

**EŞ İŞLEM STRATEJİSİ YÖNTEMİYLE
İMKB'DE PORTFÖY OLUŞTURMADA
VERİ MADENCİLİĞİ UYGULAMASI**

Levent TERLEMEZ
Doktora Tezi

İstatistik Anabilim Dalı

Şubat - 2008

JÜRİ VE ENSTİTÜ ONAYI

Levent TERLEMEZ'in “Eş İşlem Stratejisi Yöntemiyle İMKB’de Portföy Oluşturmada Veri Madenciliği Uygulaması” başlıklı İstatistik Anabilim Dalındaki, Doktora Tezi 07/01/2008 Tarihinde, aşağıdaki jüri tarafından Anadolu Üniversitesi Lisansüstü Eğitim Öğretim ve Sınav Yönetmeliğinin ilgili maddeleri uyarınca değerlendirilerek kabul edilmiştir.

	Adı Soyadı	İmza
Üye (Tez Danışmanı)	: Yard. Doç. Dr. ATILLA ASLANARGUN	
Üye	: Prof. Dr. ALİ FUAT YÜZER	
Üye	: Prof. Dr. AHMET ÖZMEN	
Üye	: Doç. Dr. CAN CENGİZ ÇELİKOĞLU	
Üye	: Doç. Dr. MUHARREM AFŞAR	

**Anadolu Üniversitesi Fen Bilimleri Enstitüsü Yönetim Kurulu’nun
..... tarih ve sayılı kararıyla onaylanmıştır.**

Enstitü Müdürü

ÖZET

Doktora Tezi

EŞ İŞLEM STRATEJİSİ YÖNTEMİYLE İMKB'DE PORTFÖY OLUŞTURMADA VERİ MADENCİLİĞİ UYGULAMASI

Levent TERLEMEZ

**Anadolu Üniversitesi
Fen Bilimleri Enstitüsü
İstatistik Anabilim Dalı**

**Danışman: Yard. Doç. Dr. Atilla ASLANARGUN
2008, 139 sayfa**

Günümüzde kalemle tutulan bir nottan uzaktan algılama yeteneğine sahip bir uyduya kadar uzanan veri kaynaklarındaki değişimler veri akışının hızlanmasına, veri yığınlarının artmasına ve verinin karmaşıklaşmasına neden olmaktadır. Bunun sonucunda, verinin derlenmesinde, işlenmesinde ve yorumlanmasında büyük zorluklarla karşılaşmaktadır. Bu nedenle, veri kaynaklarından gelen veriyi derlemek, yorumlamaya açık hale getirmek ve işlerlik kazandırmak için otomatik sistemlere ihtiyaç duyulmaktadır.

Veri madenciliği akış halindeki veriyi derlemek, görselleştirmek, analiz etmek ve karar sürecine uyarlamak için çeşitli modelleri bir araya getirerek bir sistem oluşturulmasını sağlayan bir süreçtir.

Bu çalışmada, İstanbul Menkul Kıymetler Borsası'nda işlem gören hisse senetleri ile bir portföy oluşturmak ve al – sat zamanlaması için eş işlem (pairs trading) stratejisini kullanan bir veri madenciliği sistemi oluşturulmuştur.

Anahtar Kelimeler: Veri madenciliği, İMKB, Portföy, Eş işlem stratejisi

ABSTRACT

PhD Dissertation

A DATA MINING APPLICATION VIA PAIRS TRADING STRATEGY ON PORTFOLIO CONSTRUCTION AT ISTANBUL STOCK EXCHANGE

Levent TERLEMEZ

**Anadolu University
Graduate School of Sciences
Statistics Program**

**Supervisor: Asst. Prof. Dr. Atilla ASLANARGUN
2008, 139 pages**

Today, changes in data sources varying from taking notes to satellites having remote sensing capability cause data streaming to accelerate, data piles to increase and make data be more complex. Hence, in data compilation, process and in interpretation are encountered difficulties. Therefore, it is needed to develop automatic systems to compile data from data sources, make clear for interpretation and put into action.

Data mining is a process to provide a system that compiles, virtualizes and analysis the streaming data and adapt to decision process.

In this study, a data mining system is created to form a portfolio with the stocks from Istanbul Stock Exchange using pairs trading strategy for stock selection and buy – sell timing.

Keywords: Data mining, ISE, Portfolio, Pair trading strategy

TEŞEKKÜR

Bu çalışmanın ortaya çıkmasında katkılarını ve yardımlarını esirgemeyen danışman hocam Sayın Yard. Doç. Dr. Atilla ASLANARGUN'a; değerli katkılarıyla araştırmama yön veren tez izleme komitesi üyesi hocam Prof. Dr. Ali Fuat YÜZER ve Sayın Prof. Dr. Ahmet ÖZMEN'e; manevi desteğini, yardımlarını ve anlayışını esirgemeyen Sayın Doç. Dr. Can Cengiz ÇELİKOĞLU ve Sayın Doç. Dr. Muharrem AFŞAR'a en içten teşekkürlerimi sunarım.

Bu bilimsel sürecin her aşamasında idari ve bilimsel desteğini esirgemeyen, bölüm başkanımız Sayın Prof. Dr. Embiye AĞAOĞLU'na saygılarımı sunarım.

Çalışma konumda deneyim ve bilgileriyle beni destekleyen, önemli bilimsel katkılar sağlayan hocam Sayın Yard. Doç. Dr. Metin ÇOŞKUN'a ve Sayın Yard. Doç. Dr. Fatih TEMİZEL'e; değerli yardımlarını ve destekleriyle yanımda olan hocalarım Sayın Yard. Doç. Dr. Fikret ER'e, Sayın Yard. Doç. Dr. Harun SÖNMEZ'e, Sayın Yard. Doç. Dr. Zerrin AŞAN'a ve Sayın Yard. Doç. Dr. K. Özgür PEKER'e ve sevgili arkadaşlarım Sayın Öğr. Gör. Dr. Sevil ŞENTÜRK'e ve Sayın Arş. Gör. Halil ERYILMAZ'a en içten teşekkürlerimi sunarım.

Ayrıca çalışmalarımnda manevi desteklerini esirgemeyen değerli bölüm hocalarıma ve arkadaşlarıma teşekkür ederim.

Levent TERLEMEZ

Şubat 2008

Sabırla, anlayıřla ve sevgiyle beni destekleyen annem Mzeyyen TERLEMEZ'e, babam Suphi TERLEMEZ'E ve kardeřim Seda TERLEMEZ'e teřekkrlerimi sunarım.

İÇİNDEKİLER

	<u>Sayfa</u>
ÖZET.....	i
ABSTRACT	ii
TEŞEKKÜR	iii
İÇİNDEKİLER	iv
ŞEKİLLER DİZİNİ	vii
ÇİZELGELER DİZİNİ	x
1. GİRİŞ.....	1
2. VERİ MADENCİLİĞİ.....	6
2.1. Veri, Bilgi, Özbilgi ve Bilgelik Kavramları.....	6
2.2. Veri Madenciliğinin Tanımı.....	8
2.3. Veri Madenciliği ve İstatistik.....	11
2.4. Veri Madenciliği Sürecinin Aşamaları.....	13
2.4.1. Amaçların Belirlenmesi	13
2.4.2. Verinin Derlenmesi	14
2.4.2.1. Verinin Seçimi	15
2.4.2.2. Ön İşleme.....	15
2.4.2.3. Veri Dönüşümü.....	22
2.4.3. Veri Madenciliği Modelinin Belirlenmesi, Veri Analizi ve Model Değerlendirmesi.....	26
2.4.4. Karar Sürecine Uyarlama.....	31
2.5. Veri Madenciliği Metotları ve Algoritmalar	33
2.5.1. İstatistiksel Metotlar.....	34
2.5.1.1. Regresyon Analizi	34
2.5.1.2. Varyans Analizi	36
2.5.1.3. Lojistik Regresyon.....	40
2.5.1.4. Log–Lineer Modeller.....	41
2.5.1.5. Doğrusal Ayırma Analizi.....	42
2.5.1.6. Kümeleme Analizi	43
2.5.2. Karar Ağaçları.....	46
2.5.2.1. ID3 Algoritması.....	50

2.5.2.2. C4.5 Algoritması	53
2.5.2.3. C5.0.....	54
2.5.2.4. Sınıflandırma ve Regresyon Ağaçları (CART)	55
2.5.2.5. Ölçeklendirilebilir Karar Ağacı Teknikleri	56
2.5.3. Birliktelik Kuralları.....	57
2.5.3.1. Büyük Öğe Setleri (Large Itemsets)	59
2.5.3.2. Apriori Algoritması	60
2.5.3.3. Örneklem Algoritması	61
2.5.3.4. Bölümlendirme	62
2.5.4. Yapay Sinir Ağları	62
2.5.5. Genetik Algoritma.....	66
2.5.6. Bulanık Kümeler ve Bulanık Mantık	68
2.5.6.1. Üçgen Üyelik Fonksiyonu	71
2.5.6.2. Yamuk Üyelik Fonksiyonu.....	71
2.5.6.3. Gauss Üyelik Fonksiyonu.....	72
2.5.6.4. Genelleştirilmiş Bell Üyelik Fonksiyonu	72
3. PORTFÖY, PÖRTFÖY YÖNETİMİ VE EŞ İŞLEM STRATEJİSİ.....	75
3.1. Portföy Kavramı.....	75
3.2. Portföy Yönetimi.....	77
3.2.1. Pasif Portföy Yönetim Stratejileri.....	80
3.2.2. Aktif Portföy Yönetim Stratejileri	81
3.3. Eş İşlem Stratejisi.....	83
4. İSTANBUL MENKUL KIYMETLER BORSASI'NDA EŞ İŞLEM	
STRATEJİSİ İLE VERİ MADENCİLİĞİ UYGULAMASI.....	85
4.1. Araştırmanın Amacı ve Önemi	85
4.2. Yöntem.....	86
4.2.1. Araştırmanın Modeli	86
4.2.2. Araştırmanın Evreni ve Örneklemi	86
4.2.3. Araştırmanın Değişkenleri ve Veri Derleme Yöntemi.....	87
4.3. Uygulama	88
4.3.1. Bankacılık Sektörü için Hisse Senedi Çiftlerinin Belirlenmesi	89
4.3.2. Tekstil Sektörü için Hisse Senedi Çiftlerinin Belirlenmesi	100

4.3.3. Eş İşlem Stratejisinin Test Verisi İle Sınanması	113
5. SONUÇ VE ÖNERİLER	116
KAYNAKLAR	121
EK – I Tekstil Sektörü İçin Uzun Dönem İlişki Matrisi.....	125
EK - I Tekstil Sektörü İçin Uzun Dönem İlişki Matrisi (Devam).....	126
EK – II Tekstil Sektörü İçin Kısa Dönem İlişki Matrisi	127
EK – II Tekstil Sektörü İçin Kısa Dönem İlişki Matrisi (Devam).....	128
EK – III Eş İşlem Strateji ile elde edilen Tekstil Sektörü Hisse Senedi Çiftleri.....	129

ŞEKİLLER DİZİNİ

2.1. Veri, Bilgi, Anlama, Bilgelik Hiyerarşisi.....	7
2.2. Özbilgi Piramidi	8
2.3. Veri Madenciliğinin Kronolojik Görünümü	10
2.4. Disiplinler arası ilişki	12
2.5. Bir Karar Ağacının yapısı ve veri setinin bölümlenmesi	47
2.6. $\log(1/p)$	51
2.7. $p \log(1/p)$	52
2.8. $H(p, 1-p)$	52
2.9. $\{A, B, C, D\}$ için öge setleri örgüsü (lattice).....	60
2.10. Yapay nöron modeli.....	63
2.11. Genelleştirilmiş Bell üyelik fonksiyonunun parametre değişimleri.....	73
2.12. Üçgen, Yamuk, Gauss ve Genelleştirilmiş Bell Üyelik fonksiyonları.....	74
4.1. Bankacılık sektörü için a) Kısa dönem ve b) Uzun dönem ilişki matrislerinin grafiksel gösterimler	90
4.2. Bankacılık Sektörü için a) Kısa dönem ve b) Uzun dönem ilişki matrislerinin sayısal grafik gösterimler.....	91
4.3. SKBNK/TEBNK hisse senedi çiftinin kapanış fiyat oran grafiği.....	92
4.4. SKBNK/TEBNK hisse senedi çiftinin kapanış fiyatları grafiği	92
4.5. SKBNK/TEBNK hisse senedi çiftinin günlük getiri grafikleri	94
4.6. SKBNK/TEBNK hisse senedi çiftinin birikimli getiri ve kapanış fiyatları grafikleri	95
4.7. SKBNK/YKBNK hisse senedi çiftinin kapanış fiyat oran grafiği.....	96
4.8. SKBNK/YKBNK hisse senedi çiftinin kapanış fiyatları grafiği.....	97
4.9. SKBNK/YKBNK hisse senedi çiftinin günlük getiri grafikleri.....	98
4.10. SKBNK/YKBNK hisse senedi çiftinin birikimli getiri ve kapanış fiyatları grafikleri	99
4.11. Bankacılık Sektörü getirisinin Ulusal 100 Endeks getirisi ile karşılaştırması	100
4.12. Tekstil Sektörü Kısa Dönem İlişki Matrisi a) Grafik ve b) Sayısal grafik gösterimi	101

4.13. Tekstil Sektörü Uzun Dönem İlişki Matrisi a) Grafik ve b) Sayısal grafik gösterimleri.....	102
4.14. AKIPD/ ARAT hisse senedi çiftinin kapanış fiyat oran grafiği	104
4.15. AKIPD/ARAT hisse senedi çiftinin kapanış fiyatları grafiği	104
4.16. AKIPD/ARAT hisse senedi çiftinin günlük getiri grafiği	105
4.17. AKIPD/ARAT hisse senedi çiftinin kümülatif getiri grafikleri.....	106
4.18. ALTIN/YUNSA hisse senedi çiftinin kapanış fiyat oran grafiği.....	107
4.19. ALTIN/YUNSA hisse senedi çiftinin kapanış fiyatları grafiği.....	108
4.20. ALTIN/YUNSA hisse senedi çiftinin günlük getiri grafiği.....	109
4.21. ALTIN/YUNSA hisse senedi çiftinin kümülatif getiri grafikleri	110
4.22. Tekstil Sektörü getirisinin Ulusal 100 Endeks getirisi ile karşılaştırması	111
4.23. Portföy getirisinin Ulusal Tüm Endeks getirisi ile karşılaştırması	112
4.24. Bankacılık Sektörü Gelecek Dönem Getirisi - Ulusal 100 Endeksi Karşılaştırması	113
4.25. Tekstil Sektörü Gelecek Dönem Getirisi - Ulusal 100 Endeksi Karşılaştırması	114
4.26. Portföy Gelecek Dönem Getirisi - Ulusal 100 Endeksi karşılaştırması.....	115
EK – III.1. DERIM/MNDRS hisse senedi çiftinin kapanış fiyat oran grafiği	129
EK – III.2. DERIM/MNDRS hisse senedi çiftinin kapanış fiyatları grafiği.....	129
EK – III.3. DERIM/MNDRS hisse senedi çiftinin günlük getiri grafikleri.....	130
EK – III.4. DERIM/MNDRS hisse senedi çiftinin kümülatif getiri grafikleri....	131
EK – III.5. DERIM/VAKKO hisse senedi çiftinin fiyat oranı grafiği.....	132
EK – III.6. DERIM/VAKKO hisse senedi çiftinin kapanış fiyatları grafiği.....	132
EK – III.7. DERIM/VAKKO hisse senedi çiftinin günlük getiri grafikleri.....	133
EK – III.8. DERIM/VAKKO hisse senedi çiftinin kümülatif getiri grafikleri ...	133
EK – III.9. KORDS/YATAS hisse senedi çiftinin fiyat oranı grafiği	134
EK – III.10. KORDS/YATAS hisse senedi çiftinin kapanış fiyatı grafiği	134
EK – III.11. KORDS/YATAS hisse senedi çiftinin günlük getiri grafikleri	135
EK – III.12. KORDS/YATAS hisse senedi çiftinin kümülatif getiri grafikleri..	136
EK – III.13. MTEKS/YATAS hisse senedi çiftinin fiyat oranı grafiği	137
EK – III.14. MTEKS/YATAS hisse senedi çiftinin kapanış fiyatı grafiği	137

EK – III.15. MTEKS/YATAS hisse senedi çiftinin günlük getiri grafikleri	138
EK – III.16. MTEKS/YATAS hisse senedi çiftinin kümülatif getiri grafikleri..	139

ÇİZELGELER DİZİNİ

2.1	Veri madenciliği gelişim zaman çizelgesi.....	9
2.2.	İstatistik ve Veri Madenciliği arasındaki farklar.....	13
2.3.	Kayıp Veri Yapıları.....	18
2.4.	Sayısal Kestirim için performans ölçüleri.....	31
2.5.	Varyans Analizi Tablosu.....	39
2.6.	Regresyon Modeli için Varyans Analizi Tablosu	40
2.7.	Kümeleme Analizinde Kullanılan Uzaklık Ölçüleri	43
2.8.	Kümeleme Analizinde Kullanılan Hiyerarşik Kümeleme Yöntemleri	44
2.9.	Birliktelik kuralı gösterimleri.....	59
2.10	Yaygın Aktivasyon Fonksiyonları	65
2.11.	Genetik algoritmalarındaki temel kavramlar.....	66
4.1	Bankacılık ve Tekstil Sektörü Veri Matrisleri	88
4.2	Bankacılık Sektörü Kısa Dönem Kapanış Fiyatı İlişki Matrisi.....	89
4.3	Bankacılık Sektörü Uzun Dönem Kapanış Fiyatı İlişki Matrisi	89
4.4.	Bankacılık Sektörü Uzun Dönem – Kısa Dönem Kapanış Fiyatı İlişkilerine Göre Eşleşmeler.....	91
4.5.	SKBNK/TEBNK çiftinin getiri çizelgesi.....	93
4.6.	SKBNK/YKBNK çiftinin getiri çizelgesi	97
4.7.	Uzun Dönem – Kısa Dönem Kapanış Fiyatı İlişkilerine Göre Eşleşmeler	103
4.8.	ALTIN/YUNSA çiftinin getiri çizelgesi.....	108
4.9.	Portföy Nakit Getirisi.....	112
EK – III.1.	DERIM/MNDRS çiftinin getiri çizelgesi	130
EK – III.2.	KORDS/YATAS çiftinin getiri çizelgesi	135
EK – III.3.	MTEKS/YATAS çiftinin getiri çizelgesi	138

1. GİRİŞ

Günümüzde yoğun veri akışı hayatın her alanında kendisini hissettirmektedir. Örneğin günlük hayatımızın bir parçası olan cep telefonlarında bir günde gönderilen ve alınan metin mesajı sayısı dünyanın nüfusundan daha fazladır. Amerika Birleşik Devletlerinde bir günde 3000'den fazla kitap yayınlanmaktadır. Ayrıca New York Times gazetesinde bir hafta içerisinde yer alan tüm bilgi (information), sıradan bir insanın 18. yüzyılda bütün hayatı boyunca karşılaşacağı bilgi miktarından fazla olduğu ileri sürülmektedir. Günümüzde üretilen teknik bilgi miktarı, her iki yılda bir, ikiye katlanmaktadır. Öte yandan bilgisayar dünyasında 2013 senesine kadar, insan beyninin hesaplama kapasitesini aşabilecek bir süper bilgisayarın ortaya çıkacağından söz edilmektedir. Bu örneklerden de kolaylıkla görülebileceği gibi veri ve bilgi kavramları hem çok güncel, hem de üzerinde titizlikle çalışılması gereken ciddi konular olarak ortaya çıkmaktadır.

Veri tek başına herhangi bir anlamı olmayan ama işlendiği zaman yeni açılımlara olanak veren bir olgu olarak karşımıza çıkmaktadır. Akıp giden verinin kontrolü ve düzenlenerek anlamlı hale getirilmesi günümüz teknoloji, eğitim ve sanayi sektörleri için giderek önemli bir hale gelmektedir. Öncelikli olan, verinin anlaşılması ve yığın olarak gelen verinin kontrol altında tutulmasıdır.

Veri akışını kontrol etmek günümüz teknolojisinde çok kolaylaşmış olsa da, veri akışının süreklilik göstermesi ve veri yapılarının etkin bir şekilde analiz edilemeyecek kadar karmaşık olması gibi nedenlerle, yönetilmesi ve anlamlı kullanılabilir gizli kalmış bilgileri ve ilişkileri ortaya çıkartarak, istenilen yönde kullanmak oldukça zorlaşmaktadır. Bu ve benzer nedenle, günümüzde veri akışının küçük bir kısmı kullanılabilir hale gelebilmektedir.

Ne şekilde olursa olsun, tüm organizasyonlar için saniyelerin bile önem taşıdığı günümüz rekabetçi dünyasında bu veri akışından en yüksek düzeyde yararlanmak ve bunun içinde ortaya çıkan zorlukların üstesinden gelmek başarı açısından önem kazanmaktadır.

Sürekli akan veri içinde de farklı örüntüler olabilir. Bazı örüntüler, örneğin, güneşin her gün doğup batması, bir otomobilin yürüyen aksamının gösterdiği arızalardaki örüntüler gibi sayılabilir. Bazı örüntüler ise; hava

olaylarının her zaman aynı şekilde ortaya çıkmaması ve insanların yaygın olan davranışlarındaki örüntüler gibi tekrarlı olamayan fakat aynı örüntünün önceki davranışlarından hareketle tanımlanabilen, örüntülerdir. Bu örüntüleri kelimelerle ve sayılarla sembolize etmek, bu örüntüleri ve davranışları sembolik nesneler olarak tanımlanmasını mümkün kılar [1].

Önceden bilinmeyen, geçerli ve uygulanabilir bilgiyi kaynağından çıkartarak karar destek sürecine uyarlayan veri madenciliği ise, bunu sağlayabilmek için istatistiksel ve matematiksel tekniklerle, örüntü tanımlama gibi teknolojileri bir araya getirerek, büyük miktarlarda statik veya dinamik veri içinde yeni ilişkiler, eğilimler ve örüntüler bulma sürecini içerir.

Günümüzde pazarlama stratejilerinin belirlenmesi, müşteri profilinin çıkarılması, risk analizi ve yönetimi, kredi taleplerinin belirlenmesi, hisse senedi fiyat tahmini ve genel piyasa analizlerinde yaygın uygulama alanları bulan veri madenciliği, pazarlama, bankacılık, sigortacılık, perakendecilik, telekomünikasyon sektörü gibi pek çok alanda bilgi üretimine katkıda bulunmaktadır. Bu uygulama alanları dışında hava tahminleri, toprak kullanımı ve şehir planlaması gibi uygulamaların yanında, sağlık, endüstri, eğitim, güvenlik ve daha birçok sektörde de uygulama alanı da bulabilmektedir [2 – 6].

Veri madenciliği, bankacılık alanında risk analizi, müşteri yapısı, kredi taleplerinin değerlendirilmesi, usulsüzlüklerin tespiti; finansal alanda ise finansal piyasaların analizi, hisse senedi fiyat tahmini, alım – satım stratejilerinin iyileştirilmesi gibi alanlarda giderek önem kazanmaktadır [2]. Veri madenciliği, hisse senedi piyasalarında da kullanılmaktadır. Bu uygulamalarda çeşitli yöntemler denenmiş, yenileri önerilmiş veya performans sınamaları gerçekleştirilmiştir.

1956 yılından günümüze kadar gelişim gösteren ve çevrimiçi veri akışı olan alanlarda etkin olan veri madenciliği finans piyasalarının içyapısını daha iyi anlamak, hisse senedi fiyatlarını temel alarak finansal araçları sınıflandırmak, yakın gelecekteki fiyat hareketlerini öngörmek gibi amaçlarla özellikle hisse senedi piyasalarında giderek artan bir önem kazanmaktadır. Özellikle çevrimiçi veri akışı olan alanlarda daha çok uygulama imkânı bulan veri madenciliği, hisse senedi piyasalarında da çeşitli uygulama örnekleri sunmaktadır.

Bu çalışmanın yapısı ve amaçlarına ilişkin en azından yüzeysel bir karşılaştırmaya olanak sağlamak amacıyla, aşağıda konuya ilişkin bazı çalışmalara ve sonuçlarına yer verilmiştir.

