen, C., Wang, L. ve Li, Q. Optimization of injection molding process parameters using combination of artificial neural network and genetic algorithm method, *Journal of Materials Processing Technology*, 183 (2-3), 412-418, (2007).

Shi, X., Schillings P., Boyd, D. Applying Artificial Neural Networks and Virtual Experimental Design to Quality Improvement of Two Industrial Processes, *International Journal of Production Research*, 42, 1,101–118, (2004).

Shor, N. Z., *Minimization Methods for Non-differentiable Functions*. Springer-Verlag, ISBN 0-387-12763-1 (1985).

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

SPSS 16.0 GPL Reference Guide, Chicago, IL: SPSS Inc, 2007. http://support.spss.com/ProductsExt/SPSS/Documentation/SPSSforWindows (accessed 26 Aug. 2009).

Steuer, R.E., *Multiple Criteria Optimization: Theory,Computation and Application*, NY: Wiley, (1986).

Suneel, T.S., Pandle, S.S., Date, P.P. A Technical Note on Integrated Product Quality Model Using Artificial Neural Networks, *Journal of Materials Processing Technology*, 121, 77-86, (2002).

T. Hastie, R. Tibshirani and J. Friedman, *The Elements of Statistical Learning, Data Mining, Inference and Prediction*, New York, NY: Springer, (2001).

Taguchi, G. ve Jugulum, R., New trends in multivariate diagnosis*. Sankhya: The Indian Journal of Statistics*. 62, 233-248, (2000).

Taguchi, G., Chowdhury, S. ve Yuin, W. *The Mahalanobis-Taguchi System.* New York: McGraw-Hill, (2001).

Tam, C.M., Tong, T. K. L., Lau, T. C. T., Chan, K.K. Diagnosis of Prestressed Concrete Pile Defects Using Probabilistic Neural Networks, *Engineering Structures*, 26, 1155–1162, (2004).

Tan, S.C., Lim, C.P., Rao, M.V.C. 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, (2007).

Tanaka H., Hayashi I., Watada J., Possibilistic linear regression analysis for fuzzy data, *European Journal of Operational Research*, 40, 389-396, (1989).

Tanaka H., Uejima S., Asai K., Fuzzy linear model, fuzzy linear regression model, *IEEE Trans. System Man Cybernet.*, 12, 903 – 907, (1982).

Tay, K.M., Butler, C., 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, (1997).

Taylan, P., Weber, G.-W., and Beck, A., New approaches to regression by generalized additive models and continuous optimization for modern applications infinance, science and technology. *Optimization* 56, 5–6, 1–24, (2007).

Tohumcu, Z. ve Karasakal, E., Project Performance Evaluation with Multiple and Interdependent Criteria, 22nd European Conference on Operational Research, (2007)

Tsai, Y., Chen, J. C., Lou, S., An In-process Surface Recognition System Based on Neural Networks in End Milling Cutting Operations, *International Journal of Machine Tools & Manufacture*, 39, 583–605, (1999).

Tseng, H.Y., Welding parameters optimization for economics design using neural approximation and genetic algorithm. *International Journal of Advanced Manufacturing Technology*, 27, 9/10,897-901, (2006).

Vasudevan, M., Murugananth, M., Bhaduri, A.K., 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, (2002).

Vasudevan, M., Rao, B.P.C., Venkatraman, B., Jayakumar, T., Raj, B., Artificial Neural Network Modeling for Evaluating Austenitic Stainless Steel and Zircaloy-2 Welds, *Journal of Materials*

Processing Technology, 169, 396–400, (2005).

Vlcek J., A Bundle-Type Algorithms for. Nonsmooth Optimization., Technical report , Institue of Computer Sciences, Academy of Sciences of The Chech Republic, No: 608, (1997).

Wang, H., Tsaur, R., Insight of a fuzzy regression model, *Fuzzy Sets and Systems*, 112, 355–369, (2000).

Wang, K., Applying data mining to manufacturing: the nature and implications, *Journal of Intelligent Manufacturing*, 18, 487-495, (2007).

Wang, K., Tong, S., Eynard, B., Roucoules, L. and Matta, N., Review on application of data mining in product design and manufacturing. Lei, J., Yu, J. ve Zhou, S. (editörler) *Proceedings of the Fourth International Conference on Fuzzy Systems and Knowledge Discovery (FSKD 2007), 24-27 August 2007Haikou Hainan China*. IEEE, 613-618, (2007a).

Wang, L., Fang, J. C., Zhao, Z. Y. ve Zeng, H. P., Application of backward propagation network for forecasting hardness and porosity of coatings by plasma spraying. *Surface and Coating Technology*, 201, 9-11, 5085-5089, (2007b).

Ward, J. H., "Hierarchical Grouping to Optimize an Objective Function". *Journal of the American Statistical Association* 58, 301, 236–244, (1963).

Weiss G.M., Provost F., Learning When Training Data are Costly: The Effect of Class Distribution on Tree Induction. *Journal of Artificial Intelligence Research*, 19, 315-354, (2003).

Werbos, P. J., Beyond Regression: New Tools for Prediction and Analysis in
the Behavioral Sciences. Unpublished doctoral dissertation, Harvard University, (1974).

West, P.M., Brockett, P. L. and Golden L. L., A Comparative Analysis of Neural Networks and Statistical Methods for Predicting Consumer Choice, *Marketing Science*, 16, 4, 370-391, 1997.

Windham, M.P., Cluster Validity for Fuzzy Clustering Algorithms, *J. Fuzzy Sets and Systems,* 5, 177-185, 1981.

Ye N., *The Handbook of Data Mining*. Lawrence Erlbaum; 1 edition, 426-440, 2003.

Yenidünya, B., *Robust design with binary response using Mahalanobis Taguchi System*. Yüksek lisans tezi. Endüstri Mühendisliği Bölümü, Orta Doğu Teknik Üniversitesi, Ankara, (2009).

Yerlikaya, F., A *New Contribution to Nonlinear Robust Regression and Classification With MARS and Its Applications to Data Mining For Quality Control in Manufacturing.* Yüksek lisans tezi. Orta Doğu Teknik Üniversitesi, Uygulamalı Matematik Enstitüsü, Ankara, 2008. (**ODTÜ, Uygulamalı Matematik Enstitüsü, Bilimsel Hesaplama Yüksek Lisans Programı En İyi Tez Ödülü, 2009**)

Yin, X., Yu, W., The Virtual Manufacturing Model of the Worsted Yarn Based on Artificial Neural Networks and Grey Theory, *Applied Mathematics and Computation,* 185, 1, 322-332, (2006).

Zhang, J.H., Xie, A.G. ve Shen, F.M., Multi-objective optimization and analysis model of sintering process based on BP neural network. *Journal of Iron and Steel Research, International*, 14, 2, 01-05, (2007).