Boginski ve arkadaşları çalışmalarında, hisse senedi piyasalarında, piyasa davranışını yansıtan piyasa grafiğinin istatistiksel analizini gerçekleştirmişler ve güç-kanun modelini takip ettiğini göstermişlerdir. Ayrıca, bu grafikte klikler ve bağımsız setler belirlenmiştir. Bu özel biçimlenmenin pratik bir yoruma sahip olduğu ve analizlerinin de, hisse senedi piyasasının içyapısına daha derin bir bakış açısı sağladığı vurgulanmıştır [7].

Mittermayer'in çalışmasında ise, metin madenciliği ile ilgili bir uygulama yapılmıştır. Çalışmada, üç bileşenden oluşan NewsCATS sistemi tanıtılmıştır. Sistem bileşenleri, basında yayınlanan bildirilerinden metin işleme teknikleri aracılığıyla anlamlı bilgi elde edilmesi, önceden belirlenmiş kategorilere göre basın bildirilerinin sıralanması ve son olarakta uygun işlem stratejilerinin kategorizasyonu olarak sıralanmıştır [8].

Yihua Philip Sheng ve arkadaşları, finansal bir teori olan *piyasa etkinlik teorisi* ile sezgisel değerlendirme algoritmaları bir veri madenciliği tekniğini birleştirmişler, yakın gelecekteki fiyat yükseliş ve düşüşlerine ilişkin sağlam ve güvenilir bir örüntü madenciliğini ortaya çıkarmıştır [9].

Wuthrihch ve arkadaşlarının uygulaması da metin madenciliğini temel almaktadır. Bu çalışmada hedef, web üzerinde yayımlanmış makalelerin içerdiği bilgiyi kullanarak hisse senedi piyasalarını tahmin etmektir. Çoğunlukla lider ve etkili finansal gazetelerde çıkan metinsel makaleler girdi olarak alınmış ve Asya, Avrupa ve Amerika'daki ana hisse senedi piyasa endeksleri için kestirim yapılmıştır. Metinsel ifadelerin sadece etkiyi değil (örneğin, hızlı hisse senedi düşüşleri) ayrıca nedeni de (örneğin, dolardaki güçsüzlük nedeniyle ve bunun sonucunda hazine bonolarının güçsüzleşmesi) içerdiğinden, sayısal zaman serisi verisine ek olarak metinsel bilginin kullanılması ile girdi kalitesinin artacağı ifade edilmiştir. Buna bağlı olarak, iyileştirilmiş kestirim beklentileri gösterilmiştir [10].

Pui Cheong Fung ve arkadaşlarının çalışması da metin madenciliğini temel alan bir çalışmadır. Bu çalışmada kitle iletişim araçlarında (gazete, televizyon,

vb.) yayınlanan gerçek zamanlı haberlerinden hareketle hisse senedi fiyat hareketlerini kestirmek amaçlanmıştır. Ancak, Pui Cheong Fung ve arkadaşları, önceki çalışmaların haber makaleleri ve hisse senetleri arasında bir ilişki olduğunu zaten ileri sürdüğünü, buna rağmen var olan yaklaşımların tümünün sadece tek bir zaman serisinin madenciliği ile ilişkilendirildiğini ifade etmişlerdir. Farklı hisse senetleri arasındaki iç ilişkilerin iyi işaretlenmediğini ve çoklu zaman serilerinin aynı zamanda madenciliğinin sadece daha çok bilgi sağlayıcı değil ayrıca daha çok zorlayıcı olduğunu ve bu yönde araştırmaların eksik kaldığını belirtmişlerdir [11].

Bu çalışmada veri madenciliği ilkeleri göz önüne alınarak finansal piyasa için bir uygulama planlanmıştır. Uygulamanın genel amacı veri madenciliği teknikleri yardımıyla İstanbul Menkul Kıymetler Borsası'nda işlem gören hisse senetlerinden, uygun hisse senedi çiftlerinin seçilmesi ile bir portföyün oluşturulması ve hisse senedi çiftlerinin fiyat pariteleri üzerinden alım – satım işlem zamanlaması için bir otomasyon sağlamaktır.

Amaca uygun olarak, İstanbul Menkul Kıymetler Borsası'nın belirlediği sektörlerden; Bankacılık ve Tekstil Sektörleri için veri madenciliği uygulaması gerçekleştirilmiştir.

Günümüzde küresel para ve ekonomi piyasalarındaki hareketlenmelerin Türkiye ekonomisini olumlu ya da olumsuz etkileme beklentileri, ülkenin kendi iç dinamikleri ile de harmanlanmıştır. Yatırımcılar, bu hareketlilikten yararlanmak amacıyla artık daha sakin kalmaya çalışmakta ve yatırım stratejilerini daha ince bir şekilde detaylandırmaktadır. Bu nedenle sermaye piyasaları için yapılan her tür çalışma büyük değer içermektedir. Özellikle veri madenciliğinin veri kümesi içerisinde açık bir şekilde ya da gizli olarak ortaya çıkabilecek örüntüler olabileceği gerçeği ve eğer varsa bunlardan yararlanarak tutarlı sonuçlara yönelmek bu çalışmanın amaçlarından birini oluşturmaktadır.

Çalışmanın izleyen bölümünde istatistik ve veri madenciliği arasındaki ilişki ana çizgileri ile gözden geçirilmiş, üçüncü bölümde veri madenciliği, aşamaları, veri madenciliğinde kullanılan metotlar ve algoritmalar ele alınmıştır. uygulama bölümünde ise, İMKB'de işlem gören Bankacılık ve Tekstil Sektörlerine ilişkin hisse senetleri göz önüne alınarak, eş işlem (pair trading)

stratejisini kullanarak bir portföy oluşturmak ve portföye dahil olan hisse senetleri üzerinde işlem yapma sinyali veren bir veri madenciliği sistemi oluşturulmuştur.

Oluşturulan sistem, ilk adımda, hisse senedi çiftlerinin oluşturulabilmesi için belirlenen 43 adet hisse senetlerinden oluşan veri seti için öncelikle eksik veri kontrolünü yapmakta ve eğer varsa bulunan eksik veriyi kübik spline tekniği ile doldurmayı gerçekleştirmektedir. İkinci adımda, portföye dahil olacak hisse senedi çiftlerini belirlemek için, iki sektör için uzun dönem ve kısa dönem ilişki matrislerini oluşturarak uzun dönemde yüksek, kısa dönemde ise düşük ilişki gösteren hisse senetleri eşlenerek hisse senedi çiftleri oluşturmaktadır. Üçüncü adımda ise, portföye dahil olan hisse senedi çiftlerinde işlem zamanlamasının belirlenebilmesi için gerekli olan hisse senedi çiftlerinin kapanış fiyatı oran dönüşümünü yapılmış, kapanış fiyat oranları için $\pm \sigma$ ve $\pm 2\sigma$ güven aralıkları hesaplanmıştır. Son adımda ise, hisse senedi çiftleri üzerinde işlem yapmak için belirlenen $\pm 2\sigma$ limitlerinin aşıp aşılmadığını belirlenerek, seçilen hisse senedi çiftlerinin test verisi ile getiri yapısının genel karşılaştırması yapılmıştır.

2. VERİ MADENCİLİĞİ

2.1. Veri, Bilgi, Özbilgi ve Bilgelik Kavramları

Veri madenciliği veri, bilgi (information), özbilgi (knowledge), kavrama (understanding) ve bilgelik (wisdom) adımlarını içeren bir yapıya sahiptir. Bağlı olarak veri madenciliği ile tek başına hiç bir anlamı olmadığı ifade edilen veri ile başlanıp, birçok tekniğin kullanılması ile bilgeliğe ulaşılmasıdır. Bu kavramlar kısaca aşağıda ortaya çıkış sıralarına göre gözden geçirilmiştir.

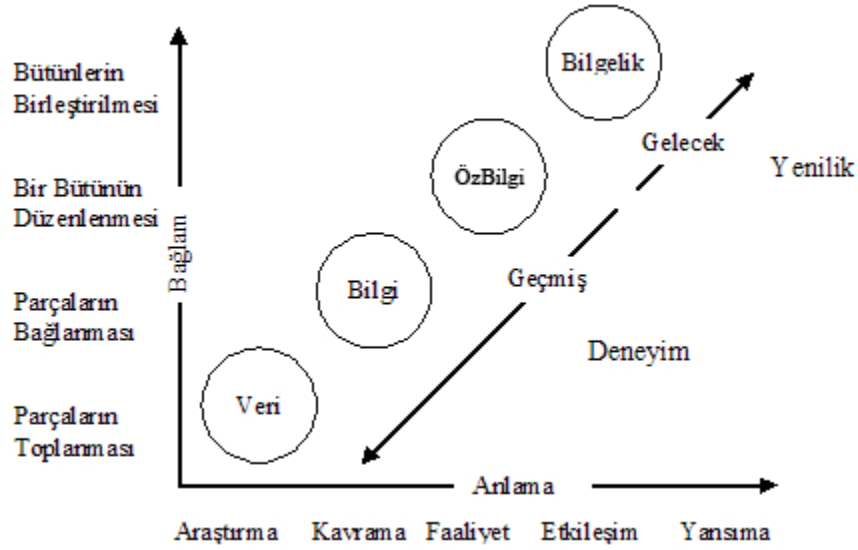
Veri: Olaylar ya da nesneler hakkında ortaya çıkan gerçekler, sembollerdir. Veri ham bir şekilde ortaya çıkar. Sadece var olur ve varlığının dışında (bir şeyin içinde var olması veya kendisi) bir öneme sahip değildir. Herhangi bir formda var olabilir, kullanılabilir veya kullanılamaz olabilir. Tek başına bir anlama sahip değildir. Ama işlenebilir, analiz edilebilir ve “bilgi” ile “özbilgi”ye dönüştürülebilir.

Bilgi (Information): Kullanışlı olması için işlenmiş veri, “kim”, “ne”, “nerede” ve “ne zaman” sorularına cevap sağlar. Bilgi, ilişkisel bağlantı yoluyla anlamlandırılmış veridir. Bu “anlam” kullanışlı olabilir, fakat olmak zorunda değildir.

Özbilgi (Knowledge): Veri ve bilginin uygulanması; “nasıl” sorularını cevaplar. Özbilgi, bilginin uygun bir derlemesidir, öyle ki amacı kullanışlı olmaktır. Bilgi “ezberlendiği” zaman, biriktirilmiş özbilgiye sahip olunur. Bu özbilgi ezberleyen için kullanışlı bir anlama sahiptir, fakat daha fazla anlama sonucunu çıkarmak gibi bir bütünlük (tamamlama) sağlamaz. Örneğin bir ilkokul öğrencisi “çarpım tablosunu” ezberler veya bilgisini biriktirir ve “ $2 \times 2 = 4$ ” olduğunu söyleyebilir çünkü o bilgiyi biriktirmiştir (çarpım tablosuna eklenmiştir). Fakat “ 1267×300 ”ün ne olduğu sorulduğunda, doğru olarak cevap veremez çünkü girdi çarpım tablosunda yoktur. Böyle bir soruyu doğru cevaplamak, bilişsel (cognitive) ve analitik yeteneği gerektirir.

Bilgelik (Wisdom): Bilgelik, sonuca ulaşıcı (extrapolative) ve deterministik (olasılıklı olmayan) bir süreçtir. Bilinçliliğin öncesindeki tüm düzeyler ve özellikle insan programlarının özel tiplerinden (moral, etik kodlar, vb.) yararlanır. Daha önce hiç anlaşılmamış olanlar hakkında aydınlatıcı bir rol

üstlenerek, kavramanın kendisinden daha ötesine ulaşılmasını hedefler. Filozofik (felsefi) incelemenin özüdür. Önceki *düzeylerden* farklı olarak, hiçbir (kolay ulaşılabilen) cevabın olmadığı ve bazı durumlarda, insanoğlu tarafından bilinen (no humanly-known) cevap cümlesinin olamayabileceği soruları sorar. Bilgelik bu nedenle, doğru ile yanlış, iyi ve kötü ayrımı yapılan veya sorgulayan bir süreçtir.



Şekil 2.1. Veri, Bilgi, Anlama, Bilgelik Hiyerarşisi

Veri, bilgi, öz bilgi ve bilgelik zinciri, Şekil 2.1’den de izlenebileceği gibi, verideki ilişkilerin anlaşılabilir olarak bilgiye, bilgideki örüntülerin anlaşılabilir olarak öz bilgiye ve prensiplerin anlaşılabilir olarak bilgelikçe geçişi gösterir. Bu geçişin ilk adımında veri bilgiye dönüşecektir. Bu adımda veri, kesikli, biriktirilebilir, işlenebilir, ölçülebilir, nesnel bir varlığa sahip ve doğal bir yapısı olmayan bir olgudur. Bilgi, miktarı belirlenebilen, nesnel, transfer edilebilen, yenilenebilen, şeffaf, ölçülebilir, bir şekle sahip, işlenebilir, üretilebilir ve tüketilebilir, kullanılabilir bir yapıya dönüşür. Hassas bilgi ise nitel veya nicel bilgi gibi farklı özelliklere sahip farklı tiplerde olabilen bir yapıdır.

İzleyen adımda ise, bilgi öz bilgiye dönüşecektir. Bu adımda öz bilgi, bilgi ve veriden oldukça farklı bir yapıdadır. Çünkü öz bilgi genellikle kişisel ve subjektiftir ve doğal olarak ta yereldir. Öz bilgi, bilen tarafından özümseilir ve var olan algıları ve tecrübeleri ile şekillenir.

Son adımda ise, özbilgi bilgeliğe dönüşür. Bilgelik, sonuca ulaşıcı (extrapolative) ve deterministik bir süreçtir. Bilinçliliğin önceki tüm düzeylerinden ve özellikle insan programlarının özel tiplerinden (moral, etik kodlar, vb.) yararlanır. Daha önce hiç anlaşılmamış olanlar hakkında anlam için işaret gönderir ve böylece de, anlam sınırlarını genişletir. Bilgelik, filozofik (felsefi) incelemenin özüdür. Önceki dört düzeyden farklı olarak, hiçbir (kolay ulaşılabilen) cevabın olmadığı ve bazı durumlarda, insanoğlu tarafından bilinen (no humanly-known) cevap cümlesinin olamayabileceği soruları sorar ve böylece bilgelik, anlamın ötesine geçer.

Şekil 2.2'den de görüleceği gibi anlama piramidinin ilk basamağını veri oluşturmakta, bilgi ve özbilgi basamakları çıkılarak bilgeliğe ulaşılmaktadır [12, 13].



Şekil 2.2. Özbilgi Piramidi

2.2. Veri Madenciliğinin Tanımı

Veri madenciliği daha önceden bilinmeyen, geçerli ve uygulamaya konabilir bilgiyi elde etme hedefine ulaşmak için, bu hedefe ulaşmaya yardımcı olacak veriyi veri kaynağından derleyerek ortaya çıkartır, analiz eder ve karar destek sürecine uyarlar [14]. Bu hedefin ortaya konabilmesi için Çizelge 2.1 deki gelişimi gösteren istatistiksel ve matematiksel tekniklerle, örüntü tanımlama, yapay zekâ, veri tabanı gibi teknolojileri bir araya getirerek, Şekil 2.3 de verilen kronolojik görünümü kazanan veri madenciliği büyük miktarlardaki statik veya

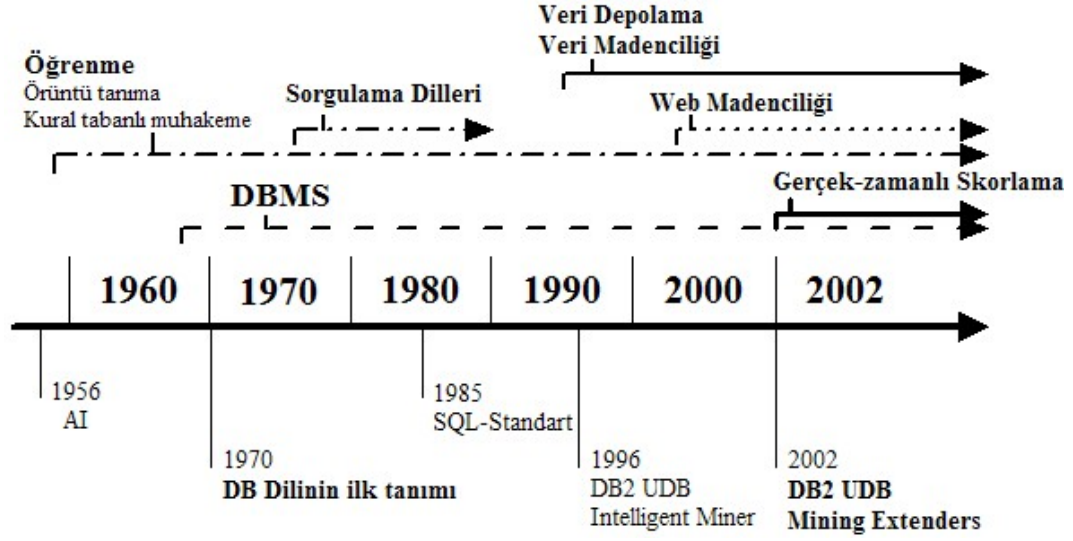
dinamik veri içinde yeni ilişkiler, eğilimler ve örüntüler bulma sürecinde ilerler [6, 15, 16, 17].

Çizelge 2.1 Veri madenciliği gelişim zaman çizelgesi [16]

Zaman	Alan	Katkı
1700'lı yılların sonu	İstatistik	Bayes Olasılık Teoremi
1900'lı yılların başı	İstatistik	Regresyon analizi
1920'li yılların başı	İstatistik	Maksimum benzerlik tahmini
1940'lı yılların başı	Yapay Zekâ	Sinir ağları
1950'li yılların başı		En yakın komşuluk
1950'li yılların başı		Tek bağlantı
1950'li yılların sonu	Yapay Zekâ	Perseptron
1950'li yılların sonu	İstatistik	Yeniden örnekleme, yanlılık indirgeme, jackknife tahminci
1960'lı yılların başı	Yapay Zekâ	ML başladı
1960'lı yılların başı	Veritabanı	Yığın raporlama
1960'lı yılların ortası		Karar ağaçları
1960'lı yılların ortası	İstatistik	Sınıflandırma için doğrusal modeller
	Bilgi erişimi	Benzerlik ölçüleri
	Bilgi erişimi	Kümeleme
	İstatistik	Açıklayıcı veri analizi
1960'lı yılların sonu	Veritabanı	İlişkisel veri modelleri
1970'li yılların başı	Bilgi erişimi	Akıllı bilgi erişimi sistemleri
1970'li yılların ortası	Yapay Zekâ	Genetik algoritmalar
1970'li yılların sonu	İstatistik	Eksik veri ile tahmin (EM algoritmaları)
1970'li yılların sonu	İstatistik	K – ortalama kümeleme
1980'li yılların başı	Yapay Zekâ	Kohonen kendini örgütleyen eşleme
1980'li yılların ortası	Yapay Zekâ	Karar ağacı algoritmaları
1990'lı yılların başı	Veritabanı	Birliktelik kuralı algoritmaları
		Web ve arama motorları
1990'lı yıllar	Veritabanı	Veri ambarlama
1990'lı yıllar	Veritabanı	Çevrimiçi analitik işleme (OLAP)

Veri madenciliği kavramı birçok teorisyen tarafından ele alınmıştır. Bu teorisyenler geldikleri bilim kollarının etkisi ile değişik veri madenciliği tanımları ortaya koymuşlardır. Temel olarak bir tek veri madenciliği kavramı olmasına rağmen farklı bilim dallarının, problemlerin çözümüne farklı öneriler getirmesi burada da kendisini göstermiştir. Farklı veri madenciliği tanımları aşağıda sırayla verilmiştir.

Fayyad'ın tanımına göre; veri madenciliği verideki geçerli, alışılmamış, imkân dâhilinde kullanışlı ve anlaşılabilir örüntüleri sıradan olmayan tanımlama sürecidir [17].



Şekil 2.3 Veri Madenciliğinin Kronolojik Görünümü [18]

Zekulin'in tanımına göre; veri madenciliği önceden bilinmeyen, anlaşılabilir ve harekete geçirilebilir bilginin, geniş veritabanlarından çıkartılması ve kritik ticari kararlar alınması için kullanımı sürecidir [17].

Parsaye'in tanımına göre; veri madenciliği bilginin bilinmeyen ve beklenmeyen örüntüleri için büyük veritabanlarında arama yapılan bir karar destek sürecidir [17].

Mehta'nın tanımına göre; veri madenciliği, Karar Ağaçları, Sinir Ağları, Kural İndirgeme, En Yakın Komşuluklar ve Genetik Algoritmadır [17].

Veri madenciliğinin iki tane birincil hedefi olduğu söylenebilir. Bunlar sırasıyla kestirim ve tanımdır. Kestirimi amaç edinen **Yönlendirilmiş veri madenciliği** (directed), **kestirimci veri madenciliği** (predictive) veya **kontrollü veri madenciliği** (supervised); bağımsız değişkenler ile hedef değişkenler arasındaki ilişkileri tanımlayarak, hedef değişkenlerin bilinmeyen veya gelecekteki değerlerini kestirerek, belli bir soruyu cevaplayacak şekilde düzenlenmesidir. Yönlendirilmiş veri madenciliği sonucunda, elde edilen veri seti ile betimlenen sistemin modeli elde edilmiş olur. [19, 20]. Başka bir şekilde ifade etmek gerekirse, araştırmacı neyi aradığını bildiğinde uygulanan teknik, yukarıdan aşağı doğru ilerleyen bir veri madenciliği türüdür [21]. Tanımlamayı amaç edinen **Yönlendirilmemiş veri madenciliği** (undirected), **tanımlayıcı veri madenciliği** veya **kontROLSÜZ veri madenciliği** (unsupervised) ise, elde edilebilir veri setine dayanarak, şimdiye kadar bilinmeyen, araştırmacı tarafından yorumlanabilecek

bilginin tanımlanabilmesi umuduyla veri üzerinde uygulaması yapılan veri madenciliği türüdür [19, 20]. Diğer bir deyişle, bir bilgi ortaya çıkarsa bu bilginin önemli olup olmadığını araştırmacıya bırakarak, bilgi çıkarsaması yapan aşağıdan yukarı doğru ilerleyen veri madenciliği türüdür [21]. Yönlendirilmemiş veri madenciliği ve yönlendirilmiş veri madenciliği bir birlerinden bağımsız iki ayrı başlık değildir. Temelde veri madenciliği her ikisinin bir bileşimini içermeye gayret eder. Çünkü kestirimci bir model kurarken bile yönlendirilmemiş teknikleri kullanmak yardımcı olabilmektedir [21].

Veri madenciliği hakkında gözden geçirilen genel bilgiler ışığında, veri madenciliği kavramını daha ayrıntılı bir şekilde incelenmeden önce, veri madenciliği ve istatistik arasındaki ilişki ana çizgileriyle ele alınacaktır.

2.3. Veri Madenciliği ve İstatistik

Veri madenciliği ve istatistik arasında önemli farklılıklar olsa da pek çok noktada örtüşmektedirler. Bu farklılıklardan biri kullanılan veri tipinden kaynaklanmaktadır. İstatistik geleneksel olarak, birinci elden veri ile, veri madenciliği ise çoğu kez farklı kaynaklardan toplanmış ikinci elden veri ile çalışır. Diğer bir fark ise, veri madenciliğinde toplanan bilginin saklanması için kullanılan geniş veri tabanlarında yatan önemli bilgi parçalarının önceden tamamen bilinemeyebileceği ve bu nedenle hipotezlerin formüle edilmesinde kullanılamayacağıdır [1, 22]. Bir başka fark ise, istatistiğin ana kütle hakkındaki çıkarsamalarını, bu ana kütlede çekilen temsili örnekten yapmasıdır. Veri madenciliğinde ise, bu durum tartışma konusu olabilmektedir, çünkü üzerinde çalışılan veritabanı örneklendiğinde, küçük fakat aşırı derecede önemli bilginin gözden kaçırılabilmesi ileri sürülmektedir [23].

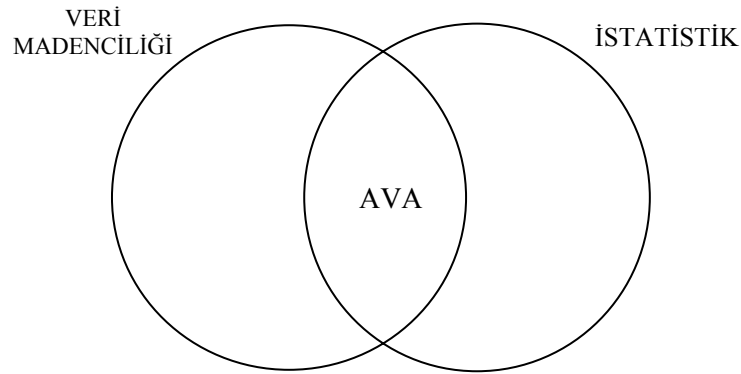
Öte yandan veri madenciliği ve açıklayıcı veri analizi arasında bir örtüşmeden söz edilebilir. Veri madenciliği genellikle büyük miktardaki veriyi, veri içinde gizli kalmış, tutarlı örüntüleri veya sistematik ilişkileri incelemek ve daha sonra tanımlanmış örüntüleri veya ilişkileri test verisine uygulayarak elde edilen sonuçların geçerliliğini sınavan bir süreçtir [24].

Açıklayıcı veri analizi ise veri hakkında ya hiç ya da çok az varsayımına bağlı olarak, genellikle basit hesaplamalarla ve grafikler kullanarak verinin başlangıç düzeyinde incelemesinde kullanılır. Geleneksel istatistik ise veri

hakkında verilen kararların, genellikle kesin ve karmaşık varsayımlara bağlı olarak, onaylanmasında kullanılır [25].

Aralarındaki en büyük fark, veri madenciliği tarafından araştırılan veri tabanlarının grafik, tablo ve diğer insan gücü görevlerle çalışmak gibi geleneksel açıklayıcı veri analizi yaklaşımlarını imkânsız hale getirmesidir. Veri madenciliği ise bilgisayar gücü görevlere ve istatistiksel düzensizlikleri bulan ve potansiyel ilginç olağanüstülükleri bulmayı amaçlayan algoritmalara başvurur. Kısaca özetlemek gerekirse, veri madenciliği çok büyük, ikinci elden, düşük kalite veriye uygulanmış oldukça otomatikleşmiş bir açıklayıcı veri analizi olarakta görülebilir [1].

Bu durumda bu iki disiplin arasında Şekil 2.4’de gösterilen Veri Madenciliği – Açıklayıcı Veri Analizi – İstatistik, gibi disiplinler arası bir ilişki kurulabilir.



Şekil 2.4 Disiplinler arası ilişki

İstatistik ve veri madenciliği arasındaki temel farklar Çizelge 2.2’de gösterilmiştir. Çizelge 2.2 de gösterilen istatistik ve veri madenciliği arasındaki farklar, veri hacmi, veri tipi, veri işleme, yerine getirilen görevler ve araştırma amaçlarından kaynaklanmaktadır.

Çizelge 2.2. İstatistik ve Veri Madenciliği arasındaki farklar

İSTATİSTİK		VERİ MADENCİLİĞİ
*Orta derecede küçük n = onlar veya yüzler p = kullanışlı, nadiren birkaç onluktan fazla	Veri Hacmi	*Geniş n = yüzlerce binlik – milyonluk p = binlerce – milyonlarca
*Veri bir modeli test etmek için veya belirli soruları cevaplamak için toplanır veya bir araya konur (birinci elden veri). °Durum – kontrol çalışmaları °Fikir anket araştırmaları °Gözlemsel çalışmalar	Veri Tipi	*Veri elektronik olarak elde edilir ve gelecekte olası kullanım için depoda tutulur (ikinci elden veri). °Süper Market işlemleri verisi °Telefon aramaları verisi °Hava kontrol verisi °Hisse senedi pazarı verisi °Internet trafiği verisi
*İnsan elinden izlek “ <i>Veri bilgisayar yardımıyla insan tarafından işlenir</i> ”	Veri İşleme	*Oldukça otomatikleşmiş izlek “ <i>Veri insan yardımıyla bilgisayar algoritmaları tarafından işlenir</i> ”
*Model uydurma *Model test etme *Güven ve kestirim aralıkları	Tipik Görevler	*Örüntü arama ve tanımlama *Sınıflandırma *Gruplandırma
*Daha iyi istatistiksel izlek geliştirme *Metotların istatistiksel ve matematiksel özelliklerini çalışma	Araştırma Amaçları	*Veri madenciliği görevlerini yerine getirmek için daha iyi/hızlı algoritmalar geliştirme *Madencilik algoritmalarının deneysel performansını çalışma

2.4. Veri Madenciliği Sürecinin Aşamaları

Daha önce de değinildiği gibi veri madenciliği, önceden bilinmeyen, geçerli ve uygulamaya konabilir bilgiyi elde etmek için, veriyi otomatik veya yarı otomatik anlamda kaynağından *çıkartarak analiz* eden ve karar destek sürecine uyarlanan bir süreçtir. Bu süreci oluşturan aşamalar, veri madenciliğinin belli bir düzende ilerlemesini ve sonuca ulaşmasını sağlar. Bu aşamalar kısaca amaçların belirlenmesi, verinin derlenmesi, veri madenciliği ve son olarak sonuçların analiz edilmesi ve karar sürecine uyarlanması olarak sınıflanabilir [14].

2.4.1. Amaçların Belirlenmesi

Bu aşamada yapılacak analizin hedefleri belirlenir. Analiz edilmek istenen olgunun tanımlanması her zaman kolay olmayabilir. Gerçekte, hedeflenen amaçlar çoğunlukla açıktır, fakat temelindeki problemlerin analiz edilmesine ihtiyaç duyulan ayrıntılı amaçlara dönüştürülmesi zor olabilir. Problemin açık bir ifadesi ve ulaşılabilecek amaçlar, analizin doğru bir şekilde kurulması için ön koşuldur. Amacın belirlenmesi daha sonra uygulanacak yöntemin nasıl organize edileceğini

belirlediğinden, kesinlikle sürecin en zor aşamasıdır. Bu nedenle amaçlar açık olmalıdır ve hiçbir şüpheye veya belirsizliğe yer verilmemelidir [22].

2.4.2. Verinin Derlenmesi

Veri madenciliğinin en önemli ögesini oluşturan veri setidir. Veri setleri veri madenciliğinin başarılı olması için uygulamalara uygun, anlamlı, düzenli ve eksiksiz olmalıdır. Veri kalitesi sistemin imajı üstünde çok derin bir etkiye sahiptir ve tam olarak tanımlanmış ilgili modelin belirlenmesinde önemli bir rol oynar. Ters bir durumun ortaya çıkmasıyla kötü kalitedeki bir veriyi kullanarak mevcut durum hakkında yeni ve geçerli çıkarımlarda bulunmak mümkün olmayacaktır [20, 26]. Veri madenciliğinde kullanılacak verinin kaliteli olması gereklidir. Verinin kalitesini tanımlayan göstergeler aşağıdaki gibi ifade edilebilir:

- Veri kesin olmalı. Analizci, örneğin ismin doğru bir şekilde yazıldığına, kodun verilen bir aralıkta olduğunu, değerin tam olduğunu, vb. kontrol etmelidir.
- Veri, tiplerine göre depolanmalı. Analizci, sayısal değerin karakter formunda gösterilmediğinden, tam sayının gerçek sayı formunda, vb. gösterilmediğinden emin olmalıdır.
- Veri bütünlüğe sahip olmalı. Güncellemeler, kullanıcılar arasındaki uyumsuzluklar nedeniyle kaybolmamalıdır.
- Veri tutarlı olmalı. Farklı kaynaklardan gelen geniş veri setlerinin bütünleşmesinden sonra form ve içerik aynı olmalıdır.
- Veri gereksiz olmamalı. Uygulamada, gereksiz veri minimize edilmeli ve iyice düşünülmüş ve mantıklı çoğaltmalar kontrol altında tutulmalıdır. Çoğaltılmış veri elimine edilmelidir.
- Veri zaman açısından yerinde olmalı. Veri incelendiğinde zaman bileşeninin ne olduğu ve amaca uygunluğu kolaylıkla anlaşılabilir olmalıdır.
- Veri iyi anlaşılmış olmalı. İsimlendirme standartları bir gerekliliktir. Fakat verinin iyi anlaşılabilmesi için tek koşul değildir. Kullanıcının verinin kurulmuş bir bilgi alanıyla uyduğunu bilmelidir.

- Veri seti eksiksiz olmalı. İlgilenilen olay için derlenen veride kayıp gözlemlerin mümkün olduğunca en aza indirgenmelidir.

Bazı veri madenciliği teknikleri her ne kadar eksik veri durumunda analizin doğru sonuçlar verebilmesini sağlayacak bir yapıya sahip olmasına rağmen genellikle eksik verinin model kalitesini azaltabileceği de unutulmamalıdır.

Veri hazırlığı, süreçte en çok kaynak harcanan süreçtir. Tipik bir projenin büyük bir kısmını oluşturur. Bu aşama öncelikle verinin tanımlanmasını ve çıkarılmasını içeren veri seçimi, ikinci olarak verinin örneklendiği ve veri kalitesinin test edildiği veri ön işleme ve son olarak da verinin analitik bir modele çevrildiği veri dönüştürme biçiminde, üç basamaktan oluşur [14].

2.4.2.1. Verinin Seçimi

Veri seçiminin hedefi, ulaşılabilir veri kaynaklarının tanımlanması ve daha ilerideki madencilik hazırlığındaki başlangıç analizi için ihtiyaç duyulan verinin çıkartılmasıdır. Veri seçimi, kullanılacak uygulama tipi ve sırasıyla belirlenen amaçlarla değişiklik gösterecektir. Genelde, iki ayrı durum söz konusudur. Birinci durumda veri oluşturma sürecinin bir uzmanın (modelleyici) kontrolü altında olduğu durumdur: bu yaklaşım *tasarlanmış deney* olarak adlandırılır. İkinci durum ise uzmanın veri oluşturma sürecine etki edemediği durumdur ve *gözlemsel yaklaşım* olarak adlandırılır. Çoğu veri madenciliği uygulamalarında rassal veri oluşturma, gözlemsel ayarlama olarak varsayılır.

Seçilmiş değişkenlerin her biri ile birlikte, her bir değişkenin ne anlama geldiğini anlamak için ilişkilendirilmiş bir üst veriye (metadata) ihtiyaç duyulur. Üst veri verinin sadece sağlam iş tanımlarını değil, ayrıca veri tiplerinin açık tanımlarını, potansiyel değerlerini, orijinal kaynak sistemini, veri tiplerini ve diğer karakteristiklerini de içermelidir [14, 20].

2.4.2.2. Ön İşleme

Veri işlemenin amacı, seçilmiş verinin kalitesini garantilemektir. Temiz ve iyi anlaşılmış veri, diğer nicel analizlerde olduğu gibi, başarılı veri madenciliği için zorunluluktur. Çünkü ister istatistiksel teknikler, ister diğer teknikler olsun, mevcut tüm tekniklerin birçoğu veri setindeki gözlemlerin temiz ve tam olduğu

durumlar için geliştirilmiştir. Kayıp veri, araştırmalarda bilgi eksikliğini temsil ettiğinden bilgi kaybına neden olur. Bunun sonucunda değerlendirilen modele yanlışlık girmiş olur ve hatalı veri madenciliği çıkarsamalarına neden olur [14, 27, 28].

Veri ön işleme teorik olarak, çoğu işlevsel verinin veri madenciliği amaçları için yakalanmamış veya modellenmemiş olmaması nedeniyle problemlili olması beklenen bir aşamadır. Örneğin, bir mağazada bulunan satış noktası (POS) veya elektronik ortamdan yakalanmış veri dışında, seçilmiş veri, tipik olarak, sayısız, tutarsız ve zayıf belgelenmiş işlevsel sistemlerden yakalanmıştır. Örneğin, bir müşteri anketi sonucunda toplanan veride cevaplayıcılar tarafından unutulmuş veya cevaplanmamış hatta araştırma konusu olan ürün yelpazesi için işlevini yitirmiş sorular olabilir.

Veri madenciliği sürecinde en çok karşılaşılan iki önemli sorun, kirli veri ve kayıp değerlerdir [14].

Kirli veri, bir veya daha çok değişkenin, o değişkenlerin alabileceği değer aralığının belirgin bir şekilde dışında olan değerlerdir. Bu kirli değerlerin ortaya çıktığı gözlemler, aykırı değerler olarak adlandırılır. Aykırı değerler, iyi veya kötü anlamda olabilir. İyi anlamda kesinlikle aranılan bilgiyi temsil etmesi, kötü anlamda ise geçersiz veriden başka bir şey olamayabileceğidir.

Farklı aykırı değer türlerine farklı yollarla yaklaşılr. Aykırı değer, insan hatası sonucu oluşabilir. Örneğin, bir kişinin yaşı 650 veya geliri negatif olarak kaydedilmiş olabilir. Eğer, geçerli bir değeri veya makul yedeği bulunabiliyorsa, bu değerler düzeltilmeli, yoksa analizden çıkartılmalıdır. İlave olarak, eğer verinin kaynağı içerde (in-house) ise, veri kontrol kısıtlamaları ve/veya işlevsel süreçler yeniden gözden geçirilmelidir. Eğer bir veri ambarı kullanımdaysa bu kontrol, veri bilgisel çevreye yüklenirken daha basit bir biçimde gerçekleştirilebilir. Aykırı değerın başka bir çeşidi ise, işlevsel sistemlerindeki değişiklikler veri madenciliği ortamına henüz yansıtılmadığı zaman ortaya çıkar. Örneğin, operasyon sistemlerine tanıtılan yeni ürün kodları başlangıçta aykırı değer olarak ortaya çıkarlar. Bu durumda gerekli olan hareket, veri bilgisinin (metadata) güncellenmesidir. Çarpık dağılımlar çok sık aykırı değerleri gösterir. Örneğin, bir görselleştirme, hedef gruptaki insanların çoğunun düşük gelire sahip olduğunu ve

çok azının yüksek kazançlı olduğunu gösterebilir. Bu aykırı değerlerin olması belki iyidir, öyle ki bu homojen gruptaki gerçek yüksek kazançlıları temsil ediyordur veya zayıf veri toplamının sonuçları da olabilirler; örneğin, grup esasen emekli insanlardan oluşuyor olabilir fakat kazara birkaç çalışan profesyoneli içerebilir.

Özet olarak, aykırı değerler onların doğasına bağlıdır. İyi aykırı değerler kötülerden ayırt edilmeli ve uygun şekilde tepki verilmelidir. Aykırı değerleri ele almak için;

- a) Öncelikle aykırı değerleri belirlemek ve nihayetinde ön - işleme sürecinin bir parçası olarak uç değerleri ortadan kaldırmak veya
- b) Uç değerlere duyarlı olmayan güçlü modelleme teknikleri geliştirmek.

olmak üzere iki strateji vardır [14].

Kayıp veri, seçilen veride yer almayan ve belirlenmiş kirli verinin silinmesi durumunda ortaya çıkan değerleri içerir. Değerler, sorulara yanıt vermeme, bulamama gibi gözlem birimi kaynaklı ya da veri toplama, kayıt ve veri giriş hataları gibi gözlem birimi dışındaki hatalar, veri girişi esnasında bilginin ulaşılır olmamasından veya heterojen kaynaklardan seçilmiş olmasından dolayı kayıp olabilir. Kayıp veriyi ele almak için, hiç birinin ideal olmadığı farklı teknikler kullanır [14, 28].

Kayıp veri mekanizmaları dört kategoriye ayrılmıştır: Tamamıyla Rassal Olarak Kayıp (Missing Completely at Random, MCAR), Rassal Olarak Kayıp (Missing at Random, MAR), İhmal Edilemez Kayıp (Nonignorable Missing Data, NI) ve Kayıp Veri Olarak İşlenen Aşırı Değerler (Outliers Treated as Missing Data) [27, 28, 29].

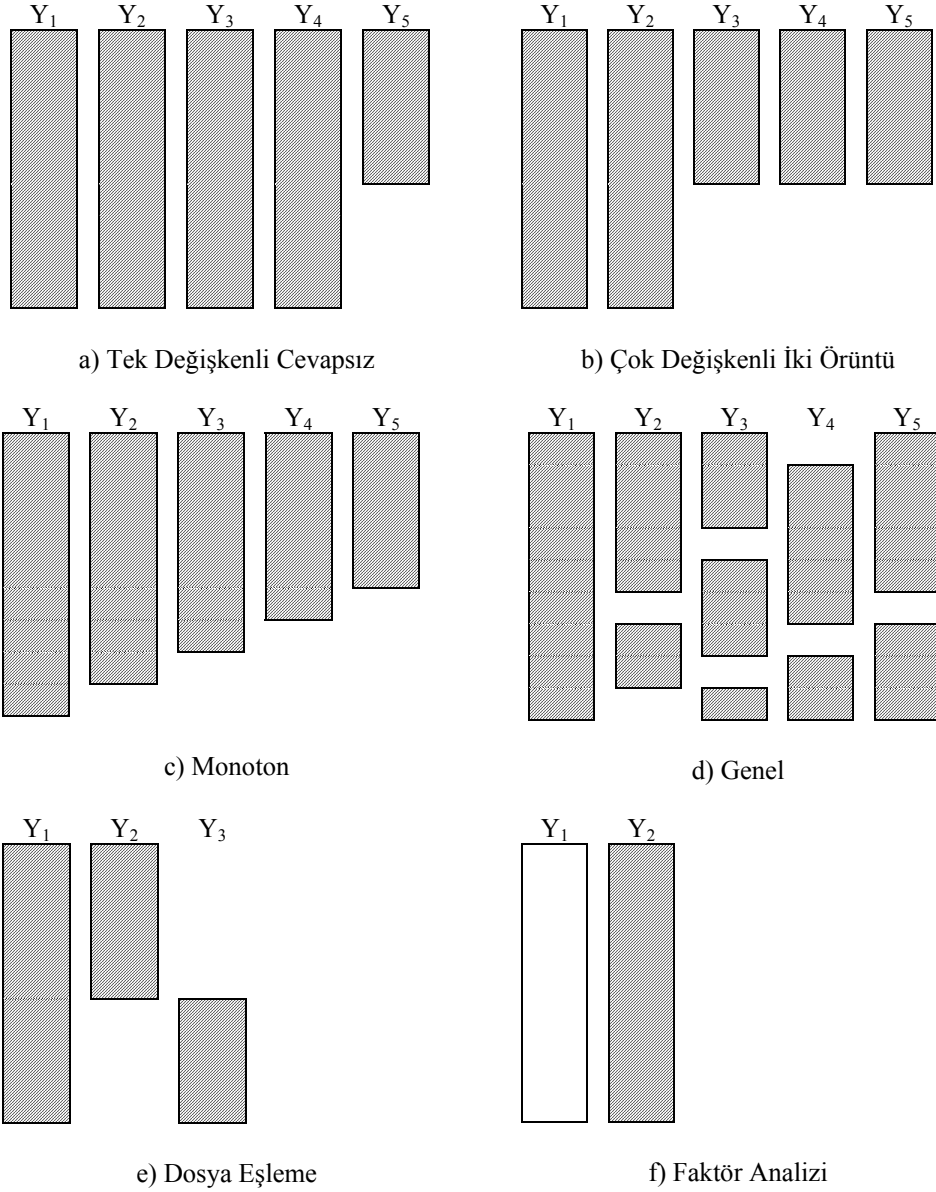
Rassal Olarak Kayıp Veri: Kayıp veri, “X ve Y değişkenleri verilmişken, cevap olasılığı Y değil fakat X’e bağlıysa” MAR olarak tanımlanmıştır [27, 28].

Tamamen Rassal Olarak Kayıp Veri: Kayıp veri, “X ve Y değişkenleri verilmişken, cevap olasılığı X ve Y arasında bağımsızlığın var olduğunu gösterdiği” zaman veri MCAR olarak sınıflandırılır. MCAR verisi MAR verisinden daha yüksek rassallık düzeyi sergiler.

İhmal Edilemez Kayıp Veri: Kayıp veri, “X ve Y değişkenleri verilmiş iken, cevap olasılığı X değişkenine ve muhtemelen Y değişkenine bağımlı” iken veri ihmal edilemez sayılır [27, 28].

Kayıp Veri Olarak İşlenen Aşırı Değerler: Ön test ve hesaplama eşik sınırları kayıp olarak sınıflandırılacak olan değerlerin belirlenmesi amacıyla verinin ön işleminde gereklidir. Değerleri önceden belirlenmiş aralıkların dışına düşen veri, test sonuçlarını çarpıtabilir. Bu aşırı değerlerin kayıp veri olarak değerlendirilmesi gereklidir [27]. Bazı kayıp veri yapıları Çizelge 2.3’de verilmiştir.

Çizelge 2.3. Kayıp Veri Yapıları



Çizelge 2.3 de verilen kayıp veri yapılarının iki tanesi örnek olarak şu şekilde açıklanabilir. Çizelge 2.3a da verilen tek değişkenli cevapsız kayıp veri yapısında, beş değişkenden oluşan veri setinde, kayıp verinin tek bir değişkendeki veri kayıplarıyla sınırlı kalması durumudur. Örneğin, tarımsal bir deneyde ürün verimliliği bağımlı değişkeni ile cins, gübre tipi ve sıcaklık gibi tam olarak gözlemlenmesi istenen bağımsız değişkenler arasındaki ilişki incelenmek istendiğinde, bazı durumlarda tohumun filizlenmemesi veya verinin yanlış kaydedilmesi bağımlı değişkende kayıp verilerin ortaya çıkmasına neden olacaktır. Çizelge 2.3c de verilen monoton kayıp veri yapısında ise, beş değişkenden oluşan veri setinde bir durum için herhangi bir değişkende kayıp olması durumunda izleyen değişkenlerde de kayıp olması durumudur. Özellikle zamana yayılmış olarak yürütülen veri derleme çalışmalarında bazı deney birimlerinin örneklem çerçevesinden araştırmacıya bağlı olmaksızın çıkmaları durumlarıyla ortaya çıkabilir. Örneğin hastaların uzun süreli izlenmelerinde, hastanın iyileşmesi, araştırmacının ulaşamayacağı bir başka merkeze gitmesi ya da hayatının sona ermesi bu tür kayıp verilerin ortaya çıkmasına sebep olabilir.

Kayıp veriyi gidermek için, sırasıyla, atama yapmadan liste veya durum bazında veri silme (list-wise or case-wise data deletion) ve çiftler bazında veri silme (pair-wise data deletion) gibi yöntemlere ve doldurma yapan yöntemlere başvurulur [27, 28]. En çok başvuru alan doldurma yöntemleri, durum ile doldurma (case substitution), ortalama ile doldurma (mean substitution), hot deck doldurma (hot deck imputation), cold deck doldurma (cold deck imputation), regresyon ile doldurma (regression imputation), çoklu doldurma (multiple imputation) ve EM yaklaşımı (Expectation Maximization – EM - approach) yöntemleridir [27, 28, 29].

Sadece Tam Verinin Kullanımı: Bu metot, sadece eksik veri içermeyen birimlerin aldıkları değerlerin kullanılmasını içerir. Kayıp veri MCAR olarak sınıflandırıldığında başarılı bir şekilde kullanılabilen bir metottur. Eğer kayıp veri MCAR olarak sınıflandırılmazsa, yanlılığı ortaya çıkar ve sonuçları tüm topluma genelleştirilemez hale sokar. Bu metottan genellikle “tam durum yaklaşımı” olarak söz edilir [27, 28].

Gözlemlerin ve/veya Değişkenlerin Silinmesi: Bu yöntemde eksik veri içeren gözlem(ler)in ve/veya değişken(ler)in veri setinden tamamıyla silinmesini içerir ve rassal olmayan kayıp veri örüntüsü mevcutken istifade edilebilir [27, 28].

Kayıp verinin giderilmesi amacıyla kullanılan doldurma metotları, araştırmada kullanılan veri setindeki tam veri içeren gözlem ve/veya değişkenlere dayanarak eksik verinin tahmin edilerek doldurulmasına dayanır. Hair ve diğerleri (1998) tarafından doldurmanın tanımı “bir gözlemin kayıp verisinin diğer değişkenlerdeki geçerli değerlerini temel alarak tahmin edilmesi sürecidir” şeklinde yapılmıştır [27].