Zhou, Q., Xiong, Z., Zhang, J. ve Xu, Y., Hierarchical neural network based product quality prediction of industrial ethylene pyrolysis process. Wang, J. v.d. (editörler) *Proceedings of the*

Third International Symposium on Neural Networks, Advances in Neural Networks (ISNN 2006), *Lecture Notes in Computer Science*, 28 Mayıs – 1 Haziran, 2006 Chengdu China. Berlin: Springer-Verlag, 3973, 1132-113, (2006).

EK

Tablo EK-1: Tahmin etme modellerinin performans sonuçları

Ölçüt	CV	METOT					
		KA	YSA	MARS	ÇDR	HUBERM (logit)	Bulanık Fonk.
MAE	R1T1	0.000	0.018	0.031	0.035	0.026	0.030
	R1T2	0.017	0.049	0.045	0.031	0.063	0.014
	R1T3	0.016	0.039	0.032	0.029	0.041	0.029
	R2T1	0.021	0.024	0.024	0.030	0.026	0.025
	R2T2	0.032	0.036	0.076	0.038	0.032	0.031
	R2T3	0.022	0.026	0.021	0.026	0.059	0.022
	R3T1	0.028	0.019	0.034	0.031	0.041	0.018
	R3T2	0.028	0.019	0.089	0.026	0.021	0.017
	R3T3	0.033	0.039	0.126	0.038	0.068	0.034
MSE	R1T1	0.000	0.001	0.003	0.003	0.004	0.003
	R1T2	0.001	0.008	0.006	0.001	0.037	0.001
	R1T3	0.001	0.004	0.002	0.002	0.011	0.002
	R2T1	0.002	0.002	0.002	0.002	0.003	0.001
	R2T2	0.006	0.006	0.020	0.004	0.007	0.004
	R2T3	0.002	0.002	0.002	0.001	0.033	0.001
	R3T1	0.003	0.002	0.003	0.001	0.019	0.001
	R3T2	0.003	0.002	0.130	0.001	0.002	0.001
	R3T3	0.006	0.005	0.086	0.004	0.026	0.003
RMSE	R1T1	0.000	0.034	0.053	0.056	0.064	0.053
	R1T2	0.037	0.091	0.074	0.037	0.193	0.030
	R1T3	0.037	0.059	0.049	0.041	0.105	0.048
	R2T1	0.047	0.040	0.041	0.039	0.055	0.038
	R2T2	0.080	0.079	0.143	0.061	0.082	0.062
	R2T3	0.045	0.047	0.040	0.030	0.181	0.036
	R3T1	0.054	0.041	0.055	0.038	0.137	0.030
	R3T2	0.058	0.039	0.360	0.031	0.044	0.030
	R3T3	0.076	0.074	0.294	0.062	0.163	0.057
r	R1T1	1.000	0.836	0.495	0.449	0.142	0.563
	R1T2	0.408	0.536	0.128	0.490	-0.063	0.650
	R1T3	0.814	0.550	0.739	0.780	-0.020	0.647
	R2T1	0.458	0.478	0.439	0.530	0.140	0.633
	R2T2	0.199	0.197	0.145	0.641	0.185	0.721
	R2T3	0.681	0.152	0.485	0.442	-0.089	0.494
	R3T1	0.087	0.539	0.562	0.574	-0.023	0.768
	R3T2	0.329	0.307	0.104	0.574	0.162	0.642
	R3T3	0.324	0.311	0.258	0.610	-0.046	0.732
R2	R1T1	1.000	0.700	0.245	0.201	0.020	0.317

	R1T2	0.167	0.288	0.016	0.240	0.004	0.423
	R1T3	0.663	0.302	0.546	0.608	0.000	0.418
	R2T1	0.209	0.228	0.192	0.281	0.020	0.401
	R2T2	0.039	0.039	0.021	0.410	0.034	0.520
	R2T3	0.463	0.023	0.236	0.195	0.008	0.244
	R3T1	0.008	0.291	0.315	0.329	0.001	0.589
	R3T2	0.108	0.094	0.011	0.330	0.026	0.412
	R3T3	0.105	0.097	0.067	0.372	0.002	0.536
Adj-R2	R1T1	1.000	0.492	0.158	4.993	-0.252	0.069
	R1T2	0.069	-8.681	-4.251	4.799	-26.741	0.213
	R1T3	0.655	27.334	0.215	2.625	-2.940	0.196
	R2T1	-0.091	0.140	-0.068	4.595	-1.164	0.184
	R2T2	-0.055	-0.113	-4.375	3.949	-0.465	0.346
	R2T3	-1.121	-8.810	-0.787	4.334	-42.717	0.188
	R3T1	-0.478	-0.071	-0.912	4.353	-10.885	0.440
	R3T2	-1.674	-0.656	-145.243	4.352	-0.976	0.370
	R3T3	0.026	-0.071	-25.587	3.601	-4.422	0.359
PWI1	R1T1	0.000	0.968	0.936	0.968	0.903	0.968
	R1T2	0.935	0.871	0.903	0.903	0.903	0.871
	R1T3	0.900	0.900	0.933	0.933	0.933	0.867
	R2T1	0.935	0.935	0.871	0.935	0.871	0.903
	R2T2	0.903	0.903	0.903	0.935	0.871	0.935
	R2T3	0.933	0.933	0.933	0.967	0.933	0.933
	R3T1	0.935	0.935	0.936	0.968	0.936	0.871
	R3T2	0.903	0.935	0.968	0.968	0.871	0.903
	R3T3	0.900	0.933	0.900	0.933	0.933	0.933
PWI2	R1T1	0.000	0.968	0.968	0.968	0.903	0.968
	R1T2	0.935	0.968	1.000	1.000	0.903	0.968
	R1T3	0.967	1.000	1.000	0.967	0.967	1.000
	R2T1	0.968	1.000	0.968	1.000	0.968	0.968
	R2T2	0.935	0.935	0.968	0.968	0.903	0.935
	R2T3	0.967	0.967	0.967	1.000	0.933	0.967
	R3T1	0.968	0.968	0.968	1.000	0.936	1.000
	R3T2	0.968	0.968	0.968	1.000	0.936	1.000
	R3T3	0.933	0.933	0.967	0.967	0.967	0.933
Stability_MSE	R1T1	1.000	-0.160	-0.490	-0.873	-0.128	-0.268
	R1T2	-0.628	-0.823	-0.944	-0.169	-0.428	0.679
	R1T3	-0.601	-0.478	-0.775	-0.728	-0.581	0.172
	R2T1	-0.859	0.274	-0.733	-0.004	0.072	0.500
	R2T2	-0.931	-0.768	-0.992	-0.771	-0.527	-0.589
	R2T3	-0.876	-0.139	-0.177	-0.308	-0.586	0.582
	R3T1	-0.922	0.296	-0.760	-0.248	-0.562	0.658
	R3T2	-0.905	0.103	-0.998	0.161	0.332	0.672
	R3T3	-0.933	-0.638	-0.997	-0.925	-0.580	-0.243