Durum İle Doldurma: Bu metot çoğunlukla tamamen kayıp verili gözlemleri değiştirmek için kullanılır. Durumlar sadece örneklenmemiş gözlemlerle değiştirilirler. Başka bir şekilde ifade edilirse, örnekleme giren ancak kendisine ulaşamayan ya da çok fazla eksik bilgiye sahip gözlemlerin, örnekte olmayan fakat örnekte yer alma olasılığı olan ve eksik bilgiye sahip gözlemlerin özelliklerine benzerlik gösteren gözlemlerle yer değiştirilmesidir.

Ortalama İle Doldurma: Bu doldurma metodu, kayıp değerlerin kaydedilmiş veya ulaşılabilir değerlerin ortalaması kullanılarak tahmin edilmesini içerir. Kayıp verinin değiştirilmesi için çok kullanılan bir doldurma metodu olmasına rağmen, bu metotta ortalamanın sadece doğruluğu kanıtlanmış ve normal dağılıma sahip olduğu onaylanmış bir toplumdan seçilmiş cevaplardan hesaplanması önemlidir. Eğer verinin çarpıklığı kanıtlanmışsa, değişiklik için ulaşılabilir verinin medyanı da kullanılabilir. Ancak kolay uygulanabilir olması ve tam bilgiye sahip gözlemleri dikkate alması gibi avantajlara sahip olmasına rağmen varyansların olduğundan küçük çıkması, hesaplanan ortalamayı içeren kategoriye gerçekte var olabilecekten daha fazla gözlem düşeceğinden gerçek dağılımının bozulması ve tek bir değerin yinelenmesinden dolayı gözlemlenmiş korelasyonların zayıflaması gibi dezavantajlara da sahiptir.

Hot Deck Doldurma: Hot deck doldurma metotları kayıp değerleri bir sonraki en benzer durumun değerleri ile değiştirilmesini, başka bir ifade ile, kayıp bir değer için benzer cevaplayıcı birimin tahmin edilmiş dağılımından seçilmiş bir değerle yer değiştirilmesini öngörür. Çoğu durumda deneysel dağılım, cevaplayan

birimlerin değerlerinden oluşur. Bu metot uygulamada çok yaygındır ancak kayıp veri literatüründe çok az ilgi çekmektedir.

Cold Deck Doldurma: Cold deck doldurma metodu, kayıp veri için o anki veri kaynağı dışındaki başka kaynaklardan elde edilen değerleri seçer veya ilişkileri kullanır. Bu metotla son kullanıcı kayıp değeri dış kaynaklardan veya bir önceki çalışmadan elde edilen sabit bir değerle değiştirir. Kullanılan değişim değerinin içsel olarak elde edilen değerden daha geçerli olduğu son kullanıcı tarafından tespit edilmelidir. Ortalama ile doldurma tekniğinde geçerli olan dezavantajların birçoğu cold deck doldurma tekniği içinde geçerlidir. Cold deck doldurma metotları tek başına doldurma metodu olarak nadiren kullanılır ve yerine genellikle hot deck doldurma teknikleri için başlangıç değerleri elde etmek için kullanılır [27, 28]. Bu metodun avantajları sırasıyla kavramsal basitlik, sürdürülebilirlik ve değişkenlerin uygun seviyelerinin kullanılmasına olanak sağlanması ve doldurma sürecinin sonunda herhangi bir tam veri gibi analiz edilebilecek tam bir veri setini sağlanabilirliği ve değişkenlere ait dağılımın muhafaza edilmesidir.

Regresyon İle Doldurma: Regresyon ile doldurma, değişkenin veri setindeki diğer değişkenlerle ilişkisine dayanarak kayıp değerlerin kestirilmesinde kullanılır. Kayıp değerlerin doldurulması için tekli ve/veya çoklu regresyon kullanılabilir. İlk basamak bağımsız değişkenlerin ve bağımlı değişkenlerin tanımlanmasından oluşur. Sırasıyla, bağımlı değişken ile bağımsız değişkenler arasında ilişki aranır. Sonuçta oluşan regresyon eşitliği daha sonra kayıp değerlerin kestirimi için kullanılır.

Çoklu Doldurma: Çoklu doldurma metodu, birkaç doldurma metodunu tek bir prosedürde birleştirilmesinden oluşur. Birçok durumda, analize veri sağlamak için beklenti maksimizasyonu, maksimum olabilirlik tahminleri ve hot deck doldurma ile birleştirilir. Metot bir maksimum olabilirlik kovaryans matrisi ve bir ortalama vektörü oluşturarak çalışır. İstatistiksel durağansızlık modele tanıtılır ve tam veritabanının doğal değişkenliği taklit edilir ve hot deck doldurma ile kayıp veri sorunu giderilir [27, 28].

2.4.2.3. Veri Dönüşümü

Veri dönüşümü sırasında ön işlenmiş veri, analitik veri modeli üretmek için dönüştürülür. Analitik veri modeli bir bilgisayar veri modelidir. Birleştirilmiş, bütünleştirilmiş, zaman bağımlı yeniden yapılandırılmış, birçok işlevsel ve dış veriden ön işlenmiş veri gerektiren bir süreçtir. Bu haliyle sonucun doğruluğu ve geçerliliği veri analizcisinin veriyi nasıl organize ettiğine bağımlı olduğundan, çok önemli bir aşamadır. Örneğin, eğer bir satış mağazası müşteri satış örüntülerini analiz etmek isterse, analizci, analizin tüm seviyede mi, mağaza düzeyinde veya bireysel satın alınan eşyalar düzeyinde mi yapılacağına karar vermelidir [14].

Kullanılan teknikler basit veri biçim dönüşümlerinden karmaşık istatistiksel veri indirgeme araçlarına kadar çeşitlilik gösterir. Basit veri dönüşümlerine örnek olarak, tarih değişkenlerinin Amerikan veya Avrupa standardından Julian biçimine dönüştürülmesini veya bir müşterinin yaş değişkeninin işlevsel sistem veritabanındaki doğum gününe bağlı hesaplanması verilebilir. Ek olarak, uygulamada orijinal girilmiş veriden yeni değişkenlerin türetilmesi de oldukça geneldir. Örnek olarak, var olan müşterilerin yeni bir kredi ürünü için uygunluğunu belirlemek için bir veri madenciliği çalışması en az 3, 6 ve 12 aylık dönem için hesap bakiye ortalamasının gerekliliği gösterilebilir.

Diğer bir dönüşüm tipi de veri indirgemedir. Birçok farklı yaklaşım içeren ve genel bir terim olmasına rağmen veri indirgemenin temel amacı, var olan birçok değişkeni, yeni bir değişkenden birleştirerek toplam değişken sayısını indirgemektir. Girdi değişkenlerinin sayısını azaltmak daha ileri analiz için daha küçük ve daha yönetilebilir bir küme oluşturur. Ancak, değişkenlerin birleştirilmesinin bazı bilgi kayıplarına neden olacağı da açıktır. Bağlı olarak sonuçların yorumlanması zorlaşacaktır.

Kesikleştirme (discretization) olarak isimlendirilen başka bir teknik, girdi değişkeninin değerlerini bölümlere ayırarak, nicel değişkenleri kategorik değişkenlere dönüştürmeyi sağlar. Gelir gibi sürekli bir değişken, gelir aralığı gibi bir kategorik değişkene dönüştürülebilir. 0 \$'dan 99,999 \$'a olan gelirler 1 aralığına atanabilir, 10,000 \$ ile 19,999 \$ aralığındakiler 2 aralığına atabilir, ve işleme bu şekilde devam edilir.

One-of-N, tipik bir sinir ağına giriş için sürekli değişkeni kategorik değişkene dönüştürmek ihtiyacı duyulduğunda kullanışlı olan başka bir genel dönüşüm tekniğidir. Örneğin kategorik değişken, araba tipi iken, orijinal değişkenin farklı olası değerlerinin sayısına eşit bir uzunlukta ve mutabık kalınmış bir kodlama sistemiyle nicel bir değişkene dönüştürülebilir. Örneğin, Ford, Lincoln ve Nissan değerleri sırasıyla 100, 010 ve 001 dönüştürülmüş değerleri ile temsil edilebilir [14].

Normalleştirme: Bazı veri madenciliği metotları, tipik olarak n boyutlu bir uzaydaki noktalar arasında uzaklık hesaplamaya dayananlar, en iyi sonuçlar için normalleştirilmiş veriye ihtiyaç duyabilir. Bu ölçülmüş değerler, örneğin $[-1, 1]$ veya $[0, 1]$, belirli bir aralığa ölçeklendirilebilir. Eğer değerler normalleştirilmezse, uzaklık ölçüleri, bir ortalama üzerinden, daha büyük değerlere sahip olan özelliklere gereğinden fazla ağırlık verecektir. Veriyi normalleştirmenin birçok yolu vardır. Üç basit ve etkin normalleştirme tekniği şu şekilde sıralanabilir:

- **Ondalık Ölçeklendirme:** Ondalık ölçeklendirme ondalık noktaları taşır fakat orijinal basamak değerinin çoğunu saklar. Tipik ölçek, değerleri -1 ile 1 aralığında korur. Aşağıdaki eşitlik, $v(i)$ özelliğinin durum i için değeri ve $v'(i)$ en küçük k öyle ki $\max(|v'(i)|) < 1$ için ölçeklendirilmiş bir değer iken ondalık ölçeklendirmeyi tanımlar.

$$v'(i) = v(i) / 10^k$$

İlk önce, veri setindeki maksimum $|v'(i)|$ bulunur ve sonra, ondalık nokta yeni, ölçeklendirilmiş, maksimum mutlak değer 1 den küçük oluncaya kadar hareket ettirilir. Bölen daha sonra bütün $v(i)$ değerlerine uygulanır. Örneğin, eğer setteki en büyük değer 455 ise ve en küçük değer -834 ise, o zaman özelliğin maksimum mutlak değeri 0.834 olur ve bütün $v(i)$ değerleri için bölen 1000 olur ($k = 3$).

- **Minimum - Maksimum Normalleştirme:** Bir v özelliği için verinin 150 ile 250 aralığında olduğunu varsayalım. Bu durumda, bir önceki normalleştirme metodu bütün normalleştirilmiş veriyi 0.15 ile 0.25 aralığında verecektir; fakat değerleri bütün aralığın küçük bir alt

aralığına toplayacaktır. Tüm değerler üzerinden daha iyi değerler dağılımı elde etmek için, normalleştirilmiş aralık, örneğin, [0, 1], bir kümedeki v özelliği için minimum ve maksimum değerlerin otomatik hesaplanmış veya verilen alanda bir uzman tarafından tahmin edilmiş iken minimum - maksimum formülünü kullanabilir:

$$v'(i) = (v(i) - \min(v(i))) / (\max(v(i)) - \min(v(i)))$$

Benzer dönüşümler normalleştirilmiş [-1, 1] aralığı içinde kullanılabilir. Minimum ve maksimum değerlerin hesaplanması tüm veri seti üzerinde ilave bir araştırma gerektirir, fakat, sayısal olarak, prosedür çok basittir. Diğer yandan, minimum ve maksimum değerlerinin uzman tahminleri normalleştirilmiş değerlerin kasıtlı olmayan toplanmalarına neden olabilir.

- **Standart Sapma ile Normalleştirme:** Standart sapma ile normalleştirme çoğu kez uzaklık ölçüleri ile iyi çalışır, fakat orijinal veriden tanınamayacak bir forma dönüştürür. Bir v özelliği için ortalama değeri $\text{ortalama}(v)$ ve standart sapma $\text{sd}(v)$ tüm veri setinden hesaplanır. Sonra, bir i durumu için, özellik değeri

$$v'(i) = (v(i) - \text{ortalama}(v)) / \text{sd}(v)$$

eşitliği kullanılarak dönüştürülür. Örneğin, eğer özelliğin başlangıç değerler seti $v = \{1, 2, 3\}$ ise, o zaman $\text{ortalama}(v) = 2$, $\text{sd}(v) = 1$ ve normalleştirilmiş değerlerin yeni seti $v^* = \{-1, 0, 1\}$ dir.

Normalleştirme, veri madenciliğinin kesin bir adımı olarak ele alınmaz. Bunun nedeni, normalleştirmelerin veri madenciliğinin birçok farklı metodunda kullanışlı olmasıdır. Burada önemli olan normalleştirmenin bir sefer veya bir aşamalık olay olmamasıdır. Eğer bir metot normalleştirilmiş veri gerektirirse, ulaşılabilir veri seçilmiş veri madenciliği tekniği için dönüştürülecek ve hazır hale getirilecektir, fakat veri madenciliğinin diğer tüm aşamalarında, tüm yeni ve gelecek veri ile birlikte özdeş bir normalleştirme kullanılmalıdır. Dolayısıyla, normalleştirme parametreleri bir çözümle birlikte kaydedilmelidir [20].

Veri Düzeltme: Bir sayısal özellik y , birçok farklı değer üzerinden dağılıbilir, bazen eğitim durumlarının sayısı kadar çok olabilir. Birçok veri

madenciliği tekniği için, bu değerler arasındaki küçük farklar önemli değildir; metodun ve son sonuçların performansını azaltabilir. Temel olarak aynı değerler rassal varyasyonları olarak varsayılabilir. Bu nedenle, değişkenin değerlerini düzeltmek bazen avantajlı olabilir. Benzer ölçülmüş değerlerin ortalamasını alan birçok basit düzeltici tayin edilebilir. Örneğin, eğer değerler bir dizi ondalık basamaklı gerçek sayıysa, değerleri verilen duyarlılıkta yuvarlamak, her örneğin kendi gerçek değerine sahip olduğu büyük sayıda örnekler için basit bir düzeltme algoritması olabilir. Eğer verilen F özelliği için değerler kümesi $\{0.93, 1.01, 1.001, 3.02, 2.99, 5.03, 5.01, 4.98\}$ ise, o zaman düzeltilmiş değerlerin $F_{Düzeltilmiş} = \{1.0, 1.0, 1.0, 3.0, 3.0, 5.0, 5.0, 5.0\}$ olacağı açıktır.

Bu basit dönüşüm bir veri setinde herhangi kalite kaybı olmadan uygulanmıştır ve aynı zamanda, özellik için farklı gerçek değerlerin sayısını sadece 3'e düşürmüştür. Bir özellik için farklı değerlerin sayısı azaltmak aynı zamanda veri uzayının çok boyutluluğunu azaltmak demektir. Azaltılmış değerler, özellikle veri madenciliğinin mantık temelli metotları için kullanışlıdır. Düzelticiler bu durumda sürekli özellikleri ikili doğru - yanlış değerli özellikler setine kesikleştirmek için kullanılabilir [20].

Farklar ve Oranlar: Özelliklerdeki küçük değişimler bile veri madenciliği performansında önemli iyileşmeler üretebilir. Girdi ve çıktı özelliklerinin küçük dönüşümlerinin etkisi diğerlerine göre, özellikle veri madenciliğinin amaçlarının tanımlamalarında, önemlidir. Basit dönüşümlerin iki tipi, farklar ve oranlar, bilhassa çıktı özelliklerine uygulandığında, amaç belirlemede iyileştirmeler yapabilir. Bu dönüşümler bazen bir sayının kestiriminin basit ve başlangıç amacından daha iyi sonuçlar üretebilir. Bir uygulamada, örneğin, amaç bir üretim süreci için kontrolleri en uygun ayara getirmektir. Fakat $s(t+1)$ için mutlak büyüklük özelliğini (belirtim) en uygun duruma getirmek yerine, o anki değerinden son en uygun durum $s(t+1) - s(t)$ 'ye getirmek amacına ayarlamak daha etkin olacaktır. Göreceli hareketlerin değerler aralığı mutlak kontrol ayarının değerler aralığından genellikle daha küçüktür. Dolayısıyla, çoğu veri madenciliği metotları için, küçük sayıdaki alternatifler etkinliğini

iyileştirecek ve sıkça daha iyi sonuçlar verecektir. Özelliklerin dönüşümünde ikinci basit dönüşüm oranlardır [20].

2.4.3. Veri Madenciliği Modelinin Belirlenmesi, Veri Analizi ve Model Değerlendirmesi

Uygun veri madenciliği tekniğinin seçilmesi ve uygulanması bu aşamadaki temel görevidir. Bu süreç kolay değildir; genellikle, pratikte, uygulanması birçok modele dayanır ve en iyisinin seçilmesi ek bir görevdir [20].

Kullanılabilecek çeşitli istatistiksel metot ve birçok algoritma vardır, bu nedenle var olan metotların bir sınıflandırmasını yapmak önemlidir. Metotların seçimi çalışılan probleme veya eldeki verinin tipine bağlıdır. Veri madenciliği süreci uygulamalar tarafından yönlendirilir. Bu nedenle kullanılan metot analiz amacına göre sınıflandırılabilir. Bu sınıflandırma üç ana başlık altında yapılabilir:

- **Açıklayıcı metotlar:** Veri grupları daha kısa bir şekilde açıklamayı hedefler; yönlendirilmemiş veya doğrudan olmayan metotlar olarak da adlandırılırlar. Gözlemler önceden bilinmeyen gruplara sınıflandırılabilir (kümeleme analizi); değişkenler önceden bilinmeyen bağlantılara göre kendi aralarında bağlanabilirler (birliktelik metotları, log–doğrusal metotlar). Bu yolla eldeki tüm değişkenler aynı düzeyde ele alınırlar ve herhangi bir nedensellik hipotezi yoktur.
- **Kestirici metotlar:** Bir veya daha fazla değişkenin diğer tüm değişkenlerle olan ilişkisini açıklamayı hedefler. Yönlendirilmiş veya doğrudan metotlar olarak da adlandırılırlar. Bu veriye dayanan sınıflandırma veya kestirim kuralları aranarak yapılır. Bu kurallar açıklayıcı veya girdi değişkenlerine ne olduğu ile ilişkili bir veya daha fazla cevap veya hedef değişkenin gelecek sonuçlarını kestirmekte veya sınıflamakta yardımcı olur. Bu tipin temel metotları sinir ağları ve karar ağaçlarının yanında doğrusal ve lojistik regresyon modelleri gibi klasik istatistiksel modellerdir.
- **Yerel metotlar:** Veritabanının ilgisindeki alt setle ilgili belirli karakteristikleri tanımlamayı hedefler; tanımlayıcı metotlar ve kestirici

metotlar yerel olmaktan çok evrenseldir. Yerel metotlara örnek olarak işlemsel veriyi analiz eden birliktelik kuralları ve aşırı değerlerin tanımlanması verilebilir.

Bu metotlar kendi başlarına veya çok aşamalı bir analizin bir basamağı olarak kullanılabilir. İstatistiksel metot belirlendiğinde, eldeki veri tabanından istenen sonuçların sentezlenmesine yardım eden hesaplamalar için uygun algoritmalar uygulanır [16, 22].

Veri madenciliğinin son aşamasında, model elde edildikten sonra, önemli bir problem hala var olmaktadır. Bu problem modelin *geçerliliğinin* ve *uyumluluğunun* sınanmasıdır.

Model *geçerliliği*, modelin uygulanabilme alanının içinde, kullanıcılar tarafından tanımlanmış amaçlarla tatmin edici kesinlikte uyumlu davrandığını kanıtlamaktır. Diğer bir deyişle, modelin geçerliliğinin sınanmasında, verinin modele dönüştürüldüğünü ve gözlemlenmiş sistemi temsil etmede yeterli kesinliğe sahip olduğunu kanıtlanır. Modelin uyumluluk sınaması ise *doğru*, sisteme uyan, modelin kurulması ile ilgilenir. Model *uyumluluk sınaması* modelin veriden yeterli kesinlikteki yeni temsillere hedeflendiği gibi dönüştürüldüğünü kanıtlar. Model uyumluluğu, veriye doğru şekilde uyan modelin doğru kurulduğunu kanıtlar.

Model geçerliliği veri madenciliği sonuçlarının güvenilirliği ve kabul edilebilirliği için gerekli fakat yetersiz bir koşuldur. Eğer başlangıç amaçları yanlış biçimde tanımlanırsa veya veri seti uygunsuz bir biçimde belirlenirse, model üzerinden ifade edilen veri madenciliği sonuçları kullanışlı olmayacaktır; buna rağmen, model hala geçerli olabilir. Veri madenciliğinin esas amacı eldeki bir problem için sadece bir model üretmek olmamalıdır, fakat yeterli düzeyde güvenilir ve kabul edilebilir, karar-yapıcılar tarafından uygulanmış bir model sağlamak olmalıdır.

Veri madenciliği sonuçlarının geçerliliği ve uyumluluğu uygun bir test süreci ile sınanır. Model testi var olan yanlışlıkları gösterir veya modeldeki hataların varlığını ortaya çıkartır. Düzgün çalıştığını görmek için model üzerinde test verisi ile çalışılabilir veya çeşitli teslere tabi tutulabilir. Bu aşamada “Test

başarısız” ifadesi genellikle testin değil, modelin başarısız olduğu anlamına gelmektedir. Bazı testler modelin davranışsal kesinliğini (ör: geçerlilik) değerlendirmek için tasarlanmıştır ve bazı testler ise verinin modele dönüşümünün kesinliğini değerlendirmeyi (ör: uyumluluk) amaçlamaktadır.

Veri madenciliği süreci yardımıyla elde edilen modelin amacı, yeni örneklerin doğru bir şekilde sınıflandırılması/kestirilmesidir [20].

Sınıflama ve kestirim modelleri için bazı değerlendirme ölçütleri aşağıda sıralanmıştır.

- **Resubstitution Metodu:** Bu en basit metottur. Eldeki tüm veri hem eğitim için hem de test için kullanılır. Diğer bir deyişle, eğitim ve test setleri aynıdır. “Bu dağılım” için hata oranı tahmini iyimser bir şekilde yanlıdır (tahmin edilmiş hata, sıklıkla modelin gerçek uygulamalarından beklenebilecekten daha küçüktür) ve dolayısıyla metot gerçek dünya veri madenciliği uygulamalarında, özellikle örnek hacminin genişlik (dimensionality) oranı küçük olduğu durumlarda çok nadiren kullanılır.
- **Holdout Metodu:** Verinin yarısı veya bazen verinin üçte ikisi eğitim için kullanılırken geriye kalan kısım test için kullanılır. Eğitim ve test setleri birbirinden bağımsızdır ve hata tahmini kötümserdir. Farklı ayırmalar farklı tahminler verecektir. Süreçte rasgele seçilmiş farklı eğitim ve test setleri kullanılması, hata sonuçlarının bir standart parametrede bütünleşmesi ile yinelenmesi, modelin tahminini iyileştirecektir.
- **Leave-one-out Metodu:** Model eğitim için $(n-1)$ örnek kullanılarak tasarlanır ve geriye kalan bir örnek ile değerlendirilir. Bu n kez $(n-1)$ hacimli farklı eğitim setleri ile tekrarlanır. Bu yaklaşım işlem sayısının artmasına yol açar, çünkü n farklı model tasarlanmakta ve karşılaştırılmaktadır.
- **Rotation Metodu (n -fold cross Geçerlilik Sınaması):** Bu yaklaşım hold-out ve leave-out-one metotlarının arasında bir uzlaşmadır. Eldeki örnekleri, $1 \leq P \leq n$ iken, P ayrı alt kümeler bölür. $P-1$ alt

set eğitim için, geriye kalan alt set ise test için kullanılır. Bu uygulamadaki, özellikle örnek sayılarının göreceli olarak küçük olduğu problemler için, en popüler metottur.

- **Bootstrap Metodu:** Bu metot verilen veri seti ile aynı sayıda bir miktar “sahte” veri üretmek için, eldeki veriyi değişikliklerle yeniden örnekler. Bu yeni setlerin sayısı her zamanki gibi yüzlercedir. Bu yeni eğitim setleri hata oranının bootstrap hata tahminlerinin tanımlanmasında kullanılabilir. Deneysel sonuçlar bootstrap tahminlerinin cross-validation tahminlerinden daha üstün olduğunu göstermiştir. Bu metot özellikle küçük örnek setleri durumunda kullanışlıdır. [20]

Şu ana kadar açıklanan değerlendirme ölçütleri sınıflandırma durumlarında kullanılmakta olanlardır. Çeşitli alternatif ölçüler Çizelge 2.4’de özetlenmiştir. Burada $p_1, p_2, \dots, p_n$ test örneklerini ve $a_1, a_2, \dots, a_n$ ise gerçek değerleri göstermektedir. *Hata kareler ortalaması* prensiptir ve en çok kullanılan ölçüdür; bazen kestirilmiş değer kendisi ile aynı boyutu vermesi için karekökü de alınır. Matematiksel olarak işlemenin en kolay ölçümü olma eğiliminde olduğu için doğrusal regresyon gibi birçok matematiksel teknik hata kareler ortalamasını kullanır; matematikçi deyimiyle “iyi huyludur”. Buna rağmen, veri madenciliğinde performans ölçüsü olarak dikkate alınır; tüm performans ölçülerinin hesaplanması kolay olduğundan hata kareler ortalamasının belli bir avantajı yoktur. Dikkat edilmesi gereken konu ele alınan görev için uygun ölçü olup olmadığıdır.