Stability_RMSE	R1T1	1.000	-0.081	-0.262	-0.586	-0.064	-0.137
	R1T2	-0.353	-0.525	-0.710	-0.085	-0.225	0.391
	R1T3	-0.334	-0.255	-0.475	-0.432	-0.320	0.087
	R2T1	-0.568	0.140	-0.436	-0.002	0.036	0.268
	R2T2	-0.682	-0.468	-0.878	-0.471	-0.285	-0.326
	R2T3	-0.591	-0.070	-0.089	-0.158	-0.324	0.321
	R3T1	-0.665	0.151	-0.461	-0.126	-0.308	0.375
	R3T2	-0.636	0.052	-0.937	0.081	0.171	0.386
	R3T3	-0.686	-0.361	-0.926	-0.670	-0.320	-0.123

Tablo EK-2: Sınıflandırma modellerinin performans sonuçları

Ölçüt	Veri	CV	METOD						
			KA	YSA	MARS	LR	DVM	MTS	BSF
PCC	Erkunt	R1T1	0.8065	0.8387	0.6774	0.6452	0.9032	0.4194	0.9355
		R1T2	0.9032	0.7419	0.6452	0.8065	1.0000	0.6452	0.8387
		R1T3	0.8000	0.8333	0.8333	0.7333	0.9667	0.6333	0.8667
		R2T1	0.8710	0.8065	0.8387	0.8065	1.0000	0.4516	0.8387
		R2T2	0.7097	0.7419	0.7097	0.6774	0.9677	0.3548	0.9032
		R2T3	0.8333	0.7333	0.8333	0.7667	0.9667	0.7333	0.9667
		R3T1	0.8065	0.8387	0.7419	0.8710	1.0000	0.4194	1.0000
		R3T2	0.9032	0.8710	0.9032	0.8387	0.9677	0.5806	0.9677
		R3T3	0.8667	0.8333	0.8667	0.7333	0.9667	0.7333	0.9000
	Tofaş	R1T1	0.8077	0.8077	0.6923	0.7308	0.7308	0.7692	0.8077
		R1T2	0.6538	0.7308	0.6538	0.7692	0.6154	0.6923	0.8462
		R1T3	0.7692	0.7692	0.6923	0.7308	0.6923	0.6923	0.7692
		R2T1	0.7692	0.6154	0.5769	0.8846	0.6538	0.7308	0.8846
		R2T2	0.6923	0.7308	0.7308	0.7692	0.6923	0.7308	0.7308
		R2T3	0.6538	0.6923	0.5769	0.7308	0.6923	0.6923	0.7692
		R3T1	0.6538	0.6538	0.6923	0.5769	0.6923	0.6923	0.7308
		R3T2	0.7692	0.7308	0.6923	0.8462	0.7692	0.6923	0.8077
		R3T3	0.6538	0.7692	0.6923	0.8077	0.6923	0.7308	0.8077
Precision	Erkunt	R1T1	0.3333	NaN	0.2727	0.0000	0.7500	0.2174	0.8000
		R1T2	1.0000	0.3333	0.0000	0.4000	1.0000	0.2000	0.5000
		R1T3	0.4000	0.5000	0.5000	0.0000	1.0000	0.2500	0.5714
		R2T1	0.5714	0.4286	NaN	0.4286	1.0000	0.1667	0.5000
		R2T2	0.2500	0.2000	0.2500	0.1429	1.0000	0.1739	0.6667
		R2T3	0.5000	0.0000	NaN	0.2500	0.8333	0.3333	1.0000
		R3T1	0.4000	NaN	0.2000	0.6000	1.0000	0.1905	1.0000
		R3T2	0.7500	0.6667	0.7500	0.5000	1.0000	0.2778	1.0000
		R3T3	0.6000	0.5000	0.5714	0.3333	1.0000	0.3333	0.6250
	Tofaş	R1T1	0.6667	0.7143	0.5000	0.6667	0.6667	1.0000	0.8000
		R1T2	0.4286	0.6000	0.4545	0.6000	0.0000	0.5000	0.7500
		R1T3	0.7500	0.7500	0.5000	0.6000	NaN	NaN	0.6250
		R2T1	0.7500	0.3750	0.2857	0.7778	0.3333	1.0000	0.7273
		R2T2	0.5000	0.6667	0.5556	0.7500	NaN	1.0000	0.6000
		R2T3	0.4545	NaN	0.3333	0.5385	NaN	NaN	0.5714
		R3T1	0.4000	0.4286	0.5000	0.3333	NaN	0.5000	0.5385
		R3T2	0.7500	0.5714	NaN	0.8333	0.7500	NaN	0.8000
		R3T3	0.4286	1.0000	0.5000	0.6667	0.5000	0.5714	0.8000
Recall	Erkunt	R1T1	0.2000	0.0000	0.6000	0.0000	0.6000	1.0000	0.8000
		R1T2	0.4000	0.6000	0.0000	0.4000	1.0000	0.4000	0.6000
		R1T3	0.4000	0.4000	0.4000	0.0000	0.8000	0.6000	0.8000

		R2T1	0.8000	0.6000	0.0000	0.6000	1.0000	0.6000	1.0000
		R2T2	0.4000	0.2000	0.4000	0.2000	0.8000	0.8000	0.8000
		R2T3	0.2000	0.0000	0.0000	0.2000	1.0000	0.6000	0.8000
		R3T1	0.4000	0.0000	0.2000	0.6000	1.0000	0.8000	1.0000
		R3T2	0.6000	0.4000	0.6000	0.4000	0.8000	1.0000	0.8000
		R3T3	0.6000	0.6000	0.8000	0.6000	0.8000	0.6000	1.0000
	Tofaş	R1T1	0.7500	0.6250	0.2500	0.2500	0.2500	0.2500	0.5000
		R1T2	0.3750	0.3750	0.6250	0.7500	0.0000	0.2500	0.7500
		R1T3	0.3750	0.3750	0.3750	0.3750	0.0000	0.0000	0.6250
		R2T1	0.3750	0.3750	0.2500	0.8750	0.1250	0.1250	1.0000
		R2T2	0.3750	0.2500	0.6250	0.3750	0.0000	0.1250	0.3750
		R2T3	0.6250	0.0000	0.3750	0.8750	0.0000	0.0000	1.0000
		R3T1	0.2500	0.3750	0.5000	0.3750	0.0000	0.1250	0.8750
		R3T2	0.3750	0.5000	0.0000	0.6250	0.3750	0.0000	0.5000
		R3T3	0.3750	0.2500	0.2500	0.7500	0.3750	0.5000	0.5000
F0.5	Erkunt	R1T1	0.2941	NaN	0.3061	NaN	0.7143	0.2577	0.8000
		R1T2	0.7692	0.3659	NaN	0.4000	1.0000	0.2222	0.5172
		R1T3	0.4000	0.4762	0.4762	NaN	0.9524	0.2830	0.6061
		R2T1	0.6061	0.4545	NaN	0.4545	1.0000	0.1948	0.5556
		R2T2	0.2703	0.2000	0.2703	0.1515	0.9524	0.2062	0.6897
		R2T3	0.3846	NaN	NaN	0.2381	0.8621	0.3659	0.9524
		R3T1	0.4000	NaN	0.2000	0.6000	1.0000	0.2247	1.0000
		R3T2	0.7143	0.5882	0.7143	0.4762	0.9524	0.3247	0.9524
		R3T3	0.6000	0.5172	0.6061	0.3659	0.9524	0.3659	0.6757
	Tofaş	R1T1	0.6818	0.6944	0.4167	0.5000	0.5000	0.6250	0.7143
		R1T2	0.4167	0.5357	0.4808	0.6250	NaN	0.4167	0.7500
		R1T3	0.6250	0.6250	0.4688	0.5357	NaN	NaN	0.6250
		R2T1	0.6250	0.3750	0.2778	0.7955	0.2500	0.4167	0.7692
		R2T2	0.4688	0.5000	0.5682	0.6250	NaN	0.4167	0.5357
		R2T3	0.4808	NaN	0.3409	0.5833	NaN	NaN	0.6250
		R3T1	0.3571	0.4167	0.5000	0.3409	NaN	0.3125	0.5833
		R3T2	0.6250	0.5556	NaN	0.7813	0.6250	NaN	0.7143
		R3T3	0.4167	0.6250	0.4167	0.6818	0.4688	0.5556	0.7143
F1	Erkunt	R1T1	0.2500	NaN	0.3750	NaN	0.6667	0.3571	0.8000
		R1T2	0.5714	0.4286	NaN	0.4000	1.0000	0.2667	0.5455
		R1T3	0.4000	0.4444	0.4444	NaN	0.8889	0.3529	0.6667
		R2T1	0.6667	0.5000	NaN	0.5000	1.0000	0.2609	0.6667
		R2T2	0.3077	0.2000	0.3077	0.1667	0.8889	0.2857	0.7273
		R2T3	0.2857	NaN	NaN	0.2222	0.9091	0.4286	0.8889
		R3T1	0.4000	NaN	0.2000	0.6000	1.0000	0.3077	1.0000
		R3T2	0.6667	0.5000	0.6667	0.4444	0.8889	0.4348	0.8889
		R3T3	0.6000	0.5455	0.6667	0.4286	0.8889	0.4286	0.7692
	Tofaş	R1T1	0.7059	0.6667	0.3333	0.3636	0.3636	0.4000	0.6154
		R1T2	0.4000	0.4615	0.5263	0.6667	NaN	0.3333	0.7500

		R1T3	0.5000	0.5000	0.4286	0.4615	NaN	NaN	0.6250
		R2T1	0.5000	0.3750	0.2667	0.8235	0.1818	0.2222	0.8421
		R2T2	0.4286	0.3636	0.5882	0.5000	NaN	0.2222	0.4615
		R2T3	0.5263	NaN	0.3529	0.6667	NaN	NaN	0.7273
		R3T1	0.3077	0.4000	0.5000	0.3529	NaN	0.2000	0.6667
		R3T2	0.5000	0.5333	NaN	0.7143	0.5000	NaN	0.6154
		R3T3	0.4000	0.4000	0.3333	0.7059	0.4286	0.5333	0.6154
F2	Erkunt	R1T1	0.2174	NaN	0.4839	NaN	0.6250	0.5814	0.8000
		R1T2	0.4545	0.5172	NaN	0.4000	1.0000	0.3333	0.5769
		R1T3	0.4000	0.4167	0.4167	NaN	0.8333	0.4688	0.7407
		R2T1	0.7407	0.5556	NaN	0.5556	1.0000	0.3947	0.8333
		R2T2	0.3571	0.2000	0.3571	0.1852	0.8333	0.4651	0.7692
		R2T3	0.2273	NaN	NaN	0.2083	0.9615	0.5172	0.8333
		R3T1	0.4000	NaN	0.2000	0.6000	1.0000	0.4878	1.0000
		R3T2	0.6250	0.4348	0.6250	0.4167	0.8333	0.6579	0.8333
		R3T3	0.6000	0.5769	0.7407	0.5172	0.8333	0.5172	0.8929
	Tofaş	R1T1	0.7317	0.6410	0.2778	0.2857	0.2857	0.2941	0.5405
		R1T2	0.3846	0.4054	0.5814	0.7143	NaN	0.2778	0.7500
		R1T3	0.4167	0.4167	0.3947	0.4054	NaN	NaN	0.6250
		R2T1	0.4167	0.3750	0.2564	0.8537	0.1429	0.1515	0.9302
		R2T2	0.3947	0.2857	0.6098	0.4167	NaN	0.1515	0.4054
		R2T3	0.5814	NaN	0.3659	0.7778	NaN	NaN	0.8696
		R3T1	0.2703	0.3846	0.5000	0.3659	NaN	0.1471	0.7778
		R3T2	0.4167	0.5128	NaN	0.6579	0.4167	NaN	0.5405
		R3T3	0.3846	0.2941	0.2778	0.7317	0.3947	0.5128	0.5405
LOR	Erkunt	R1T1	1.0986	NaN	1.2164	Inf	3.6243	Inf	4.6052
		R1T2	Inf	1.6094	-Inf	1.6314	Inf	0.4055	2.4423
		R1T3	1.5870	2.0369	2.0369	Inf	Inf	0.9808	3.3787
		R2T1	3.4232	2.1102	NaN	2.1102	Inf	0.0953	Inf
		R2T2	0.7985	0.3185	0.7985	-0.1823	Inf	0.3878	3.8712
		R2T3	1.7918	-inf	NaN	0.6061	Inf	1.5581	Inf
		R3T1	1.6314	NaN	0.3185	2.8904	Inf	0.7503	Inf
		R3T2	3.6243	2.8134	3.6243	2.0794	Inf	Inf	Inf
		R3T3	2.8478	2.3979	3.3787	1.5581	Inf	1.5581	Inf
	Tofaş	R1T1	2.7081	2.5903	0.9808	1.7346	1.7346	Inf	2.8332
		R1T2	0.7419	1.5686	1.2040	2.3514	-Inf	0.9808	3.1781
		R1T3	2.3224	2.3224	1.0986	1.5686	NaN	NaN	2.1203
		R2T1	2.3224	0.4447	-0.1431	4.0254	0.1335	Inf	Inf
		R2T2	1.0986	1.7346	1.7636	2.3224	NaN	Inf	1.5686
		R2T3	1.2040	NaN	0.1823	2.6391	NaN	NaN	Inf
		R3T1	0.5108	0.7419	1.2528	0.1823	NaN	0.8873	2.6391
		R3T2	2.3224	1.6094	NaN	3.3440	2.3224	NaN	2.8332
		R3T3	0.7419	Inf	0.9808	2.7081	1.0986	1.6094	2.8332
Kappa	Erkunt	R1T1	0.1468	0.0000	0.1969	-0.2135	0.6109	0.1254	0.7615