Ortalama Mutlak Hata bir alternatiftir: hataların her birini işaretlerini hesaba katmadan *ortalamasını* alır. Hata kareler ortalaması kestirim hataları, diğerlerinkinden daha büyük olan durumların, aşırı değerlerin etkilerini şişirir. Fakat mutlak hata bu etkiye sahip değildir: tüm hata büyüklüklerini değer olarak büyüklüklerine göre eşit olarak ele alır. Bazen önemin *mutlak* hata değerlerinden çok hata *orandır*. Örneğin; Eğer %10’luk hata, tahminin gerçek değer kabul edilen 500 değerinden 50 birim sapması ya da gerçek 2 değerinden 0,2 birim sapması durumlarının her biri için eşit ölçüde önemli ise mutlak sapmaların ortalamaları

anlamsız hale gelecektir, fakat oransal hatalar uygun olacaktır. Bu etki hata kareler ortalaması veya ortalama mutlak hata hesaplamalarında hata oranları kullanılarak hesaba katılacaktır.

Oransal karesel hata 'nın biraz daha farklı bir yapısı vardır. Hata, basit bir kestirici kullanıldığında elde edilebilecek değere oranlanarak hesaplanmaktadır. Söz konusu basit kestirici eğitim verisinden oluşturulan gerçek değerlerin ortalamasıdır. Dolayısıyla oransal karesel hata temel olarak, toplam karesel hatayı o anda geçerli kabul edilen tahminci yardımıyla hesaplanan toplam karesel hataya bölerek normalleştirmektedir.

Oransal Mutlak hata ise, yukarıda bahsedilen normalleştirme yapısına sahip toplam mutlak hatadır. Bu üç oransal hata ölçüsünde, hatalar ortalama değerleri kestiren basit kestiricinin hatasıyla normalleştirilir.

Korelasyon katsayısı ise, gerçek değer a ile kestirilmiş değer p arasındaki istatistiksel ilişkiyi ölçer. İlişki katsayısı tam ilişkiyi gösteren 1 ile ilişki olmadığını gösteren 0 ve tam ters ilişki olduğunu gösteren -1 değerleri arasında değer alır. Tabi ki, makul kestirim metotları için eksi değerler ortaya çıkmamalıdır. İlişki katsayısı diğer ölçülerden oldukça farklıdır çünkü ölçeği bağımsızdır; belirli bir kestirimciler setindeki tüm kestirimciler sabit bir faktörle çarpıldığında ve gerçek değerler değişmediğinde, hatada değişmeyecektir. İyi performansın büyük ilişki katsayısı değerlerine neden olması nedeniyle farklıdır. Çünkü iyi performansı ölçen diğer metotlar, hatayı küçük değerlerle gösterilirler.

Verilen bir duruma bu ölçülerden hangisinin uygun olduğu da, sadece uygulamanın çalışarak belirlenebileceği bir konudur. Neyin minimize edilmek istendiği, farklı hataların maliyetinin ne olduğu konularında karar vermek çoğu zaman kolay değildir. Hata kareler ölçüleri ve hata karelerinin karekök değeri ölçüleri büyük tutarsızlıkları küçük olanlardan daha ağırlıklı tartar, mutlak hata ölçülerinde böyle bir durum söz konusu değildir. Karekök almak (hata kareler ortalamasının karekök değeri), kestirilmiş miktarla aynı boyutluluğa sahip olması için sadece rakamları küçültür. Hata oranı rakamları çıktı değişkenin kestirilebilirliğini veya kestirilemezliğini kontrol altında tutmaya çalışır: eğer

ortalama değerine oldukça yakın olma eğilimindeyse, kestirimin iyi olması beklenir ve oran rakamı bunu kontrol altında tutar.

Diğer taraftan, bir durumdaki hata değerinin diğer durumdaki değerden çok daha büyük olmasının sebebi, kestiricinin daha kötü olmasından değil, ilk durumdaki niceliğin doğasında daha fazla değişkenlik olması ve dolayısıyla kestirmesinin de daha zor olması olabilir.

Çoğu uygulamalı durumunda en iyi sayısal kestirimi veren metot, hangi hata ölçüsü kullanılırsa kullanılsın, hala en iyi kestirim metodudur [30].

Çizelge 2.4. Sayısal Kestirim için performans ölçüleri

Performans Ölçüsü	Formül
Hata Kareler Ortalaması	$\frac{(p_1 - a_1)^2 + \dots + (p_n - a_n)^2}{n}$
Hata Kareler Ortalamasının Karekök Değeri	$\sqrt{\frac{(p_1 - a_1)^2 + \dots + (p_n - a_n)^2}{n}}$
Ortalama Mutlak Hata	$\frac{ p_1 - a_1 + \dots + p_n - a_n }{n}$
Oransal Karesel Hata (Relative Squared Error)	$\frac{(p_1 - a_1)^2 + \dots + (p_n - a_n)^2}{(a_1 - \bar{a})^2 + \dots + (a_n - \bar{a})^2}, \bar{a} = \frac{1}{n} \sum_i a_i$
Oransal Karesel Hatanın Karekök Değeri	$\sqrt{\frac{(p_1 - a_1)^2 + \dots + (p_n - a_n)^2}{(a_1 - \bar{a})^2 + \dots + (a_n - \bar{a})^2}}$
Oransal Mutlak Hata	$\frac{ p_1 - a_1 + \dots + p_n - a_n }{ a_1 - \bar{a} + \dots + a_n - \bar{a} }$
İlişki Katsayısı	$\frac{S_{PA}}{\sqrt{S_P S_A}}, S_{PA} = \frac{\sum_i (p_i - \bar{p})(a_i - \bar{a})}{n-1},$ $S_P = \frac{\sum_i (p_i - \bar{p})^2}{n-1} \text{ ve } S_A = \frac{\sum_i (a_i - \bar{a})^2}{n-1}$

2.4.4. Karar Sürecine Uyarılama

Veri madenciliği sadece verinin analiz edilmesi değildir, ayrıca sonuçların firmanın karar verme sürecine bütünleşmesidir. Ticari (business) bilgi, kuralların elde edilmesi ve karar sürecine katılması analitik aşamadan karar mekanizmasının oluşturulmasına geçişi sağlar. Model bir kez seçildiğinde ve bir veri seti ile test edildiğinde, sınıflandırma kuralı referans toplumun tümüne uygulanabilir.

Örneğin, hangi müşterilerin daha karlı olacağını önceden ayırabilecek veya farklı hedef müşteri grupları için farklılaştırılmış ticari poliçeleri ayarlanabilir, dolayısıyla da firmanın karlarını arttırabilir.

Veri madenciliğinden elde edilebilecek getirileri gördükten sonra, tüm potansiyelinden yararlanmak için süreci doğru bir şekilde tamamlamak çok önemlidir. Veri madenciliği sürecinin firma organizasyonunun içine dahil edilmesi kademeli olarak yapılmalıdır, süreç boyunca gerçekçi hedefler gösterilmeli ve alınan sonuçlar düzgün bir şekilde incelenmelidir. Veri madenciliğinde son hedef firmanın bütün aktiviteleri ile tam bütünleşmesini sağlanarak firma kararlarının uygun şekilde desteklenmesidir.

Firmanın veri madenciliğini bütünleşme sürecini dört aşamada açıklayabiliriz:

- **Stratejik aşama (strategic phase):** Bu ilk aşamada firma tarafından kullanılan, veri madenciliğinin en iyi sonuçları verebileceği ticari süreçlerin belirlenmesi için çalışılır. Bu aşamanın sonunda elde edilen sonuçlar, pilot veri madenciliği projesi için ticari amaçların ve projenin kendisinin değerlendirme ölçütlerinin tanımlarıdır.
- **Eğitim aşaması (training phase):** Bu aşama veri madenciliği aktivitesinin daha dikkatli bir şekilde değerlendirilmesine izin verir. Pilot proje kurulur ve bir önceki aşamada saptanan ölçüt kullanılarak sonuçlar değerlendirilir. Pilot projenin seçimi temel bir durumdur. Basit ve kullanımı kolay fakat ilgi yaratacak kadarda önemli olmalıdır. Eğer pilot projenin sonucu pozitifse, iki olası sonuç vardır: Birincisi farklı veri madenciliği tekniklerinin faydasının ilk değerlendirilmesidir, ikincisi ise veri madenciliği sistemi ilk örneğinin (prototype) tanımıdır.
- **Oluşum aşaması (Creation phase):** Eğer pilot projenin pozitif değerlendirmesi bütün bir veri madenciliği sisteminin uygulanması ile sonuçlanıyor ise, ticari sürecin veri madenciliği aktivitesini de içerecek şekilde organize edilmesi için detaylı bir plan kurulmasını gerektirir. Daha belirgin şekilde, ticari veritabanının bir veri ambarı ile yeniden

organize edilmesi, veri madenciliği sisteminin ilk işlevsel sürümünü elde edilinceye kadar önceki veri madenciliği ilk örneğini geliştirmek ve projenin takip edilebilmesi için personel ve zamanın tahsis edilmesi gerekecektir.

- **Geçiş aşaması (migration):** Bu aşamada ihtiyaç duyulan tek şey organizasyonun uygun bir şekilde hazırlanmasıdır, böylece veri madenciliği süreci başarılı bir şekilde bütünleştirilebilir. Bunun anlamı yeni sistemin potansiyelinin uygun kullanıcılara öğretilmesi ve getireceği yararları olan inançlarını arttırmaktır. Bu da veri madenciliği sürecinden elde edilen etkin sonuçların sürekli olarak değerlendirilmesi (ve iletilmesi) demektir.

Bir şirket içerisinde veri madenciliğinin geçerli bir süreç olabileceğine karar vermek için iyi iletişim ve ortak çalışma gücüne sahip en az üç farklı kişiliğe ihtiyaç duyulacaktır. Bu kişi ya da kişilerin sahip olması beklenen farklı özellikleri;

- Bir işletme uzmanı; hedeflerin belirlenmesini gerçekleştirebilecek ve veri madenciliğinin sonuçlarını yorumlayabilecek,
- Bilgi teknolojisi uzmanı; veri ve ihtiyaç duyulabilecek teknolojiler hakkında bilgi sahibi olan,
- İstatistik bilgisine sahip uzman; Veri analizi aşaması için istatistiksel metotlarda bir uzman

olarak sıralayabiliriz [22].

2.5. Veri Madenciliği Metotları ve Algoritmalar

Veri madenciliği farklı görevleri yerine getirmek için birçok farklı algoritmayı gerektirir. Bu algoritmaların tümü veriye bir model uydurmaya çalışır. Algoritmalar veriyi inceler ve incelenmiş verinin karakteristiklerine en yakın olan modeli belirler. Bu algoritmalar, veri madenciliği modellerine göre değişiklik gösterecektir: farklı veriden bulunmuş bilinen sonuçların kullanılarak verinin değerleri hakkında kestirim yapan **kestirici model** veya verideki örüntüleri veya ilişkileri tanımlayan **tanımlayıcı model** [16].

Veri madenciliğinde kullanılan metotlar şu şekilde sıralanabilir: regresyon, varyans analizi, lojistik regresyon, log–lineer modeller, ayırma analizi, kümeleme analizi, karar ağaçları ve karar kuralları, birliktelik kuralları, yapay sinir ağları, genetik algoritmalar, bulanık kümeler ve bulanık mantık ve görselleştirme metotları [20].

2.5.1. İstatistiksel Metotlar

İstatistiksel metotlar veri madenciliğinin her aşamasında ortaya çıkmaktadır. Temel olarak veri madenciliği kimi durumlarda bir sınıflama işlemini gerektirirken kimi durumlarda ise tahmin sürecini içerebilir. İşte bu amaçların gerçekleştirilmesinde farklı istatistik metotlarına ihtiyaç duyulmaktadır. Bu bölümde sıklıkla kullanılan bazı istatistik metotları veri madenciliği bakış açısı ile kısaca ele alınacaktır.

2.5.1.1. Regresyon Analizi

Değişkenler arasında bulunan nedensellik ilişkilerinde regresyon analizinin uygulanması ile karşı karşıya kalınmaktadır. Bir ya da birden fazla bağımsız değişkenin bir bağımlı değişkeni açıkladığı düşünülüyor ise bu ilişkinin modellenmesi regresyon analizi yardımıyla yürütülebilir. Bu modeller bir değişkenle ele alınan diğer değişkenlerin eğilimleri arasındaki ilişkiyi tanımlamak için kullanılır. Kimi durumlarda açıklanan yani bağımlı değişken sayısı da birden fazla olabilir. Bu tür durumlarda çok değişkenli regresyon analizine ihtiyaç duyulur. Değişkenler arasındaki ilişkinin doğrusal olduğu varsayımı altında doğrusal regresyon analizi işlemi yürütülür [20].

Regresyon analizi için istatistiksel bir ifade tarzına ihtiyaç duyulduğunda, var olan bir bağımlı Y değişkeni ile bir veya daha fazla bağımsız X değişkeni arasındaki ortalama ilişkinin matematiksel bir fonksiyonla ifade edilmesi ihtiyacı ortaya çıkmaktadır [31]. Bu ilişkiyi karakterize eden kestirim modeli regresyon denklemi olarak adlandırılır [20, 31]. Regresyon modelinin en çok kullanılan şekli genel doğrusal regresyon modelidir. Bu model

$$Y = \alpha + \beta_1 \cdot X_1 + \beta_2 \cdot X_2 + \beta_3 \cdot X_3 + \dots + \beta_n \cdot X_n \quad (3.1)$$

olarak ifade edilir.

Bu model verilmiş tüm örneklere uygulandığında, ε_j verilmiş m örneğin her biri için regresyon hata terimi ve $j = 1, \dots, m$ iken, aşağıdaki gibi bir ifade elde edilir.

$$\hat{y}_j = \hat{\alpha} + \hat{\beta}_1 \cdot x_{1j} + \hat{\beta}_2 \cdot x_{2j} + \hat{\beta}_3 \cdot x_{3j} + \dots + \hat{\beta}_n \cdot x_{nj} + \hat{\varepsilon}_j \quad (3.2)$$

Doğrusal regresyonun en basit şekli bir bağımlı Y , ve bir bağımsız X , değişkenin yer aldığı biçimdir. Basit doğrusal regresyon modeli olarak adlandırılan bu yapıda regresyon denklemi,

$$Y = \alpha + \beta \cdot X \quad (3.3)$$

olarak ifade edilir, burada n örnek veya veri noktası $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$ ikililerinden oluşmaktadır. Ayrıca basit doğrusal regresyon modelinde α ve β , regresyon katsayılarını ifade etmektedir [20], [31]. Bu modelin çözümünde sıklıkla kullanılan teknik en küçük kareler tekniğidir. Bu teknikte Y 'nin varyansı sabittir ve α ve β katsayıları tahmin edilen regresyon doğrusu ile gerçek veri noktaları arasındaki hatayı en küçük yapacak değerler olacak şekilde en küçük kareler tekniği yardımıyla tahmin edilirler.

Regresyon doğrusunun Hata Kareler Toplamı (HKT), y_i veri setinde yer alan gözlemlenmiş değer ve $\hat{y}_i$ modelden elde edilen tahmin değeri iken, aşağıdaki gibi ifade edilir.

$$HKT = \sum_{i=1}^n e_i^2 = \sum_{i=1}^n (y_i - \hat{y}_i)^2 = \sum_{i=1}^n (y_i - \alpha - \beta x_i)^2 \quad (3.4)$$

Hata kareler toplamının α ve β 'ya göre kısmi türevleri sıfıra eşitlenerek, yani toplam hatanın en küçüklenmesi ile *normal denklemler* elde edilir. Aşağıda verilen normal denklemler yardımıyla α ve β katsayıları elde edilir.

$$\begin{aligned}
n\alpha + \beta \sum_{i=1}^n x_i &= \sum_{i=1}^n y_i \\
\alpha \sum_{i=1}^n x_i + \beta \sum_{i=1}^n x_i^2 &= \sum_{i=1}^n x_i y_i
\end{aligned} \tag{3.5}$$

Çoklu doğrusal regresyon modeli ise doğrusal regresyonun bir koludur. Basit doğrusal regresyondan farkı ise, bir bağımlı Y ve birden fazla bağımsız X değişkenin yer almasıdır. Örneğin X_1 , X_2 ve X_3 bağımsız değişkenlerinin ve Y bağımlı değişkeninin yer aldığı çoklu doğrusal regresyon modeli, en küçük kareler metodu kullanılarak tahmin edilmiş α , β_1 , β_2 ve β_3 katsayılar iken,

$$Y = \alpha + \beta_1 \cdot X_1 + \beta_2 \cdot X_2 + \beta_3 \cdot X_3 \tag{3.6}$$

şeklinde ifade edilir.

Çoklu doğrusal regresyon modelinin Regresyon doğrusunun **Hata Kareler Toplamı** (HKT) matris işlemleri ile ifadesi, $\mathbf{Y}$ bağımlı değişken matrisi, $\mathbf{X}$ bağımsız değişkenler matrisi ve $\boldsymbol{\beta} = \{\beta_0, \beta_1, \beta_2, \beta_3\}$, $\beta_0 = \alpha$ iken,

$$\begin{aligned}
\mathbf{Y} &= \boldsymbol{\beta} \cdot \mathbf{X} \\
HKT &= (\mathbf{Y} - \boldsymbol{\beta} \cdot \mathbf{X})' (\mathbf{Y} - \boldsymbol{\beta} \cdot \mathbf{X})
\end{aligned} \tag{3.7}$$

şeklinde ifade edilir. Yine, toplam hatanın en küçüklenmesi sonucunda, $\boldsymbol{\beta}$ regresyon modelinde tahmin edilmiş katsayılar vektörü iken,

$$\boldsymbol{\beta} = (\mathbf{X}' \cdot \mathbf{X})^{-1} (\mathbf{X}' \cdot \mathbf{Y}) \tag{3.8}$$

şeklinde elde edilir [31].

2.5.1.2. Varyans Analizi

Sıklıkla tahmin edilmiş regresyon doğrusunun kalitesinin ve bağımsız değişkenlerin son regresyon modeli üzerindeki etkisinin analiz edilmesi ve k bağımsız veya bağımlı grubun grup ortalamalarının veya işlem ortalamalarının farklılıklarının test edilmesi problemleri **varyans analizi** yaklaşımı ile ele alınır. Bu süreç bağımlı değişkendeki toplam değişimi, gözlemlendikten ve sistematik

bir şekilde ele alındıktan sonra anlamlı bileşenlere böler. Varyans analizi birçok veri madenciliği uygulamasında kullanılan güçlü bir araçtır.

Tek yönlü varyans analizinde tek faktör ve bu faktörün çeşitli düzeyleri ya da denemeleri söz konusu olmaktadır. Burada amaç denemelerin bağımlı değişken üzerindeki etkilerini incelemektir. Varyans analizinin esası deney sonucunda elde edilen verilerin genel ortalamadan olan sapmalarının kareleri toplamının, söz konusu sapmalara neden olan unsurlara göre kısımlara ayırmak ve analiz etmektir [32].

Her gözlem değerinin genel ortalamadan olan sapması iki nedenden ileri gelmektedir. Birincisi gözlem değerinin ait olduğu grubun ortalamasının genel ortalamadan farklı olması, ikincisi ise aynı gruptaki gözlemler arasında bir farklılık söz konusu olduğundan her terimin kendi grup ortalamasından bir sapma göstermesidir [33].

Deney sonucu elde edilen verileri, deney planlamasında göz önüne alınan faktörler ve bunların etkileşimlerinin bir fonksiyonu olarak tanımlamak mümkündür [32, 34].

y_{ij} : i. düzeyin j. tekrarındaki gözlem değeri,

μ_i : i. düzeyin ortalaması,

ε_{ij} : i. gözlemin j. tekrarındaki rassal hata

olmak üzere model denklemi,

$$y_{ij} = \mu_i + \varepsilon_{ij} \begin{cases} i=1,2,\dots,a \\ j=1,2,\dots,n \end{cases} \quad (3.9)$$

şeklinde ifade edilmektedir. Burada a düzey sayısını ve n gözlem sayısını göstermektedir. Hatalar sıfır ortalamaya sahiptir ve $E[y_{ij}] = \mu_i$ ' dir. Eşitlik (3.9) ortalamalar modeli olarak adlandırılmaktadır.

$$\mu_i = \mu + \tau_i, \quad i=1,2,\dots,a \quad (3.10)$$

olmak üzere ortalamalar modelini

$$y_{ij} = \mu + \tau_i + \varepsilon_{ij} \begin{cases} i = 1, 2, \dots, a \\ j = 1, 2, \dots, n \end{cases} \quad (3.11)$$

şeklinde tanımlamak da mümkün olmaktadır.

Modelin bu formunda, μ genel ortalama olarak adlandırılan tüm işlemlerde ortak parametredir. τ_i i . düzey etkisi olarak adlandırılan i . işleme ait bir parametredir. Eşitlik (3.10) etkiler modeli olarak adlandırılmaktadır.

Ortalamalar modeli tek yönlü veya tek faktörlü varyans analizi modeli olarak adlandırılmaktadır. Bu modelde sadece tek bir faktörün etkisi incelenmektedir. İşlemlerin (sıklıkla deneysel birimler olarak adlandırılır) uygulandığı çevrenin mümkün olduğunca tekdüze olması için deneylerin rassal sırada gerçekleştirilmesi gerekmektedir. Böylece, uygulanan deneysel tasarım tamamen rassallaştırılmış tasarım olacaktır. Burada amaç, işlem ortalamaları ile ilgili uygun hipotezleri test ve tahmin etmektir. Hipotezleri test etmek için gerekli temel varsayım, model hatalarının normal dağılımlı, sıfır ortalamalı ve σ^2 varyanslı bağımsız dağılmış rassal değişkenler olduğu varsayılmaktadır.

Modelin hatalarının normal dağılımlı, σ^2 varyanslı bağımsız dağılmış rassal değişkenler olduğu varsayımını sınamak üzere model yeterlilik kontrolüne başvurulur. Normal dağılım varsayımında eğer hataların üzerinde $\varepsilon \sim N(0, \sigma^2)$ olması varsayımı geçerliyse, örneklem, sıfır ortalamalı normal bir dağılımdan elde edilmiş bir örneklem olacaktır [34].

İlgilenilen faktör düzeyleri arasında istatistiksel olarak anlamlı bir farklılığın olup olmadığı testi F dağılımı yardımıyla yürütülür. Faktör düzeyleri için meydana gelen değişkenliklerin ölçülmesi amacı ile grup içi değişkenlikler ve gruplar arası değişkenlikler hesaplanır. Daha sonra bu değerler Çizelge 2.5’de gösterildiği gibi bir tabloda bir araya getirilir. Bu tablo varyans analizi tablosu olarak adlandırılır. Tabloda hesaplanan F istatistiği daha sonra teorik F istatistiği değeri ile karşılaştırılarak karar verilir.

Tek faktörlü varyans analizi modeli için oluşturulan varyans analizi tablosuna Çizelge 2.5’de yer verilmiştir:

Çizelge 2.5. Varyans Analizi Tablosu

Değişim Kaynağı	Serbestlik Derecesi	Kareler Toplamı	Kareler Ortalaması	F
Gruplar Arası	$k - 1$	$GAKT = n \sum_{i=1}^k (\bar{y}_{i.} - \bar{y}_{..})^2$	$GAKO = \frac{GAKT}{k - 1}$	$F = \frac{GAKO}{GIKO}$
Gruplar İçi	$n - k$	$GIKT = GKT - GAKT$	$GIKO = \frac{GIKT}{n - k}$	
Toplam	$n - 1$	$GnKT = \sum_{i=1}^a \sum_{j=1}^n (y_{ij} - \bar{y}_{..})^2$		

Çizelge 2.5’de $\bar{y}_{i.}$: i.düzeyin altındaki gözlemlerin ortalamasını, $\bar{y}_{..}$: tüm gözlemlerin genel ortalamasını, y_{ij} : i.düzeyin j.tekrarındaki gözlem değerini, N: toplam birim sayısını, a : düzey sayısını, n: her gruptaki gözlem sayısını ifade etmektedir.