		R1T2	0.5279	0.2791	-0.2135	0.2846	1.0000	0.0658	0.4484
		R1T3	0.2800	0.3478	0.3478	-0.1429	0.8696	0.1538	0.5862
		R2T1	0.5894	0.3841	0.0000	0.3841	1.0000	0.0113	0.5753
		R2T2	0.1362	0.0462	0.1362	-0.0265	0.8703	0.0282	0.6690
		R2T3	0.2105	-0.1429	0.0000	0.0870	0.8889	0.2727	0.8696
		R3T1	0.2846	0.0000	0.0462	0.5231	1.0000	0.0638	1.0000
		R3T2	0.6109	0.4312	0.6109	0.3515	0.8703	0.2439	0.8703
		R3T3	0.5200	0.4444	0.5862	0.2727	0.8696	0.2727	0.7097
	Tofaş	R1T1	0.5638	0.5324	0.1613	0.2353	0.2353	0.3158	0.4961
		R1T2	0.1583	0.2946	0.2642	0.4935	-0.1404	0.1613	0.6389
		R1T3	0.3710	0.3710	0.2239	0.2946	0.0000	0.0000	0.4583
		R2T1	0.3710	0.0972	-0.0288	0.7383	0.0168	0.1651	0.7547
		R2T2	0.2239	0.2353	0.3893	0.3710	0.0000	0.1651	0.2946
		R2T3	0.2642	0.0000	0.0403	0.4615	0.0000	0.0000	0.5517
		R3T1	0.0930	0.1583	0.2778	0.0403	0.0000	0.0877	0.4615
		R3T2	0.3710	0.3453	0.0000	0.6119	0.3710	0.0000	0.4961
		R3T3	0.1583	0.3158	0.1613	0.5638	0.2239	0.3453	0.4961
Specificity	Erkunt	R1T1	0.9231	1.0000	0.6923	0.7692	0.9615	0.3077	0.9615
		R1T2	1.0000	0.7692	0.7692	0.8846	1.0000	0.6923	0.8846
		R1T3	0.8800	0.9200	0.9200	0.8800	1.0000	0.6400	0.8800
		R2T1	0.8846	0.8462	1.0000	0.8462	1.0000	0.4231	0.8077
		R2T2	0.7692	0.8462	0.7692	0.7692	1.0000	0.2692	0.9231
		R2T3	0.9600	0.8800	1.0000	0.8800	0.9600	0.7600	1.0000
		R3T1	0.8846	1.0000	0.8462	0.9231	1.0000	0.3462	1.0000
		R3T2	0.9615	0.9615	0.9615	0.9231	1.0000	0.5000	1.0000
		R3T3	0.9200	0.8800	0.8800	0.7600	1.0000	0.7600	0.8800
	Tofaş	R1T1	0.8333	0.8889	0.8889	0.9444	0.9444	1.0000	0.9444
		R1T2	0.7778	0.8889	0.6667	0.7778	0.8889	0.8889	0.8889
		R1T3	0.9444	0.9444	0.8333	0.8889	1.0000	1.0000	0.8333
		R2T1	0.9444	0.7222	0.7222	0.8889	0.8889	1.0000	0.8333
		R2T2	0.8333	0.9444	0.7778	0.9444	1.0000	1.0000	0.8889
		R2T3	0.6667	1.0000	0.6667	0.6667	1.0000	1.0000	0.6667
		R3T1	0.8333	0.7778	0.7778	0.6667	1.0000	0.9444	0.6667
		R3T2	0.9444	0.8333	1.0000	0.9444	0.9444	1.0000	0.9444
		R3T3	0.7778	1.0000	0.8889	0.8333	0.8333	0.8333	0.9444
Stability_PCC	Erkunt	R1T1	0.0990	-0.0016	0.1923	0.1915	0.0508	0.3951	0.0333
		R1T2	0.0342	0.0693	0.2157	0.0466	-0.0083	0.1831	0.0877
		R1T3	0.1111	0.0909	0.0746	0.0670	-0.0078	0.1842	0.0714
		R2T1	0.0173	0.0990	0.0362	0.0735	0.0000	0.3560	0.0877
		R2T2	0.1618	0.0972	0.1536	0.1679	-0.0088	0.4495	0.0508
		R2T3	0.0828	0.1038	0.0746	0.0992	0.0088	0.0858	0.0169
		R3T1	0.0990	-0.0016	0.1481	0.0173	-0.0252	0.4091	0.0000
		R3T2	0.0426	0.0173	0.0426	0.0452	0.0164	0.2335	0.0164
		R3T3	0.0551	0.0221	0.0714	0.1295	-0.0078	0.1038	0.0526

	Tofaş	R1T1	0.0156	0.0877	0.1818	0.1078	0.1556	0.1304	0.1064
		R1T2	0.1818	-0.0189	0.1421	0.0722	0.1047	0.1444	0.0246
		R1T3	0.0510	0.1118	0.1139	0.1078	-0.0052	-0.0052	0.0924
		R2T1	-0.0065	0.2202	0.1924	0.0127	0.2093	0.1276	0.0228
		R2T2	0.0924	0.1078	0.0872	0.0722	-0.0052	0.1556	0.1556
		R2T3	0.1624	0.0338	0.2128	0.0976	0.1818	-0.0052	0.0924
		R3T1	0.1315	0.1421	0.1033	0.2322	-0.0052	0.1345	0.1556
		R3T2	-0.0065	0.1078	-0.0052	0.0141	0.1304	-0.0052	0.0266
		R3T3	0.1096	0.0400	0.1033	0.0581	0.1818	0.0765	0.0156
AUC	Erkunt	R1T1	0.7730	0.7690	0.6462	0.6154	0.7808	0.4923	0.8141
		R1T2	0.8580	0.7690	0.6154	0.6090	1.0000	0.4462	0.6923
		R1T3	0.7760	0.6240	0.6600	0.5600	0.9000	0.6080	0.7733
		R2T1	0.8230	0.8920	0.5000	0.6731	1.0000	0.3615	0.8205
		R2T2	0.7270	0.6080	0.5846	0.5513	0.9000	0.3385	0.7949
		R2T3	0.8360	0.8000	0.5000	0.5233	0.9800	0.6480	0.8333
		R3T1	0.8460	0.9000	0.5231	0.7949	1.0000	0.4385	0.9167
		R3T2	0.9000	0.7620	0.7808	0.6282	0.9000	0.6462	0.8333
		R3T3	0.8880	0.8400	0.8400	0.6300	0.9000	0.6240	0.8567
	Tofaş	R1T1	0.7120	0.7780	0.5556	0.5724	0.5972	0.4931	0.6974
		R1T2	0.6110	0.7220	0.7708	0.7434	0.5556	0.4583	0.7961
		R1T3	0.6650	0.8090	0.5903	0.6086	0.5000	0.5139	0.7336
		R2T1	0.6560	0.5760	0.6667	0.8586	0.5069	0.4236	0.8947
		R2T2	0.5970	0.7990	0.7951	0.6349	0.5000	0.4792	0.6086
		R2T3	0.6420	0.6940	0.5069	0.7533	0.5000	0.5972	0.8158
		R3T1	0.5490	0.7080	0.6389	0.5296	0.5000	0.5694	0.7533
		R3T2	0.6600	0.8060	0.6181	0.7599	0.6597	0.5139	0.6974
		R3T3	0.6390	0.7640	0.5660	0.7697	0.6042	0.5625	0.7237