Bazı araştırmalarda yalnızca bir faktörle ilgilenilmeyebilir. Olayın kendi doğası elde edilen gözlem birimleri üzerinde ikinci bir faktörün etkisini de taşıyor olabilir. Bu tür durumlar için iki yönlü varyans analizi çözümlemesi yapılması söz konusu olmaktadır. Böylece iki faktörün çeşitli düzeylerinin bir bağımlı değişken üzerindeki etkilerinin incelendiği modele, iki yönlü varyans çözümlemesi modeli denilmektedir.

Bu model için oluşturulacak olan varyans analizi tablosu verinin genel değişkenliğinin temel sebepleri olarak görülen faktör düzeyi, faktör alt düzeyleri ve ikinci faktör etkilerini içerecek şekilde düzenlenmektedir [35].

Bunun dışında, varyans analizi, doğrusal regresyon modelinde yer alan, başka bir ifadeyle regresyon denkleminde değişkenler arasındaki ilişkiyi gösteren, β katsayılarının sıfırdan farklılıklarını incelemek için kullanılan birincil metottur [20, 36].

Regresyon doğrusunun sınanmasında kullanılan varyans analizi için hesaplanacak F istatistiği

$$F = \frac{(\hat{\beta}'X'Y - n\bar{Y}^2)/k}{(Y'Y - \hat{\beta}'X'Y)/(n - k - 1)} \quad (3.12)$$

şeklinde ifade edilir.

Doğrusal regresyon modeli için varyans analizi tablosu Çizelge 2.6’de ki gibidir [36].

Çizelge 2.6. Regresyon Modeli için Varyans Analizi Tablosu

Değişim Kaynağı	Serbestlik Derecesi	Kareler Toplamı	Kareler Ortalaması	F
Toplam	$n - 1$	$\mathbf{Y}'\mathbf{Y} - n\bar{\mathbf{Y}}^2$		
Regresyon	k	$RKT = \hat{\boldsymbol{\beta}}'\mathbf{X}' - n\bar{\mathbf{Y}}^2$	RKT / k	$F = \frac{RKO}{HKO}$
Hata	$n - k - 1$	$HKT = \mathbf{Y}'\mathbf{Y} - \hat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{Y}$	$HKT / (n - k - 1)$	

2.5.1.3. Lojistik Regresyon

Doğrusal regresyon sürekli değişkenleri modellemekte kullanılır. Genelleştirilmiş regresyon modelleri kategorik cevap değişkenlerini modellemek için uygulanabilen doğrusal regresyon yaklaşımının teorik altyapısını gösterir. Genelleştirilmiş doğrusal modelin yaygın bir türü lojistik regresyondur. Lojistik regresyon kestirici değişkenler kümesinin doğrusal bir fonksiyonu olarak ortaya çıkan bazı olayların olasılıklarını modeller. Böylece lojistik regresyon, bağımlı değişkenin tahmin değerlerinin olasılık olarak hesaplayarak, olasılık kurallarına uygun şekilde sınıflandırma yapılmasına olanak sağlar [20, 37].

Lojistik regresyon metodu, bağımlı değişkenin değerini kestirmek yerine, bağımlı değişkenin verilmiş bir değeri alma olasılığını tahmin etmeye çalışır. Örneğin, bir müşterinin iyi veya kötü kredi değerlendirmesine sahip olup olmadığını tahmin etmek yerine, lojistik regresyon yaklaşımı iyi kredi derecesi olasılığını tahmin etmeye çalışır. Burada bağımlı değişken gerçek durumu, tahmin edilen olasılık değerine bakılarak belirlenir. Verilen örnek için açıklamak gerekirse, eğer tahmin edilen olasılık değeri 0,5 değerinden büyükse kestirim müşterinin kredi derecesi *iyi bir kredi derecesine* yakındır, diğer durumda ise *kötü bir kredi derecesine* yakındır. Bu nedenle, lojistik regresyonda olasılık değeri *başarı olasılığı* olarak adlandırılır [20].

Lojistik regresyon, sadece modelin bağımlı kukla değişkeni ikili kategorik olarak tanımlandığında kullanılır. Bağımlı değişken Y , sadece 0 ve 1 değerlerini alırken, mevcut veriye bağlı olarak verilen örnek için her iki değer olasılıkları

$P(Y_j = 0) = 1 - p_j$ ve $P(Y_j = 1) = p_j$ şeklinde hesaplanabilir. Bu olasılıkların uydurulacağı modeli sağlayan, *doğrusal lojistik model* olarak bilinen, regresyon denklemi,

$$\log(p_j / 1 - p_j) = \alpha + \beta_1 \cdot X_{1j} + \beta_2 \cdot X_{2j} + \beta_3 \cdot X_{3j} + \dots + \beta_n \cdot X_{nj} \quad (3.13)$$

şeklinde ifade edilir [20, 31, 38].

Lojistik regresyon basit fakat veri madenciliği uygulamaları için güçlü bir sınıflandırma aracıdır. Tek bir eğitim seti ile lojistik regresyon modeli kurularak, tek bir test seti ile modelin kategorik değerleri kestirim kalitesi analiz edilebilir.

Bunun dışında, regresyon analizinin uygulanabilmesi için gerekli bazı varsayımların, doğrusal ayırma analizinin de, veri setinin normal dağılıma sahip olması varsayımı gibi bazı varsayımların sağlanması zorunluluğunu taşınamaması nedeniyle lojistik regresyon analizi, bu metotların alternatifi olabilmektedir [20, 37].

2.5.1.4. Log-Linear Modeller

Log-linear modeller kategorik değişkenler arasındaki ilişkiyi analiz etmenin bir yoludur. Log-linear model kesikli, çok boyutlu olasılık dağılımları tahmin eder. Çıktı Y_i 'in, beklenen değeri μ_j olan, Poisson dağılımına sahip olduğu varsayılan bir genelleştirilmiş doğrusal model türüdür. μ_j 'in doğal logaritmasının girdilerin doğrusal fonksiyonu olduğu varsayılır.

$$\log(\mu_j) = \alpha + \beta_1 \cdot X_{1j} + \beta_2 \cdot X_{2j} + \beta_3 \cdot X_{3j} + \dots + \beta_n \cdot X_{nj} \quad (3.14)$$

İlgilenilen tüm değişkenler kategorik değişkenler olduğundan, verinin evrensel dağılımını gösteren bir frekans tablosu kullanılır. Log-linear modellemede amaç kategorik değişkenler arasındaki birliktelikleri tanımlamaktır. Birliktelik modeldeki etkileşim terimleriyle ilgilidir; dolayısıyla çözümlenmesi gereken problem modeldeki β katsayılarından hangilerinin sıfır olduğunun bulunmasıdır. Eğer log-linear modeldeki değişkenler arasındaki bir etkileşim varsa, etkileşimdeki değişkenlerin bağımsız olmadıkları fakat ilişki oldukları ve ilgili β katsayılarının sıfırdan farklı oldukları anlamına gelmektedir [20].

2.5.1.5. Doğrusal Ayırma Analizi

Doğrusal ayırma analizi, bağımlı değişkenin kategorik (nominal veya ordinal) ve bağımsız değişkenlerin metrik olduğu, k tane özelliği bilinen bireylerin, bilinen özelliklerine göre iki veya daha çok gerçek gruba veya sınıfa ayrılması problemleri ile ilgilenir [20, 37, 38, 39].

Doğrusal ayırma analizinin amacı, bilinen k özellikleri bakımından farklı grupların grup profillerini belirlemeye yarayan bir ayırma fonksiyon bulmaktır. Bu fonksiyonlar, yeni gözlemlenen k özelliğe sahip birimlerin hata payı minimum olacak şekilde sınıflandırmasında kullanılır [37, 39].

Doğrusal ayırma fonksiyonu aşağıdaki gibi bir forma sahiptir:

$$z_i = w_1x_{i1} + w_2x_{i2} + w_3x_{i3} + \dots + w_kx_{ik} \quad (3.15)$$

Burada, $x_1, x_2, x_3, \dots, x_k$ bağımsız değişkenleri, z niceliği ayırma skorlarını ve $w_1, w_2, w_3, \dots, w_k$ ise ağırlıkları ifade etmektedir. Bir veri örneği için ayırma skorunun geometrik yorumu ağırlık parametreleri kümesi ile tanımlanan doğruya izdüşümleri temsil etmesi olarak yapılır.

Ayırma fonksiyonu bulunurken, gruplar arası varyansın grup içi varyansa göre en büyüklenmesi gerekir. Bu amaçla, $\mathbf{a}$ $p \times 1$ boyutlu katsayılar matrisini, $\mathbf{B}$ $p \times p$ boyutlu gruplar arası varyans analizi ve $\mathbf{W}$ $p \times p$ boyutlu grup içi varyans matrisi iken

$$f(a_1, a_2, \dots, a_p) = \frac{\mathbf{a}'\mathbf{BA}}{\mathbf{a}'\mathbf{WA}} \quad (3.16)$$

şeklinde elde edilir [20, 39].

Model kurulduktan sonra ayırma fonksiyonu z sınıflandırılmamış yeni örneğin sınıfının kestirilmesinde kullanılır. Sınır noktaları her bireyin ayırma skoru ile karşılaştırılacak bir ölçüt olarak kullanılır. Optimal sınır noktası, gruplardaki birim sayısı eşit olduğu biliniyorken

$$z_c = \frac{z_1 + z_2}{2} \quad (3.17)$$

şeklinde hesaplanır. Yeni bir örneğin gruplara sınıflandırılması, ilgili bireyin ayırma değerinin $z > z_c$ veya $z < z_c$ olmasına bağlıdır [20, 39].

2.5.1.6. Kümeleme Analizi

Kümeleme analizinin genel amacı, bir ilişki ölçüsü kullanarak bir gruptaki örnekleri, grup içinde benzer ve gruplar arasında benzemez şekilde otomatik olarak gruplayarak, özetleyici bilgi sağlamaktır. Kümeleme analizinin girdisi bir örnekler kümesi ve iki örnek arasındaki bir benzerlik (veya benzemezlik) ölçüsüdür. Çıktı ise, bir veri setinin ayrılması ile oluşan kümeler veya ayrılma yapısıdır. Kümeleme analizinde veri setinin ayrılma yapısı önceden bilinmemektedir ve mevcut duruma ilişkin sonuçlar vermesi nedeniyle gelecekte kullanılması mümkün olmamaktadır. Kümeleme analizinin diğer bir sonucu da, özellikle veri setinin karakteristiklerinin daha derin olarak araştırılmasında önemli olan, her kümenin genelleştirilmiş bir tanımıdır. Kümeleme analizinde, verilerin normal dağılımlı olması gerekliliği varsayımı ile birlikte prensipte kalmakta, uzaklık değerlerinin normalliği yeterli görülmektedir. Ayrıca kümeleme analizinde kovaryans matrisine ilişkin herhangi bir varsayım da bulunmamaktadır. Bazı uzaklık ölçüleri Çizelge 2.7’de verilmiştir [20, 37, 39].

Çizelge 2.7. Kümeleme Analizinde Kullanılan Uzaklık Ölçüleri

Minkowski Uzaklığı	$d_{\lambda}(\mathbf{x}_i, \mathbf{x}_j) = \left[\sum_{k=1}^p x_{ik} - x_{jk} ^{\lambda} \right]^{1/\lambda}, \lambda \geq 1$ için
Manhattan City-Block Uzaklığı ($\lambda=1$ durumu)	$d_1(\mathbf{x}_i, \mathbf{x}_j) = \sum_{k=1}^p x_{ik} - x_{jk} $
Öklit (Euclidean) Uzaklığı ($\lambda=2$ durumu)	$d_2(\mathbf{x}_i, \mathbf{x}_j) = \left[\sum_{k=1}^p x_{ik} - x_{jk} ^2 \right]^{1/2}$
Ölçekli Öklit Uzaklığı	$d_2(\mathbf{x}_i, \mathbf{x}_j) = \left[\sum_{k=1}^p w_k^2 (x_{ik} - x_{jk})^2 \right]^{1/2}$
Mahalanobis Uzaklığı	$d_{ij}(\mathbf{x}_i, \mathbf{x}_j) = \mathbf{D}^2 = (\mathbf{x}_i - \mathbf{x}_j)' \mathbf{S}^{-1} (\mathbf{x}_i - \mathbf{x}_j)$
	$d_{ij}(\bar{\mathbf{x}}_i, \bar{\mathbf{x}}_j) = \mathbf{D}^2 = (\bar{\mathbf{x}}_i - \bar{\mathbf{x}}_j)' \mathbf{S}^{-1} (\bar{\mathbf{x}}_i - \bar{\mathbf{x}}_j)$
Hotelling T^2 Uzaklığı	$T^2 = \frac{n_1 n_2}{n} (\bar{\mathbf{x}}_i - \bar{\mathbf{x}}_j)' \mathbf{S}^{-1} (\bar{\mathbf{x}}_i - \bar{\mathbf{x}}_j)$
Canberra Uzaklığı	$d(\mathbf{x}_i, \mathbf{x}_j) = \sum_{k=1}^p x_{ik} - x_{jk} / \sum_{k=1}^p (x_{ik} + x_{jk})$

Kümeleme analizinde pek çok yöntem bulunmakta ve bu yöntemler farklı başlıklar altında toplanmaktadır. Ancak, en çok bilinen kümeleme yöntemleri; hiyerarşik ve hiyerarşik olmayan yöntemler biçiminde iki ana başlık altında toplanmaktadır.

Hiyerarşik (hierarchical) ya da aşama sıralı kümeleme yöntemlerinde işleyişin kolay anlaşılabilmesi için ağaç diyagram (dendogram) örneğinden yararlanılmaktadır. Kümeleme sürecinin başlangıcında her birey bir kümedir, süreç sonunda ise tüm bireyler bir kümede toplanır. Bu işleyiş daha ayrıntılı bir biçimde aşağıdaki dört adımlı bir algoritma ile ifade edilebilir:

1. n tane birey, n tane küme olmak üzere işleme başlanır.
2. En yakın iki küme birleştirilir.
3. Küme sayısı bir indirgenerek yinelenmiş uzaklıklar matrisi bulunur.
4. 2 ve 3. adımlar n-1 kez tekrarlanır.

Bu algoritmaya dayalı sekiz farklı hiyerarşik teknikten söz edilmektedir. Bu teknikler, sırasıyla en yakın komşuluk, en uzak komşuluk, ortalama bağlantı, McQuitty bağlantı, Ward (minimum varyans), küresel ortalama bağlantı, ortanca ve Lance Willims teknikleridir. Hiyerarşik kümeleme yöntemleri Çizelge 2.8’de verilmiştir.

Çizelge 2.8. Kümeleme Analizinde Kullanılan Hiyerarşik Kümeleme Yöntemleri

En Yakın Komşuluk (Nearest Neighbor Single Linkage)	$d_{k(i,j)} = \min(d_{ki}, d_{kj})$
En Uzak Komşuluk (Farthest Neighbour Complete Linkage)	$d_{k(i,j)} = \max(d_{ki}, d_{kj})$
Ortalama Bağlantı (Average Linkage)	$d(A, B) = \frac{\sum_i \sum_k d_{ik}}{N_{(AB)} N_C}$
McQuitty Bağlantı (McQuitty Linkage)	$d_{mj} = (d_{kj} + d_{lj})/2$
Ward Kümeleme (Ward’s Method)	$DW_{pq} = W_t - W_p - W_q$
Küresel Ortalama Bağlantı (Centroid Method)	$S_{tr} = \frac{n_m}{2n_m} S_{mr} + \frac{n_q}{n_m+n_q} S_{qr} + \frac{n_m \cdot n_q}{n_m+n_q} S_{mq}$
Ortanca (Medyan) Bağlantı (Median Method)	$S_{tr} = \frac{1}{2} (S_{mr} + S_{qr}) - \frac{1}{4} S_{mq}$
Lance Williams Esnek Kümeleme (Lance & William’s Flexible Clustering Method)	$S_{tr} = \alpha_m S_{rm} + \alpha_q S_{rq} + \beta \cdot S_{mq} + \gamma S_{rm} - S_{rq} $

Hiyerarşik olmayan (non-hierarchical) ya da aşama sıralı olmayan kümeleme teknikleri ise, küme sayısı hakkında ön bilgi olması durumunda veya araştırmacının anlamlı olacak küme sayısına karar vermiş olması durumunda tercih edilen tekniklerdir.

Bu süreçte birden çok gözlemleri kümenin vektör olarak gösterilebilmesi amacıyla değişkenlerin ortalama değerlerinden yeni vektör oluşturmakta ya da bu kümedeki tüm gözlemler ile başka kümedeki gözlemlerin uzaklık ortalamaları da kullanılabilir.

Sosyal bilimlerde, tıpta, ziraat başta olmak üzere tüm mühendislik bilimlerinde yaygın uygulama imkânı bulunan kümeleme analizi, çok değişkenli varyans analizi, lojistik regresyon analizi, çok boyutlu ölçekleme gibi diğer çok değişkenli analizlerle de sıkı ilişkisi olan bir tekniktir [37, 39].

Kümeleme analizinde ilk aşama, veri girişi aşamasıdır. Verilerin, kümelemeye uygun biçimde girilmesi ile ilgili olan bu aşamada uzaklıklar matrisi elde edilir. İkinci aşama, kümeleme tekniğinin seçilmesi ve uygulanmasıdır. Son aşama ise sonuç aşaması olup, bu aşamada sonuçların duyarlılığı ve anlamlılığı incelenir. Sonuçların uygun olmaması durumunda (değişkenlerin uygun olmaması ve/veya küme sayısının doğru belirlenmemiş olması nedeniyle) tekrar ikinci aşamaya dönülmektedir.

Kümeleme analizinin genel amacının yanı sıra aşağıdaki özel amaçlar içinde kümeleme analizine başvurulabilir:

- Gerçek tiplerin (cinslerin-ırkların) belirlenmesi
- Model uydurmanın kolaylaştırılması
- Gruplar için ön tahmin
- Hipotezlerin testi
- Veri yapısının netleştirilmesi
- Veri indirgenmesi (veriler yerine kümelerin değerlendirilmesi)
- Aykırı değerlerin (outliers) bulunması [39].

Hiyerarşik olmayan kümeleme teknikleri başlığı altında birçok teknikten söz edilmektedir. Bu tekniklerden en çok kullanılan iki tanesi k-ortalama tekniği

ve en çok olabilirlik tekniğidir. Ancak, en çok olabilirlik tekniği, teorik dayanağı güçlü olmasına rağmen uzun hesaplama süresi gerektirdiği için yaygın kullanılmamaktadır [39].

K – Ortalamalar Tekniği: Bu teknikte bireyler, kümeler içi kareler toplamı minimum olacak şekilde k kümeye bölünmektedir. $\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n$ her biri p değişkenli gözlem vektörleri, çok boyutlu $\mathbf{X}$ uzayında birer nokta iken, aynı uzayda $a_{1n}, \dots, a_{kn}$ her grup birey için küme merkezleri olarak seçildiğinde,

$$W_n = \frac{1}{n} \sum \min_{1 \leq j \leq k} \|x_i - a_{jn}\|^2 \quad (3.18)$$

kuralı gereğince bireyler en yakın kümeye sınıflandırılmaktadır. Bu tekniği dört adımdan oluşan işleyiş algoritması ile ifade edersek,

1. İlk k gözlemin her biri bir gözlemlilik küme olarak alınır.
2. Kalan $n-k$ gözlemin her biri, ortalaması en yakın olan kümeye atanır ve her atamadan sonra küme ortalamaları tekrar hesaplanır.
3. Tüm gözlemlilerin kümelere atanması bittikten sonra, n gözlemin son bulunmuş küme ortalamalarına göre yeniden atamaları yapılır.
4. Bir önceki kümelemeye göre son elde edilen kümelemede kümeler arası gözlem geçişi durana kadar üçüncü adım tekrarlanır [39].

2.5.2. Karar Ağaçları

Karar ağaçları sınıflandırmada, kümelemede ve kestirim görevlerinde kullanılan kestirimci modelleme tekniğidir [16]. Karar ağaçları sınıflandırma problemlerini, test verisinin özellikleri hakkında dikkatle seçilmiş bir dizi soruyu sorarak çözer. Ağaçların kökünü oluşturan ilk sorudan sonra, her düzey o aşamada sorulacak sorulardan oluşturur ve her seferinde bir cevap alındığında, kaydın sınıf etiketi hakkında bir çıkarsamaya ulaşıncaya kadar bir takip sorusu sorulur. Bir soru serisi ve olası cevapları düğümlerden ve yönlendirilmiş uçlardan oluşan karar ağaçları hiyerarşik bir yapıya sahiptir [16, 20, 40].

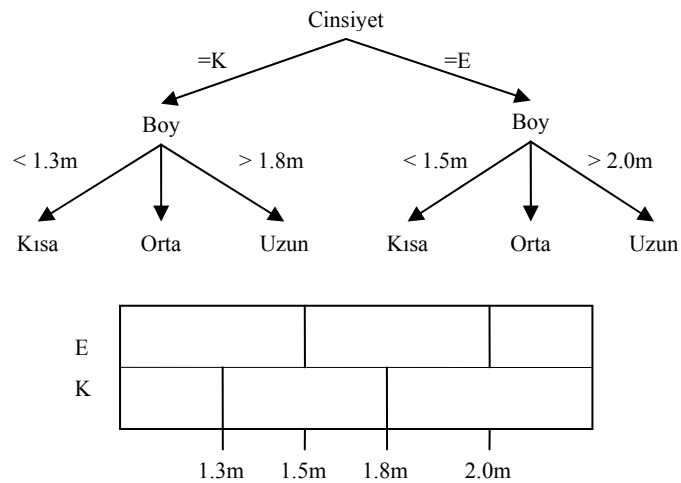
Karar ağaçları üç tip düğüme sahiptir:

1. **Kök düğüm**; gelen bir uca sahip olmayan ve sıfır veya daha fazla giden uca sahip olan bir düğümdür.
2. **İç düğüm**; her birine sadece bir ucun geldiği ve çıkan bir veya daha fazla ucun olduğu düğümlerdir.
3. **Yaprak (terminal) düğüm**; her birine sadece bir ucun geldiği ve çıkan bir ucun olmadığı düğümlerdir [16, 40].

Karar ağacında, her yaprak düğümü sınıf etiketine atanır. Kök ve diğer iç düğümleri içeren terminal olmayan düğümler, farklı karakteristiklere sahip olan kayıtları ayırmak için test koşullarını taşır [16, 20, 40].

Karar ağacı bir kez oluşturulduğunda, sınıflandırma, kök düğümden başlayarak test koşulları uygulanarak, sonuca göre uygun dal takip edilir. Bu işlem bir test koşulunun uygulandığı başka bir iç düğüme veya yaprak düğüme götürür. Eğer yaprak düğüme ulaşılmışsa, yaprak düğümle ilişkilendirilmiş sınıf etiketi kayda atanır. Böylece, kayıt sınıflandırılmış olur. Yaprak düğümleri kestirilmiş nesne için başarılı bir tahmini temsil eder [16, 40].

Sınıflandırma sonucunda, karar ağacı Şekil 2.5'deki gibi, araştırma uzayını yani N hacimli veri setini dikdörtgen bölgelere ayırmış olur [16, 20].



Şekil 2.5. Bir Karar Ağacının yapısı ve veri setinin bölünmesi

Karar ağacını bir tanımını yapmak gerekirse; $t_i = \{t_{i1}, t_{i2}, \dots, t_{ih}\}$ iken $D = \{t_1, t_2, \dots, t_h\}$ veritabanı verilmiş olsun ve veritabanı şeması $A_1, A_2, \dots, A_h$ özelliklerini içersin. Ayrıca $C = \{C_1, C_2, \dots, C_m\}$ sınıf seti verilmiş olsun. Karar ağacı, D ile ilişkilendirilmiş olan aşağıdaki özelliklere sahip bir ağaçtır:

- Her iç düğüm bir A_i özelliği ile etiketlenir.
- Her atlama üstteki ile ilişkilendirilmiş özelliğe uygulanabilen karşılaştırma belirtimi ile etiketlenir.
- Her yaprak düğüm bir C_j sınıfıyla etiketlenir. [16]

Sınıflandırma için Karar Ağaçlarının kullanılmasının birçok avantajı vardır. Karar ağaçlarının kullanımı kolay ve etkindir. Yorumlanması kolay ve anlaşılabilir kurallar yaratılabilir. Büyük veritabanları için ölçeklenebilirler çünkü ağacın boyutu veritabanının boyutundan bağımsızdır. Veritabanındaki her kaydı oluşturan veri kümesi ağaçlar vasıtasıyla filtrelenebilir. Bu işlem, ağacın yüksekliğiyle orantılı olarak, sabit olan bir zaman alır. Ağaçlar birçok özelliğe sahip veri için oluşturulabilir. Özellik değerlerinin dağılımı hakkındaki ve özelliklerin bağımsızlığı hakkındaki varsayımlarından bağımsızdır.

Karar Ağacı algoritmalarının dezavantajları da vardır. Birincisi, sürekli veriyi kolaylıkla ele alamazlar. Bu özellik tanım kümeleri (domain) ele alınabilecek kategorilere bölünmelidir. Kullanılan yaklaşım tanım kümesini (domain space) dikdörtgen bölgelere bölmektir. Sınıflandırma problemlerinin tümü bu tipte değildir. Kayıp verinin işlenmesi de zordur çünkü ağaçtaki doğru dallar alınamaz. Karar Ağacı eğitim verisiyle oluşturulduğundan, gereğinden fazla uydurma meydana gelebilir. Son olarak, veritabanındaki özellikler arasındaki ilişkiler Karar Ağacı süreci tarafından göz ardı edilir [16, 20].

Karar ağacı oluşturmak için çok çeşitli algoritmalar vardır. Ağaç inşası aşamasını basit Karar Ağacı inşa algoritması ile açıklamak gerekirse, bölümlerin yer alacağı ağaç ve çevresindeki düğümlerin adlandırılmasında kullanılacak olan veritabanı şemasındaki özellikler ayırma (splitting) özellikleri olarak adlandırılır. Ağaçtaki atlamalarla (arc) etiketlenen karşılaştırma belirtimleri (predicate)

ayırma (splitting) karşılaştırma belirtileri olarak adlandırılır. Bu tekrarlı (recursive) algoritma, ağacı, eğitim verisini inceleyerek yukarıdan aşağı biçimde (fashion) inşa eder. Başlangıç eğitim verisini kullanarak, ilk önce “en iyi” ayırma özelliği seçilir. Algoritmalar ayırma için kullanılacak “en iyi özelliği” ve onun “en iyi karşılaştırma belirtim”lerini nasıl belirlediğine göre değişir. En iyi ayırma özelliği belirlendikten sonra düğüm ve atlamalar yaratılır ve ağaca eklenir. Algoritma tekrarlı olarak her dallanma atlamasına (arc) yeni alt ağaçlar ekleyerek devam eder. Algoritma bazı “durma ölçüt”lerine ulaştığında işlemleri durdurur. Yine, her algoritma ağacı ne zaman durduracağını farklı biçimde belirler. Basit bir yaklaşım azaltılmış bir eğitim setindeki kayıtları oluşturan tüm veri setleri aynı sınıfa ait olduklarında durmak olacaktır. Bu sınıf daha sonra yaratılmış yaprak düğümünün etiketlenmesinde kullanılır.

Karar ağacı inşa algoritmalarının performansındaki temel faktörleri eğitim setinin büyüklüğü ve en iyi ayırma özelliğinin nasıl seçildiğidir. Çoğu karar ağacı algoritmalarının karşılaştığı konular şunlardır:

- **Ayırma özelliklerinin seçilmesi:** Ayırma özellikleri için hangi özelliklerin kullanılacağı inşa edilmiş karar ağacının uygulanmasındaki performansını etkiler (impact). Bazı özellikler diğerlerinden daha iyidir. Örneğin, bir veritabanında yer alan kaydın isim özelliği kesinlikle kullanılmamalıdır. Özelliğin seçimi, eğitim setindeki verinin incelenmesinin yanında alan uzmanlarının bilgilendirilmiş girdisini de gerektirir.
- **Ayırma özelliklerinin sıralanması:** Özelliklerin seçildiği sırada önemlidir. Bazı durumlarda birinci olarak seçilen bir özellik, gereksiz karşılaştırmaları içerebileceğinden, ikinci bir sefer incelenmeyi gerektirebilir.
- **Ayırmalar:** Özelliklerin sıralanması ile ilişkilendirilenler alınacak ayırmaların sayısıdır. Bazı özelliklerle birlikte, tanım kümesi (domain) küçüktür, dolayısıyla ayırmaların sayısı açıkça tanım kümesine dayanır. Buna rağmen, eğer tanım kümesi sürekli ise veya çok sayıda değere sahipse, kullanılacak ayırmaların sayısı kolaylıkla belirlenemez.

- **Ağaç yapısı:** Sınıflandırma için ağacın uygulanma performansını iyileştirmek için, en az seviyeye sahip dengelenmiş bir ağaç arzu edilir. Buna rağmen, bazı durumlarda çok yollu dallanmalı daha karmaşık karşılaştırmalara ihtiyaç duyulabilir. Bazı algoritmalar sadece ikili (binary) ağaçlar inşa eder.
- **Durma ölçütü:** Ağacın oluşturulması, veri seti mükemmel bir şekilde sınıflandırıldığında kesinlikle durur. Daha büyük ağaçların yaratılmasını önlemek için daha önce durulmasının arzu edilebileceği durumlar olabilir. Bu sınıflandırmanın doğruluğu (accuracy) ve performansı arasındaki bir alış verıştır. Buna ilave olarak daha erken durmak, gereğinden fazla uyumu bertaraf etmek içinde uygulanabilir.
- **Eğitim verisi:** Oluşturulan karar ağacının yapısı eğitim verisine bağlıdır. Eğer eğitim verisi çok küçük ise, o zaman oluşturulmuş ağaç daha genel veriyle düzgün çalışacak kadar belirli (specific) olmayabilir. Eğer eğitim verisi çok büyük ise, o zaman yaratılan ağaç gereğinden fazla uyum gösterebilir.
- **Budama:** Ağaç oluşturulduktan sonra, sınıflandırma aşaması esnasında ağacın performansını iyileştirmek için ağaçta bazı değişikliklere ihtiyaç duyulabilir. Budama aşaması daha iyi performansa ulaşabilmek için, gereksiz (redundant) karşılaştırmaları veya alt ağaçları kaldırabilir.

Bazı popüler Karar Ağacı teknikleri ID3, C4.5, C5.0, CART ve Ölçeklenebilir Karar Ağacı Teknikleri olarak sıralanabilir [16].

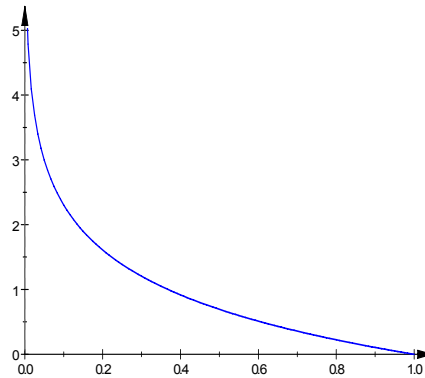
2.5.2.1. ID3 Algoritması

Karar ağaçları kurmak için ID3 tekniği bilgi teorisi üzerine kurulmuştur ve karşılaştırmalarının beklenen sayısı minimize etmeye çalışır. Başlatma (induction) algoritmasının temel fikri en fazla bilgiyi sağlayan cevapların sorusunu sormaktır. ID3'ün kullandığı temel strateji ilk elde edilen en yüksek bilgiye sahip bölme özelliklerini seçmektir. Bu özellik değeri ile ilişkilendirilmiş bilginin miktarı ortaya çıkma olasılığına ilişkindir.

Bilgiyi nitelemek için kullanılan kavram entropidir. Entropi bir veri setindeki düzensizliğin veya rassallığın miktarını ölçmek için kullanılır.

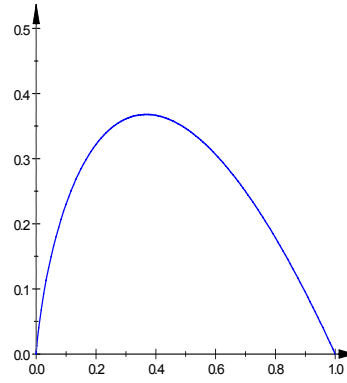
Bir setteki tüm verinin tek bir sınıfa ait olması durumunda, düzensizlik yoktur. Bu durumda entropi sıfırdır. Karar ağacı sınıflandırmasının amacı verilen veri setini her bir final altkümesindeki tüm elemanların aynı sınıfa ait olduğu altkümelere tekrarlamalı olarak ayırmaktır.

Şekil 2.6'da $\log(1/p)$ $[0, 1]$ aralığında p olasılığı olarak gösterilir. Bu sürprizin miktarını olasılığı temel alarak sezgisel olarak gösterir. $p = 1$ olduğunda sürpriz yoktur. Bunun anlamı eğer bir olayın olasılığı 1 ise, olay ortaya çıktığında, sürpriz olmayacaktır. p sıfıra yaklaştığında ise, sürpriz artacaktır. Böl ve keşfet yaklaşımı karar ağaçlarındaki gibi kullanıldığında, bölünme sonuçları toplamı 1 olan çoklu olasılıklardır. Bölünmeyle ilişkilendirilmiş bilgiyi ölçmek için, her iki olayla ilişkilendirilmiş bilgi birleştirilebilmelidir. Öyle ki, bölünmeyle ilişkilendirilmiş ortalama bilgiyi hesaplayabilmelidir. Bu iki değer birlikte toplanarak ve her birinin ortaya çıktığındaki olasılığı hesaba katarak gerçekleştirilir.



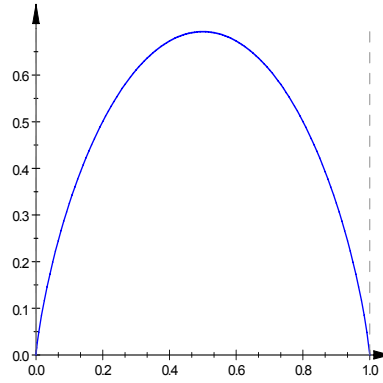
Şekil 2.6. $\log(1/p)$

Şekil 2.7'de bir olayın olasılığına dayanan beklenen bilgi olan $p \log(1/p)$ fonksiyonunu gösterir.



Şekil 2.7. $p \log(1/p)$

İki olayla ilişkilendirilmiş beklenen bilgiyi belirlemek için, bireysel değerler toplanır. Bu fonksiyon $p \log(1/p) + (1-p) \log(1/(1-p))$ dir ve Şekil 2.8’de gösterilmiştir. Buradaki önemli nokta maksimum bilgi iki olasılığın eşit olması durumunda meydana gelir.



Şekil 2.8. $H(p, 1-p)$

Entropinin tanımı, $\sum_{i=1}^s p_i = 1$ iken $p_1, p_2, \dots, p_s$ olasılıkları verildiğinde,

$$H(p_1, p_2, \dots, p_s) = \sum_{i=1}^s p_i \log_e(1/p_i) \quad (3.19)$$

olarak tanımlanır. Olasılıkların hepsi eşit olduğunda maksimum değere ulaşır.

Verilmiş bir veritabanı durumunda, D , $H(D)$ o durumda bulunma (veya bulunmama) miktarını bulur. Bu durum yeni s durumlarına bölündüğünde $S = \{D_1, D_2, \dots, D_s\}$, bu durumların entropisine bakılır. ID3’te her adımda

düzenleri en çok bölen durumu seçer. İçindeki kayıtları oluşturan tüm veri setleri aynı sınıfta olan bir veritabanı durumu tamamen düzenlidir. ID3 bilgideki en yüksek kazanıma sahip bölme özelliğini seçer. Kazanım bölünmeden önce doğru bir sınıflandırma yapmak için ne kadar bilgiye ihtiyaç duyulduğuna karşı bölünmeden sonra ne kadar bilgiye ihtiyaç duyulduğu arasındaki fark olarak tanımlanır. Ayırma en büyük miktar tarafından ihtiyaç duyulan bilgiyi azaltmalıdır. Bu bilgi orijinal veri setinin entropileri ve alt bölünmüş veri setlerinin her birinin entropilerinin ağırlıklı toplamaları arasındaki farklar belirlenerek hesaplanır. Bölünmüş veri setlerinin entropilerine ilgili bölüme yerleştirilmiş veri setinin parçası ile ağırlık verilir. ID3 algoritması belirli bir bölümün kazanımını aşağıdaki formül ile hesaplar:

$$Kazanım(D, S) = H(D) - \sum_{i=1}^s P(D_i) H(D_i) \quad (3.20)$$

[16]

2.5.2.2. C4.5 Algoritması

Karar ağacı algoritması C4.5 aşağıdaki yollarla ID3 algoritmasını iyileştirir:

- **Kayıp Veri:** Karar ağacı kurulduğunda, kayıp veri basitçe göz ardı edilir. Öyle ki, kazanım oranı sadece ilgilenilen özellik için değere sahip diğer kayıtlara bakılarak hesaplanır. Kayıp özellik değerine sahip bir kaydı sınıflamak için, o öge için değer diğer kayıtların özellik değerleri hakkında ne bilindiğine dayanarak kestirilebilir.
- **Sürekli Veri:** Temel fikir veriyi, eğitim örneğinde ilgili öge için bulunan özellik değerlerine dayanarak aralıklara bölmektir.
- **Budama:** C4.5'te önerilen iki birincil budama stratejisi vardır:
 1. *Alt ağaçla yer değiştirme* ile eğer yer değiştirme orijinal ağaçtakine yakın bir hata oranıyla sonuçlanıyorsa yaprak düğümü bir alt ağaç ile yeri değiştirilir.
 2. *Alt ağaç yükseltme* olarak adlandırılan başka bir budama stratejisi de, bir alt ağacı bu ağacın en çok kullanılan alt ağacı ile değiştirir.

Burada bir alt ağaç o anki konumundan ağaçtaki daha yüksek bir konuma yükseltilir. Yine, bu yerleştirme için hata oranındaki artış belirlenmelidir.

- **Kurallar:** C4.5 karar ağaçları ile veya karar ağaçlarından yaratılmış kurallarla sınıflandırmaya izin verir. İlave olarak, karmaşık kuralları basitleştirmek için bazı teknikler sunar. Bir yaklaşım, eğer eğitim setindeki tüm kayıtlar özdeş olarak işleniyorsa, kuralın sol tarafını daha basit bir sürümü ile değiştirmektir. Bir “diğer durum” tipindeki kural eğer diğer başka kurallar uygulanmayacaksa ne yapılması gerektiğini göstermekte kullanılabilir.
- **Bölümleme:** ID3 yaklaşımı çok bölmeli özellikleri destekler (favors) ve bu nedenle gereğinden fazla uyuma neden olabilir. Uç noktada, eğitim setindeki her tuple için benzersiz bir değere sahip bir özellik en iyisi olacaktır çünkü her bölme için tek bir tuple (ve bu yüzden tek sınıf) olacaktır. İyileştirme her bölmenin en önemliliği hesaba katılarak yapılabilir. Bu yaklaşım Kazanımın aksine Kazanım oranını kullanır. Kazanım oranı;

$$Kazanım\ Oranı(D, S) = \frac{Kazanım(D, S)}{H\left(\frac{|D_1|}{D}, \dots, \frac{|D_s|}{D}\right)} \quad (3.21)$$

ile tanımlanır.

Bölümleme amaçları için, C4.5 ortalama kazanımından daha büyüğünü sağlayan en büyük Kazanım Oranını kullanır. Bu Kazanım Oranının bir alt kümenin o alt kümenin başladığının hacmine yakın olduğu bölümlemelere karşı çarpık olması gerçeğini dengelemek içindir [16].

2.5.2.3. C5.0

C5.0 Clementine ve RuleCuest gibi çokça kullanılan birçok veri madenciliği paketinde kullanılan C4.5’in ticari sürümüdür. Büyük veritabanlarına karşı kullanımı hedeflenmiştir. Karar ağacı indüklemesi C4.5’tekine yakındır,

fakat kural oluřturma farklıdır. C4.5'ten farklı olarak, C5.0 için kullanılan kesin algoritmalar ifřa edilmemiřtir. C5.0 kural oluřturma için iyileřtirmeler içerir. Sonular C5.0'in hafıza kullanımını %90 iyileřtirdiğini, C4.5'ten 5,7 kez ve 240 kez arasında daha hızlı alıřtığını ve daha kesin kurallar ürettiğini göstermektedir.

C5.0'in kesinliğı için temel bir iyileřtirme arttırma (boosting) üzerinedir. Arttırma farklı sınıflandırıcıları birleřtirmek için bir yaklařımdır. Arttırma normalde belli bir sınıflandırıcıyı alıřtırmak için aldığı zamanı arttırırken, kesinliğı iyileřtirir. Hata oranının bazı veri setlerinde C4.5 ile bulunanın yarısından daha az olduğı gösterilmiřtir. Veri seti ok fazla gürültüye sahip olduğunda arttırma her zaman yardımcı olmamaktadır. Arttırma bir eğitim setinden oklu eğitim setleri yaratarak alıřır. Eğitim setindeki her nesneye bir ağırlık atanır. Ağırlık bu nesnenin sınıflandırmadaki önemini gösterir. Sınıflandırıcı kullanılan ağırlıkların her kombinasyonu için oluřturulur. Böylece, oklu sınıflandırıcılar fiilen oluřturulur. C5.0 bir sınıflandırma yaptığında, her sınıflandırıcı bir oya (vote) atanır, oylama yapılır ve hedef tuple en ok oya sahip sınıfa atanır [16].

2.5.2.4. Sınıflandırma ve Regresyon Ağları (CART)

Sınıflandırma ve regresyon ağları (Classification and Regression Trees - SRA) ikili karar ağları yaratan bir tekniktir. ID3'teki gibi, en iyi bölme özelliğini belirlemek için entropi kullanılır. Her alt kategori için bir ocuğın oluřturulduğı ID3'ün aksine yalnızca iki tane ocuk oluřturulur. Bölümleme en iyi bölme noktasının ne olacağı etrafında gerekleřtirilir. Her adımda, en iyi bölümü belirlemek için, en “iyi” tanımlamasının

$$\Phi(s/t) = 2P_L P_R \sum_{j=1}^m |P(C_j | t_L) - P(C_j | t_R)| \quad (3.22)$$

ile yapıldığı, ayrıntılı bir araştırma kullanılır.

Bu formül o anki düğüm t 'de ve her olası bölümleme özelliğı ve ölçütü s 'de değıerlendirilir. Burada L ve R ağta o anki düğümün sağı ve sol alt ağlarını ifade etmek için kullanılır. P_L ve P_R eğitim setindeki bir kayıt veri kümesinin ağın sol veya sağı tarafında olacağının olasılığıdır. Bu olasılık

Alt ağaçtaki tuple'lar / Eğitim setindeki tuple'lar ile tanımlanır. Eşitliğe sağ dalın alındığı varsayıldığında $P(C_j | t_L)$ veya $P(C_j | t_R)$, kaydı oluşturan veri kümesinin C_j sınıfında olması ve sol veya sağ alt ağaçta olma olasılığıdır. Bu olasılık $|Alt ağaçtaki j sınıfının ver kümeleri| / Hedef düğümdeki veri kümeleri$ ile tanımlanır. Her adımda, tüm olası ölçütlerden sadece bir tane ölçüt en iyi olarak seçilir.

SRA kullanılacak özelliklerin bir sıralaması için zorlar. SRA eksik veriyi, o özelliğin bölünme iyiliğinin hesaplanmasında, göz ardı ederek ele alır. Ağaç, bölünmenin performansının iyileşmesinin durması durumunda, büyümesini durdurur. Dikkat edilmesi gereken bir durumda, eğitim verisi için en iyisi olmasına rağmen, gelecekte eklenecek olası tüm veri için en iyisi olmayabileceğidir. SRA algoritması bir budama stratejisine de sahiptir [16].

2.5.2.5. Ölçeklendirilebilir Karar Ağacı Teknikleri

SPRINT (Scalable PaRallelizable INduction of desicion Trees) algoritması SRA tekniğinin kullanılabilir ana hafızayı dikkate almadan uygulanabilirliğini sağlayarak ölçeklendirilebilirlik konusundan bahseder. Buna ek olarak, kolaylıkla paralelleştirilebilir. SPRINT'te en iyi bölümün bulunması için gini indeksi kullanılır. D veritabanı için *gini*, p_j D veritabanındaki C_j sınıfının frekansı iken,

$$gini(D) = 1 - \sum p_j^2 \quad (3.23)$$

ile tanımlanır. D 'in D_1 ve D_2 alt kümelerine bölümünün iyiliği,

$$gini_{split}(D) = \frac{n_1}{n}(gini(D_1)) + \frac{n_2}{n}(gini(D_2)) \quad (3.24)$$

ile tanımlanır. En iyi gini değerine sahip bölüm seçilir. Daha önceki yaklaşımlardan farklı olarak, SPRINT karar ağacının indüklenmesi süreci esnasında verinin iyilik değerine göre sıralanmasına ihtiyaç duymaz. Sürekli veriyle, bölüm noktası eğitim setinin her ardışık değerler çiftinin orta noktası olacak şekilde seçilir.

Veritabanı özelliklerini ilgilendiren metadata kümesini sağlayarak, *RainForest* yaklaşımı eğitim setine ihtiyaç duymadan bölme özelliğinin seçimine izin verir. Karar ağacının her düğümü için, özellik – değer sınıfı (attribute value class – AVC label group) etiket grubu olarak adlandırılan bir tablo kullanılır. Tablo bir özellik için sınıf başına girdilerin sayısını veya özellik değer gruplanmasını özetler. Böylece, özellik – değer sınıfı tablosu bölümlene özelliklerinin belirlenmesi için ihtiyaç duyulan bilgiyi özetler. Tablonun boyutu veritabanı veya eğitim setinin boyutuyla orantılı değil, fakat aksine sınıf, benzersiz özellik değerlerinin veya potansiyel bölümlene özelliklerinin sayılarının çarpımıyla orantılıdır. Boyuttaki bu indirgeme (büyük eğitim setleri için) karar ağacı indüklemeye algoritmalarının aşırı derecede büyük eğitim setlerine ölçeklenebilmesine olanak sağlar. Ağaç inşası aşaması esnasında, eğitim verisi taranır, özellik – değer sınıfı kurulur ve en iyi bölümlene özelliği seçilir. Algoritma eğitim setini bölerek ve bir sonraki düğüm için özellik – değer sınıfını oluşturarak devam eder [16].

2.5.3. Birliktelik Kuralları

Birliktelik kuralları veri madenciliğinin temel tekniklerinden bir tanesidir ve belki de yönlendirilmemiş öğrenme sistemlerindeki yerel – örüntü keşfinin en yaygın şeklidir. Bu teknik genellikle diğer olayların meydana gelişi verildiğinde belirli bir olayın ortaya çıkışının olasılıklarını veya meyillerinin ölçülmesiyle ilgilidir. Bu metodolojiler veritabanındaki tüm olası ilginç örüntülere erişir. Bu bakılmamış nokta bırakmadığından bir anlamda bir güçtür, fakat güçsüzlük olarak ta görülebilir, çünkü, kullanıcı büyük miktarda yeni bilgi altında kolaylıkla ezilebilir ve yeni bilginin kullanılabilirlik analiz edilmesi zor ve zaman alıcıdır [19, 20].

Başka bir ürün satın alınırken bir ürünün satın alınması birliktelik kuralını tarif eder. Birliktelik kuralları pazarlamaya, reklamcılığa, yer düzenlemeleri ve stok kontrolüne yardımcı olmak için sıklıkla perakende satış mağazaları tarafından kullanılır. Perakendecilik sektöründe direk uygulanabilirliği olmasına rağmen, iletişim ağlarındaki kusurları kestirmek gibi farklı amaçlar içinde kullanılmıştır. Birliktelik kuralları veri öğeleri arasındaki ilişkileri göstermek için kullanılır. Ortaya çıkartılmamış bu ilişkiler, fonksiyonel bağımlılıklarda olduğu

gibi, verinin özünde değildir ve her hangi bir nedensellik türünü veya ilişkiyi temsil etmez. Onun yerine, birliktelik kuralları öğelerin genel kullanımını belirler. İçinde birliktelik kuralı bulunabilecek bir veritabanı, her kümenin bir öğeler kümesi taşıdığı, kaydı oluşturan veri kümesi olarak görülür.