TÜBİTAK
PROJE ÖZET BİLGİ FORMU

Proje No: 105M138

Proje Başlığı: Kalite İyileştirmede Veri Madenciliği Kullanımı ve Geliştirilmesi

Proje Yürütücüsü ve Araştırmacılar: Prof.Dr. Gülser KÖKSAL, Doç.Dr. İnci BATMAZ, Prof.Dr. Bülent KARASÖZEN, Prof. Dr. Sinan KAYALIGİL, Doç.Dr. Murat Caner TESTİK, Prof. Dr. Nur Evin ÖZDEMİREL, Prof. Dr. Gerhard Wilhelm WEBER, Berna BAKIR, Fatma GÜNTÜRKÜN, İlker Arif İPEKÇİ, Başak ÖZTÜRK, Fatma YERLİKAYA

Projenin Yürütüldüğü Kuruluş ve Adresi: Orta Doğu Teknik Üniversitesi, Endüstri Mühendisliği Bölümü, 06531 Ankara

Destekleyen Kuruluş(ların) Adı ve Adresi:

TÜBİTAK, Tunus Caddesi No:80 06100 Kavaklıdere, Ankara
ODTÜ, İnönü Bulvarı, 06531 Ankara
Veri desteği sağlayan kuruluşlar:
ERKUNT Sanayi A.Ş., İstanbul Yolu 8. km. Ankara
TOFAŞ Türk Otomobil Firması A.Ş., Yeni Yalova yolu Cad. No:574 16369 Bursa
VESTEL Elektronik A.Ş., Organize Sanayi Bölgesi, 45030 Manisa
Bilgi desteği sağlayan kuruluşlar:
AKSA Akrilik Kimya Sanayi A.Ş., Denizçalı Köyü Karamürsel Yolu 13 Km. P.K.115 Yalova
PETKİM Petrokimya A.Ş., PK. 12 35800 Aliağa, İzmir

SPAC Ltd. Şti., ODTÜ Kampüsü, Silikon Binası ZK 19, 06531 ODTÜ, ANKARA

Projenin Başlangıç ve Bitiş Tarihleri: 1.5.2006-30.6.2009

Öz (en çok 70 kelime)

Bu projede amaç, imalat sanayi kuruluşlarında ürün ve süreçlerin kalitesini iyileştirmeye yönelik veri madenciliği (VM) yaklaşımlarını belirlemek ve daha etkili yaklaşımlar geliştirmektir. Buna yönelik olarak geniş bir literatür taraması yapılmış ve çeşitli kuruluşlar ziyaret edilmiştir. Bunlardan elde veriler üzerinde yapılan uygulamalarla en uygun VM metotları belirlenmiştir. Ayrıca, uygulama aşamasında karşılaşılan bazı problemlerin giderilmesi ve mevcut yöntemlerin kullanım kolaylığı ve/veya etkililiğinin artırılması yönünde çeşitli çözümler ve metotlar geliştirilmiştir.

Anahtar Kelimeler: Kalite iyileştirme, kalite kontrol, veri madenciliği, kümeleme, tahmin etme, regresyon, sınıflandırma, parametre optimizasyonu, birliktelik analizi

Fikri Ürün Bildirim Formu Sunuldu mu? **Evet** ☐ **Gerekli Değil** ☒
Fikri Ürün Bildirim Formu'nun tesliminden sonra 3 ay içerisinde patent başvurusu yapılmalıdır.

Projeden Yapılan Yayınlar:

Uluslararası makale

UAM1) Köksal, G., Batmaz, İ. ve Testik, M.C., Haziran 2009, "A review of data mining applications for description, prediction, classification and optimisation of quality in manufacturing industry", *International Journal of Management Reviews* dergisinde, birinci revizyonu değerlendirilmektedir.

UAM2) Weber, G.W., Batmaz, İ., Köksal, G., Taylan, P, Yerlikaya Özkurt, F., Eylül 2009, CMARS: A New Contribution to Nonparametric Regression with Multivariate Adaptive Regression

Splines Supported by Continuous Optimisation, *Journal of Computational and Applied Mathematics* dergisinde değerlendirilmektedir.

Ulusal Makale

ULM1) Köksal, G., Batmaz, İ, Testik, M.C. ve Güntürkün, F., Mart 2010, İmalat Sektöründe Kalite İyileştirmede Veri Madenciliği Tekniklerinin Kullanımı, *Verimlilik Dergisi*'nde yayımlanacaktır.

Uluslararası Konferans Bildirileri

UAB1) Özer, G., Köksal, G., Batmaz, İ., Türker Bayrak, Ö., Kılıç, T., Kartal, E., Ekim 2009, "Classification Models Based on Tanaka's Fuzzy Linear Regression Approach: The Case of Customer Satisfaction Modeling", 1st International Fuzzy Systems Symposium Proceedings, Ankara. (yayımlanacaktır)

UAB2) Köksal, G., Batmaz, İ. ve Kartal, E., 2008, Developing a Classification Model For Customer Satisfaction with a Driver's Seat: A Comparative Case Study, Proceedings of 6th International Symposium on Intelligent and Manufacturing Systems, Sakarya, 520-530.

UAB3) Taylan, P., Weber, G.W. ve Yerlikaya, F., Mayıs 20-23, 2008, Continuous optimization applied in MARS for modern applications in finance, science and technology, ISI Proceedings of 20th Mini-EURO Conference Continuous Optimization and Knowledge-Based Technologies, Neringa, Lithuania, 317-322.

UAB4) Bakır, B., Batmaz, İ., Güntürkün, F.A., İpekçi, İ. A., Köksal, G., Özdemirel, N. E., "Defect Cause Modeling with Decision Tree and Reression Analysis", Proceedings of XVII. International Conference on Computer and Information Science and Engineering, Cairo, Egypt, December 8-10, 2006, Vol. 17, pp. 66-269, ISBN 975-00803-7-8.

UAB5) Akteke-Öztürk, B., Weber, G.-W. and Kropat, E., Continuous Optimization Approaches for Clustering via Minimum Sum of Squares, ISI Proceedings of 20th Mini-EURO Conference Continuous Optimization and Knowledge-Based Technologies, Neringa, Lithuania, May 20-23, 2008, pp.253-258.

Ulusal Konferans Bildirileri

ULB1) Köksal, G., Ayhan, D. ve Yenidünya, B., Kasım 2009, Kalite Sınıflandırma ve Eniyileme için Mahalanobis Taguchi Sistemi Yaklaşımları, *18. Kalite Kongresi*'nde sunulacaktır.

ULB2) Jabarnejad, M. ve Testik, M.C., Haziran 2009, Grouping and Pruning Association Rules in Data Mining, *Yöneylem Araştırması ve Endüstri Mühendisliği 29. Ulusal Kongresi Bildirileri*, CD-ROM, Ankara.

ULB3) Yerlikaya, F., Weber, G. W., Taylan, P., Batmaz, İ. ve Köksal, G., MARS Algoritmasında Tikhonov Düzenlemesi ve Çok Amaçlı Optimizasyon Kullanımı, Yöneylem Araştırması ve Endüstri Mühendisliği 28. Ulusal Kongresi Bildirileri, İstanbul, 1 Temmuz 2008, CD-ROM.

ULB4) Akteke-Öztürk, B., Weber, G.W., Kayalıgil, S., Temmuz 2007, Kalite İyileştirmede Veri Kümeleme: Döküm Endüstrisinde Bir Uygulama, Yöneylem Araştırması ve Endüstri Mühendisliği 27. Ulusal Kongresi (YA/EM 2007) Bildiriler Kitabı, 1207-1212, İzmir.

Uluslararası Konferans Bildiri Özetleri / Sunumları

UAS1) Köksal, G., Anaklı, Z., Batmaz, İ., Yerlikaya Özkurt, F., Testik, M.C., Kartal, E., Kayalıgil, S., Bakır, B., Karasakal, E., Temmuz 2009, "Comparison of Data Mining Algorithms for Classification and Prediction in Quality Improvement", *23rd European Conference on Operational Research*, Bonn, 39.

UAS2) Bakır, B., Ayhan, D., Yenidünya, B., Köksal, G., Temmuz 2009, "A Re-sampling Approach and Its Applications for MTS Classification Based on Imbalanced Data", *23rd European*

Conference on Operational Research, Bonn, 39-40.

UAS3) Akteke-Öztürk, B., Köksal, G., Weber, G.W., Temmuz 2009, "Nonsmooth Optimization of Desirability Functions by MSG Algorithm", *23rd European Conference on Operational Research*, Bonn, 40.

UAS4) Yenidünya, B., Köksal, G., Temmuz 2009, "Design Parameter Optimization Using MTS", *23rd European Conference on Operational Research*, Bonn, 63.

UAS5) Özer, G., Kılıç, T., Kartal, E., Batmaz, İ., Türker Bayrak, Ö., Köksal, G., Türkşen B., Temmuz 2009, "A Nonparametric Improved Fuzzy Classifier Function Approach for Classification Based on Customer Satisfaction Survey Data", *23rd European Conference on Operational Research*, Bonn, 63.

UAS6) Ayhan, D., Köksal, G., Temmuz 2009. "Multi-class MTS Classification Algorithms and Their Applications", *23rd European Conference on Operational Research*, Bonn, 63-64.

UAS7) Yerlikaya Özkurt, F., Taylan, P., Batmaz, İ., Köksal, G., Weber, G.W., Temmuz 2009, "A Modification of MARS by Tikhonov Regularization and Conic Quadratic Programming for Modeling Quality Data", *23rd European Conference on Operational Research*, Bonn, 299.

UAS8) İpekçi, İ., Bakır, B., Batmaz, İ. ve Testik, M.C., Optimization of Casting Process Using Artificial Neural Networks, International Conference on Multivariate Statistical Modeling and High Dimensional Data Mining, Kayseri, 19-23 Haziran 2008, konferans bildirileri kitabında basımı için değerlendirilmektedir. (?)

UAS9) Köksal, G., Testik, M.C., Güntürkün, F.A., Batmaz, İ., Temmuz 2007, Data Mining Applications in Quality Improvement: A Tutorial and a Literature Review, Book of Abstracts: 22nd European Conference on Operational Research, Prag, 148.

UAS10) İpekçi, A.İ., Köksal, G., Karasakal, E., Özdemirel, N.E., Testik, M.C., Temmuz 2007, Multi Response Decision Tree Approach Applied To A Discrete Manufacturing Quality Improvement Problem, Book of Abstracts: 22nd European Conference on Operational Research, Prag, 148.

UAS11) İpekçi, A.İ., Bakır, B., Batmaz, İ., Testik, M.C. ve Özdemirel, N.E., Ağustos 2007, Defect Cause Modeling with Data Mining: Decision Trees and Neural Networks, Book of Abstracts of 56th Session of the International Statistical Institute, Lizbon, Portekiz, 412.

Ulusal Konferans Bildiri Özetleri / Sunumları

ULS1) Yenidünya, B., Köksal, G., Haziran 2009. "Mahalanobis Taguchi Sistemi ile Tasarım Optimizasyonu: İki Düzeyli Çıktı Değişkeni Durumu", *Yöneylem Araştırması ve Endüstri Mühendisliği 29. Ulusal Kongresi*, Ankara.

ULS2) Kılıç, T., Özer, G., Kartal, E., Türker Bayrak, Ö., Batmaz, İ., Köksal, G., Haziran 2009. "Bir Döküm Sürecinin Kalitesinin Bulanık Regresyon Yöntemi ile Modellenmesi", *Yöneylem Araştırması ve Endüstri Mühendisliği 29. Ulusal Kongresi*, Ankara.

ULS3) Akteke Öztürk, B., Köksal, G., Weber, G.W., Haziran 2009. "Çekicilik Fonksiyonlarının Bileşke Fonksiyonları Olarak Çözümlenmesi", *Yöneylem Araştırması ve Endüstri Mühendisliği 29. Ulusal Kongresi*, Ankara.

ULS4) Yerlikaya Özkurt, F., Batmaz, İ., Taylan, P., Köksal, G., Weber, G.W., Haziran 2009. "CMARS ile Doğrusal Olmayan Veri Yapılarının Modellenmesi", *Yöneylem Araştırması ve Endüstri Mühendisliği 29. Ulusal Kongresi*, Ankara.

ULS5) Akteke-Öztürk, B., Köksal, G. ve Weber, G. W., Çekicilik Fonksiyonları: Pürüzlü Optimizasyon ile Yeni Yaklaşımlar, Yöneylem Araştırması ve Endüstri Mühendisliği 28. Ulusal Kongresi Bildiri Özetleri Kitabı, İstanbul, 30 Haziran 2008, sayfa 25.

ULS6) Anaklı, Z., Bakır B., İpekçi, İ., Köksal, G. ve Kayalıgil, S., Kalite Geliştirmede Veri Madenciliği: Sınıflandırmada Performans Ölçüleri Etkileşim Değerlendirmesi, Yöneylem

Araştırması ve Endüstri Mühendisliği 28. Ulusal Kongresi Bildiri Özetleri Kitabı, İstanbul, 2 Temmuz 2008, sayfa 141.

ULS7) Bakır, B. Ve Köksal, G., Ürün Tasarımı ile İlgili Müşteri Sesinin Modellenmesi: Bayes İnanç Ağları, Yöneylem Araştırması ve Endüstri Mühendisliği 28. Ulusal Kongresi Bildiri Özetleri Kitabı, İstanbul, 2 Temmuz 2008, sayfa 142.

ULS8) Ayhan, D. ve Köksal, G., Sınıflandırmada Mahalanobis Taguchi Sistem Yaklaşımı ve Lojistik Regresyon ile Karşılaştırılması, Yöneylem Araştırması ve Endüstri Mühendisliği 28. Ulusal Kongresi Bildiri Özetleri, İstanbul, 2 Temmuz 2008, sayfa 142.

ULS9) Köksal, G., Testik, M.C., Güntürkün, F.A., Batmaz, İ., Kasım 2007, Kalite İyileştirmede Veri Madenciliği Yaklaşımları ve Bir Uygulama, 16. Ulusal Kalite Kongresi Sunumları, http://www.kalder.org/genel/16kongre/GULSER%20KOKSAL.ppt, İstanbul.

Tezler

T1) Anaklı, Zeynep, Aralık 2009, A Comparison of DM Algorithms for Prediction and Classification of Quality Data. Yüksek lisans tezi. Orta Doğu Teknik Üniversitesi, Endüstri Mühendisliği Bölümü, Ankara. (yayımlanması beklenmektedir)

T2) Jabarnejad, Masood, Aralık 2009, A Method for Grouping and Pruning of Association Rules. Yüksek lisans tezi. Orta Doğu Teknik Üniversitesi, Endüstri Mühendisliği Bölümü, Ankara. (yayımlanması beklenmektedir)

T3) Öztürk, Başak, Aralık 2010, Non-smooth Optimization of Desirability Functions. Doktora tezi, Orta Doğu Teknik Üniversitesi, Uygulamalı Matematik Enstitüsü, Ankara. (yayımlanması beklenmektedir)

T4) Avcı, Ezgi, Eylül 2009, A Comparison of Robust Regression Methods for Outliers. Yüksek lisans tezi. Orta Doğu Teknik Üniversitesi, Endüstri Mühendisliği Bölümü, Ankara.

T5) Kılıç, Tuna, Eylül 2009, Fuzzy Linear Regression: Performance Analysis of HBS2 Method for Several Variables, and Nonparametric Improved Fuzzy Functions. Yüksek lisans tezi. Çankaya Üniversitesi, Endüstri Mühendisliği Bölümü, Ankara.

T6) Özer, G., 2009, Fuzzy Classification Models Based on Tanaka's Fuzzy Linear Regression Approach and Nonparametric Improved Fuzzy Classifier Functions. Yüksek lisans tezi. Orta Doğu Teknik Üniversitesi, Endüstri Mühendisliği Bölümü, Ankara.

T7) Yenidünya, B., 2009, Robust Design With Binary Response Using Mahalanobis Taguchi System. Yüksek lisans tezi. Orta Doğu Teknik Üniversitesi, Endüstri Mühendisliği Bölümü, Ankara.

T8) Ayhan, D., 2009, Multi-Class Classification Methods Utilizing Mahalanobis Taguchi System and a Re-Sampling Approach For Imbalanced Data Sets. Yüksek lisans tezi, Orta Doğu Teknik Üniversitesi, Endüstri Mühendisliği Bölümü, Ankara.

T9) Yerlikaya, F., 2008, A New Contribution to Nonlinear Robust Regression and Classification With MARS and Its Applications to Data Mining For Quality Control in Manufacturing. Yüksek lisans tezi. Orta Doğu Teknik Üniversitesi, Uygulamalı Matematik Enstitüsü, Ankara. (**ODTÜ, Uygulamalı Matematik Enstitüsü, Bilimsel Hesaplama Yüksek Lisans Programı En İyi Tez Ödülü, 2009**)

T10) Güntürkün, F., 2007, A Comprehensive Review of Data Mining Applications in Quality Improvement and A Case Study. Yüksek Lisans Tezi. ODTÜ, İstatistik Bölümü, Ankara.

T11) Bakır, B., 2007, Defect Cause Modeling with Decision Tree and Regression Analysis: A Case Study in Casting Industry. Yüksek lisans tezi. ODTÜ, Enformatik Enstitüsü, Ankara.

Diğer

D1) Bakır, B., Köksal, G., Ayhan, D. ve Yenidünya, B., Nisan 2009, A SMOTE Based Re-sampling Approach Optimized for MTS Classification of Imbalanced Data, *Workshop on Recent Developments in Applied Probability and Statistics*, Ankara. (bildiri özeti / sunum)

D2) Weber, G. W., Taylan, P., Özöğür, S. ve Akteke-Öztürk, B., 2007, Statistical Learning and Optimization Methods in Data Mining, *Recent Advances in Statistics*, Editörler: H.Ö. Ayhan, İ. Batmaz. TÜİK, 181-195. (bildiri)

D3) Batmaz, İ., 2007, Data Mining Applications on Manufacturing Data: A Casting Quality Improvement Case, *Recent Advances in Statistics*, Editörler: H.Ö. Ayhan, İ. Batmaz. TÜİK, 197-206. (bildiri)

D4) Akteke-Öztürk, B. ve Weber, G. W., "A Survey and Results on Semidefinite and Nonsmooth Optimization for Minimum Sum of Squared Distances Problem", ön baskı no. 1, Uygulamalı Matematik Enstitüsü, ODTÜ, 2006.

D5) Bakır, B. Defect Cause Modeling with Data Mining: Decision Trees and Neural Networks, İstatistik Bölümü semineri, ODTÜ, 19 Nisan 2007. (sunum)

D6) Weber, G. W., Köksal, G. ve Kayalıgil, S., Ağustos 2007, Data Mining in Quality Improvement, 2nd Open Summer School-Seminar, Kiev. (sunum)

D7) Köksal, G. ve Kayalıgil, S., Data Mining in Quality Improvement, METU-IE and TU/e 3rd Joint Workshop, Ankara, 16 Haziran 2007. (sunum)