Bir öğenin (veya öğeler setinin) *desteği*, o öğenin (veya öğelerin) ortaya çıktığı işlemlerin yüzdesidir.

Tanım 1: $I = \{I_1, I_2, \dots, I_m\}$ öğeler seti ve $t_i = \{I_{i1}, I_{i2}, \dots, I_{ik}\}$, $I_{ij} \in I$ iken $D = \{t_1, t_2, \dots, t_n\}$ işlemler veritabanı verilsin, **birliktelik kuralı** $X, Y \subset I$ öğe kümeleri olarak adlandırılan öğe kümeleri ve $X \cap Y = \emptyset$ iken, $X \Rightarrow Y$ formunun gerektirmesidir.

Tanım 2: Bir birliktelik kuralı $X \Rightarrow Y$ için **destek** (s) $X \cup Y$ içeren bir veritabanındaki işlemlerin yüzdesidir.

Tanım 3: Bir birliktelik kuralı $X \Rightarrow Y$ için **güven** veya **güç** (α), $X \cup Y$ içeren işlemlerin sayısının X taşıyan işlemlerin sayısına oranıdır. Genellikle tüm gerekliliklerle (implication) ilgilenilmez. Burada önem, Tanım 2 ve Tanım 3'te tanımlanan sıklıkla *destek* ve *güven* olarak isimlendirilen iki özellik ile ölçülür.

Birliktelik kurallarının seçimi Tanım 4'te birliktelik kuralı probleminin tanımında açıklandığı gibi bu iki değere dayanır. *Destek* kuralın veritabanında hangi sıklıkla ortaya çıktığını ölçerken, *güven* kuralın gücünü ölçer. Tipik olarak, büyük güven değerleri ve küçük bir destek değeri kullanılır. Destek kuralın veritabanının tümünde kaç kez ortaya çıktığının yüzdesini gösterdiğinden destek için küçük değerlere izin verilebilir. Birliktelik kuralında önceki ortaya çıktığında bağlı olan her zaman ortaya çıkar.

Tanım 4: $I = \{I_1, I_2, \dots, I_m\}$ öğeler seti ve $t_i = \{I_{i1}, I_{i2}, \dots, I_{ik}\}$, $I_{ij} \in I$ iken $D = \{t_1, t_2, \dots, t_n\}$ işlemler veritabanı verilsin, **birliktelik kuralı problemi** tüm $X \Rightarrow Y$ birliktelik kurallarının bir minimum destek ve güvenle tanımlamaktır. Bu değerler (s, α) probleme girdi olarak verilir.

Birliktelik kural algoritmalarının etkinliği genellikle gereken veritabanının taranması sayısına ve sayılması gereken öge setlerinin maksimum sayısına göre incelenir [16].

2.5.3.1. Büyük Öge Setleri (Large Itemsets)

Birliktelik kuralları bulmak için en uygun yaklaşım problemi iki parçaya bölmektir:

1. Tanım 5’te tanımlandığı gibi büyük öge seti bulmak
2. Sık öge setlerinden kurallar yaratmak

Bir öge seti I tüm öğeler setinin bir alt setidir.

Tanım 5: Büyük (sık) öge seti ortaya çıkma sayısı s eşiğinin üzerinde olan bir öge setidir. Büyük öge setlerinin tümünü göstermek için L ve belirli bir büyük öge setini göstermek için l gösterimi kullanılır.

Büyük öge seti bulunduktan sonra, herhangi ilgi çekici ilişki kuralı, $X \Rightarrow Y$, bu sık öge setleri setinde $X \cup Y$ ’a sahip olmalıdır. Herhangi bir büyük öge setinin alt seti de büyüktür. Birliktelik kuralı algoritmalarında kullanılan gösterimler Çizelge 2.9’daki gibidir.

Çizelge 2.9. Birliktelik kuralı gösterimleri

Terim	Tanım
D	İşlemler veritabanı
t_i	Veritabanı D ’deki işlem
s	Destek
α	Güven
X, Y	Öge setleri
$X \Rightarrow Y$	Birliktelik kuralı
L	Büyük öge setleri seti
l	L ’deki büyük öge seti
C	Aday öge setleri seti
P	Bölümlerin sayısı

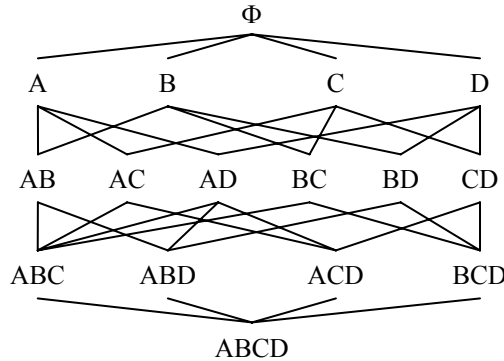
Büyük öge setlerini bulmak genellikle oldukça basittir fakat çok maliyetlidir. Saf yaklaşım herhangi bir işlemde görünen tüm öge setlerini saymak

olacaktır. m hacimli bir öğeler seti verilmişken 2^m alt set vardır. Boş setle ilgilenilmeyeceğinden, büyük öğeler setinin potansiyel sayısı $2^m - 1$ 'dir. Bu sayının patlarcasına büyümesi nedeniyle, birliktelik kuralı problemin, çözmeye davet sıklıkla tüm büyük öğe setlerinin etkin bir şekilde nasıl belirleneceği gibi görülür. Örneğin, $m = 5$ için potansiyel öğe seti sayısı 31'dir. Bu sayı $m = 30$ için bu sayı 1.073.741.823 olmaktadır. Çoğu birliktelik kuralı algoritması sayılacak öğe seti sayısını indirgemenin akıllıca yollarına dayanır. Bu potansiyel büyük öğe setleri *aday* ve tüm sayılmış (potansiyel olarak büyük) öğe setleri seti *aday öğe seti* (l) olarak adlandırılır. Birliktelik kuralı algoritmaları için kullanılan bir performans ölçümü C 'in hacmidir. Birliktelik kuralı algoritmaları tarafından çözülecek başka bir problem sayma süreci esnasında hangi veri yapısının kullanılacağıdır. Tüm büyük öğe setleri bulunduğunda birliktelik kurallarının yaratılması basittir [16].

2.5.3.2. Apriori Algoritması

Apriori algoritması en iyi bilinen birliktelik kuralı algoritmasıdır ve çoğu ticari ürünlerde kullanılır. *Büyük öğe seti özelliği* olarak adlandırılan özellik kullanılır: büyük öğe setinin herhangi bir alt seti büyük olmalıdır.

Büyük öğe setlerinin *aşağı doğru kapalı* olduğu da ifade edilir çünkü eğer bir öğe seti minimum destek gerekliliklerini sağlarsa, tüm alt setleri de sağlar. Bunun devriğine bakarak, eğer bir öğe setini küçük olduğu biliniyorsa, adaylar gibi süper setlerinin oluşturulmasına gerek yoktur çünkü onlarda küçük olmalıdır. Bu özellik Şekil 2.9'da gösterilmiştir.



Şekil 2.9 {A, B, C, D} için öğe setleri örgüsü (lattice)

Apriori algoritmasının temel fikri belirli bir hacimde aday öge setleri oluşturmak ve sonra eğer büyük iseler bunları görmek için veritabanını taramaktır. i taraması sırasında, i hacminin adayları, C_i sayılır. Sadece büyük olan adaylar bir sonraki geçiş için adayların oluşturulmasında kullanılır. Öyle ki L_i 'ler C_{i+1} 'lerin yaratılması için kullanılır. Bir öge seti sadece eğer tüm alt setleri de büyükse bir aday olarak görülür. $i+1$ hacminin adaylarını yaratmak için, bir önceki geçişte bulunan büyük öge setlerinin birleşimleri yapılır.

İlk geçişten sonraki her geçiş için aday öge setleri yaratmak için Apriori-Gen olarak adlandırılan bir algoritma kullanılır. Tüm tek öge setleri ilk geçişte aday olarak kullanılır. Burada önceki geçişin, L_{i-1} , büyük öge setleri seti adayları belirlemek için kendisiyle birleşmiştir. Her öge seti tüm öğeleri içermelidir fakat birleştirilebilmeleri için bir öğeleri ortak olmalıdır.

Apriori algoritması veritabanının hafızada yerleşik olduğu varsayar. Veritabanı taramalarının maksimum sayısı en geniş büyük öge setinin en önemliliğinden bir fazladır. Veritabanı taramalarının potansiyel büyük sayısı Apriori yaklaşımının bir zayıflığıdır [16].

2.5.3.3. Örnekleme Algoritması

Öge setlerinin büyük veritabanlarındaki etkin sayımına yardımcı olmak için, veritabanının örneklenmesi kullanılabilir. Orijinal örnekleme algoritması veritabanı taramalarının sayısını en iyi durumda bir ve en kötü durumda ikiye indirir. Veritabanı örneği hafızada yerleşik olabilecek şekilde çekilir. Sonra, Apriori gibi, herhangi bir algoritma örnek için büyük öge setlerinin bulunmasında kullanılabilir. Bunlar *potansiyel büyük* (PL) öge setleri olarak görülür ve tüm veritabanını kullanarak sayılacak adaylar olarak kullanılır. İlave adaylar *negatif sınır* (BD^-) fonksiyonu örnekten gelen büyük öge setlerine uygulanarak belirlenir. O zaman adaylar setinin tümü $C = BD^-(PL) \cup PL$ olur. Negatif sınıf fonksiyonu Apriori – Gen algoritmasının bir genelleştirmesidir. PL 'de olmayan, fakat alt setlerinin tümü PL 'de olan minimal öge setleri seti olarak tanımlanır [16].

2.5.3.4. Bölümlendirme

Büyük öge setlerinin oluşturulması için işlemler setinin bölünmesi dayanan çeşitli yaklaşımlar sunulmuştur. Bu durumda, $D, D^1, D^2, \dots, D^p$ gibi p tane bölüme bölünür. Bölümleme büyük öge setlerinin bulunma performansını birçok yoldan iyileştirebilir:

- Büyük öge seti özelliğinin avantajını alarak, bir büyük öge setinin bölümlerden en az birinde büyük olması gerektiği bilinmektedir.
- Bölümleme algoritmaları kısıtlı ana hafızaya daha iyi adapte olabilir. Buna ek olarak, bölüm başına sayılacak öge seti sayısının tüm veritabanı iyi ihtiyaç duyulandan daha az olabileceği beklenebilir.
- Bölümleme kullanarak, her bir bölümün ayrı bir makine tarafından ele alınabileceği, paralel ve/veya dağıtılmış algoritmalar kolaylıkla yaratılabilir.
- Birliktelik kurlarının artımlı yaratımları veritabanının o anki durumunu bir bölüm olarak ve yeni girdileri ikinci bir bölüm gibi ele alarak uygulanması daha kolay olabilir.

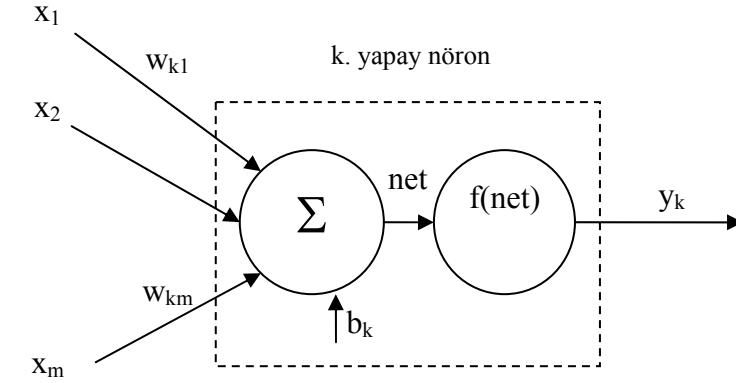
Temel bölümleme algoritmaları veritabanı taramalarının taramalarını ikiye indirir ve veritabanını, her biri ana hafızaya yerleştirilebilecek şekilde bölümlere böler. Veritabanını tararken, veritabanının o bölümü hafıza getirir ve o bölümdeki öğeleri tek başına tarar. Birinci veritabanı taraması esnasında, algoritma her bölümdeki tüm büyük öge setlerini bulur. Bu amaçla herhangi bir algoritma kullanılabilmesine rağmen, orijinal sunum, Apriori gibi, seviyeye bağlı algoritmalar kullanıldığını varsayar. Burada L^i, D^i bölümünden büyük öge setini temsil eder. İkinci tarama esnasında, sadece en az bir bölümde büyük olan öge setleri adaylar olarak kullanılır ve tüm veritabanında da büyük olup olmadıklarını belirlemek için sayılırlar [16].

2.5.4. Yapay Sinir Ağları

Bir sinir ağı, isminin de ifade ettiği gibi, tek yönde sinyal gönderen bağlantılarla bağlanmış birçok düğümden oluşan bir ağ yapısıdır. Her düğüm bir işlem birimini temsil eder ve düğümler arasındaki bağlantılar bağlanmış düğümler

arasındaki nedensel ilişkiyi belirler. Tüm düğümler, bu düğümlerin çıktılarının bu düğümlere mahsus olan değiştirilebilir parametrelere bağlı olduğu anlamında, uyarlanabilirler.

Bir yapay sinir ağının en basit bileşeni **yapay nöron**dur. Yapay nöron bir yapay sinir ağı operasyonuna temel olan bilgi işlem birimidir. Bir yapay nöron modeli Şekil 2.10'daki gibidir.



Şekil 2.10 Yapay nöron modeli

Bu bileşen üç temel öğeden oluşur. Bunlardan ilki, her birinin bir ağırlık veya güç w_{ki} ile karakterize edildiği farklı x_i girdilerden (veya sinaps) *bağlantılar setidir*. İlk indis sorgudaki nöronu işaret eder ve ikinci indis ise ağırlığın ilgilendiği sinapsın girdisini gösterir. Genelde, bir yapay nöronun ağırlıkları pozitif sayıları içerdiği kadar negatif sayıları da içeren bir aralıkta olabilir. İkinci öğe, bağlı olduğu sinaptik w_{ki} güçleri ile ağırlık verilmiş x_i girdi sinyallerini toplayan *toplayıcıdır*. Burada tanımlanan operasyon bir doğrusal birleştirici meydana getirir. Üçüncü ve son temel öğe ise, bir nöronun y_k çıktısının genliğini sınırlandıran f *aktivasyon fonksiyonudur*. Şekil 2.10'da da görüldüğü gibi yapay nöron modelinde bir de dışarıdan uygulanan b_k sapması yer almaktadır. Bu sapma aktivasyon fonksiyonunun pozitif veya negatif olmasına göre ağ girdisini artırma veya azaltma etkisine sahiptir.

Yapay nöron, doğal nöronun soyut bir modelidir. Birçok $x_i, i=1,2,\dots,m$ girdisi vardır. Her bir x_i girdisi, k yapay sinir ağındaki verilmiş bir nöronun indisi iken, ilgili ağırlık w_{ki} ile çarpılır. Ağırlıklar doğal nörondaki biyolojik

sinaptik güçleri taklit eder. Yapay sinir ağları literatüründe, $i = 1, 2, \dots, m$ için $x_i w_i$ çarpımlarının ağırlıklı toplamlar $ağ$ olarak gösterilir.

$$net_k = x_1 w_1 + x_2 w_2 + \dots + x_k w_k + b_k \quad (3.25)$$

$w_{k0} = b_k$ ve varsayılan girdi $x_0 = 1$ için benimsenmiş gösterimi kullanarak, ağ toplamının yeni düzenli bir versiyonu aşağıdaki gibi olacaktır.

$$net_k = x_0 w_0 + x_1 w_1 + x_2 w_2 + \dots + x_k w_k = \sum_{i=1}^m x_i w_{ki} \quad (3.26)$$

Aynı toplam $\mathbf{X} = \{x_0, x_1, x_2, \dots, x_m\}$ ve $\mathbf{W} = \{w_{k0}, w_{k1}, w_{k2}, \dots, w_{km}\}$ iken, vektör gösterimde iki m – boyutlu vektörün çarpımı olarak ta gösterilebilir.

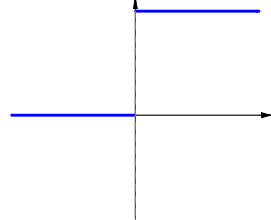
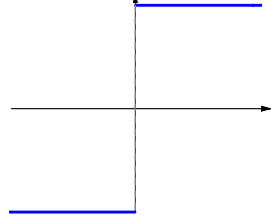
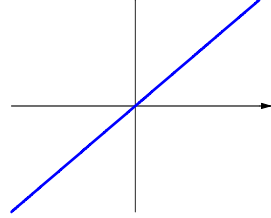
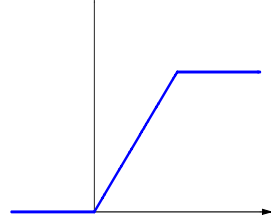
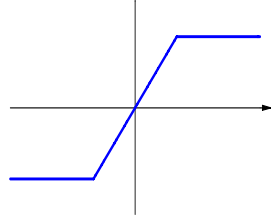
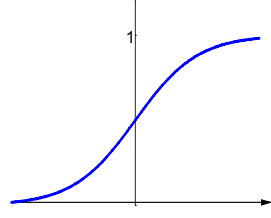
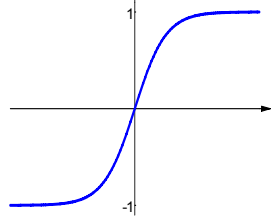
$$net_k = \mathbf{X} \cdot \mathbf{W} \quad (3.27)$$

Son olarak yapay nöron, y_k çıktısını net_k değerinin kesin bir fonksiyonu olarak hesaplar.

$$y_k = f(net_k) \quad (3.28)$$

Burada f aktivasyon fonksiyonu olarak adlandırılır ve çeşitli formlarda tanımlanabilir. Yaygın olarak kullanılan bazı aktivasyon fonksiyonları Çizelge 2.10’da verilmiştir [20].

Çizelge 2.10 Yaygın Aktivasyon Fonksiyonları

Aktivasyon Fonksiyonu	Girdi – Çıktı ilişkisi	Grafik
Hard Limit	$y = \begin{cases} 1, & net \geq 0 \\ 0, & net < 0 \end{cases}$	
Simetrik Hard Limit	$y = \begin{cases} 1, & net \geq 0 \\ -1, & net < 0 \end{cases}$	
Doğrusal	$y = net$	
Doyuran Doğrusal	$y = \begin{cases} 1, & net > 1 \\ net, & 0 \leq net \leq 1 \\ 0, & net < 0 \end{cases}$	
Simetrik Doyuran Doğrusal	$y = \begin{cases} 1, & net > 1 \\ net, & -1 \leq net \leq 1 \\ -1, & net < -1 \end{cases}$	
Log – Sigmoid	$y = 1 / (1 + e^{-net})$	
Hiperbolik Tanjant Sigmoid	$y = (e^{net} - e^{-net}) / (e^{net} + e^{-net})$	

2.5.5. Genetik Algoritma

GENETİK algoritmalar doğal seçim ve evrimsel sürece sıkı sıkıya bağlı olmayan biçim bağımsız, stokastik optimizasyon metotlarıdır. İlk defa 1975 yılında Michigan Üniversitesi'nde John Holland tarafından önerilmiş ve araştırılmıştır. Genetik algoritmalar parametre veya çözüm uzayındaki her bir noktayı kromozom olarak adlandırılan ikili dizgilere döndürür. n – boyutlu uzaydaki bu noktalar, diğer veri madenciliği metotlarındaki gibi eğitim ve test için önceden verilen veri setleri gibi değildirler. Genetik algoritmalarındaki n – boyutlu noktalar seti bir genetik algoritma parçasıdır ve optimizasyon sürecinde tekrarlanarak üretilirler. Her nokta veya ikili dizge çözülecek probleme potansiyel bir çözümü gösterir. Genetik algoritmalarda, optimizasyon probleminin karar değişkenleri doğal genetik sistemlerdeki kromozomlara benzer olan bir veya daha çok dizgi yapısıyla kodlanır. Kodlama dizgileri genlere benzer olan özelliklerden oluşur. Özellikler her bir özelliğin kendi pozisyonuna (konumuna) ve kesin bir alel değerine sahip iken, önerilen kodlama metoduyla uyum sağlayan, dizgede farklı konumlardadır. Kromozomlardaki dizge yapıları daha iyi alternatif çözümler üretmek için doğal evrim sürecine benzer farklı operasyonları gözden geçirir. Yeni kromozomların kalitesi, optimizasyon problemi için amaç fonksiyonu olarak düşünülebilen “uygunluk” değerine dayanarak tahmin edilir. Doğal evrim ve genetik algoritma kavramları arasındaki temel ilişkiler Çizelge 2.11'teki gibidir [20].

Çizelge 2.11. Genetik algoritmalarındaki temel kavramlar

Doğal Evrimdeki Kavram	Genetik Algoritmalarındaki Kavram
Kromozom	Dizge
Gen	Dizgedeki özellikler
Lokus	Dizgedeki pozisyon
Alel	Pozisyon değeri (genellikle 0 veya 1)
Genotip	Dizge yapısı
Fenotip	Karakteristikler seti (özellikler)

Genetik algoritmaların özellikle kullanışlı olduğu tipik durumlar analitik metotların iyi çalışmadığı zor optimizasyon durumlarıdır. Genetik algoritmalar kablo dağıtımı, zaman çizelgelemesi, uyarlamalı kontrol, oyun oynama, taşıma problemleri, seyahat eden satış elemanı problemleri, veritabanı sorgulama optimizasyonu, makine öğrenme, robotbilimi, üretim planlama, ekonomi, zaman

serisi analizi, örüntü tanımlama gibi optimizasyon problemlerinde ve sınıflandırma, kümeleme, birliktelik kuralları oluşturulmasını içeren çoğu problemin çözülmesinde oldukça başarılı bir biçimde uygulanmıştır [16, 41, 42]. Son yıllarda, optimizasyonunu önemi daha ileri düzeyde büyümüştür çünkü birçok önemli geniş ölçekli, tümleşik – optimizasyon problemleri ve son derece kısıtlı mühendislik problemleri sadece yaklaşık olarak çözülebilmektedir. Genetik algoritmalar bu tip karmaşık problemleri hedefler. Olasılıklı algoritmalar sınıfına aittirler, yinede yönlendirilmiş ve stokastik araştırma öğelerini birleştirdiklerinden rassal algoritmalarından çok farklıdır. Genetik temelli arama metotlarının önemli başka bir özelliği de diğer tüm metotlar araştırma uzayının bir tek noktasını işlerken, potansiyel çözümler topluluğu sağlamalarıdır. Bu özellikleri nedeniyle, genetik algoritmalar var olan yönlendirilmiş arama metotlarından çok daha istikrarlıdır (robust).

Genetik algoritmalar fonksiyonel biçimlere bağımlı olmamaları ve aşağıdaki karakteristiklere sahip olmalarından dolayı popülerdir:

1. İşlemlerini oldukça fazla hızlandırmak için paralel – işleme makinelerine tatbik edilebilen paralel – arama prosedürleridir.
2. Genetik algoritmalar sürekli veya kesikli problemlerinin her ikisine de uygulanabilir.
3. Genetik algoritmalar stokastiktir ve çoğu tatbiki, optimizasyon uygulamasında kaçınılmaz şekilde var olan, yerel minimumda tuzağa düşme olasılığı daha azdır.
4. Genetik algoritmaların esnekliği karmaşık modellerdeki yapı ve parametre tanımlamasının her ikisini de sağlar [20].

Genetik algoritmaları kullanmanın temel avantajı paralelleştirilebilir olmalarıdır. Fakat kullanımlarının birçok dezavantajı da vardır. Bu dezavantajlar aşağıdaki gibi sıralanabilir:

1. Genetik algoritmaların anlaşılması ve son kullanıcıya açıklanması zordur.

2. Problemin soyutlanması ve bireyleri (öğeleri) temsil etmesi için kullanılacak metodu tanımlamak oldukça zordur.
3. En iyi uyum fonksiyonunun belirlenmesi zordur.
4. Çaprazlama ve mutasyonun nasıl yapılacağının belirlenmesi zordur [16].

2.5.6. Bulanık Kümeler ve Bulanık Mantık

Şimdiye kadar sunulan yaklaşımlar verinin kesin (precise) olduğunu varsayar. Öyle ki, daha ileri analiz için tamı tamına ölçümlerin ele alındığını varsayar. Tarihsel olarak, klasik matematikte düşünüldüğü gibi, insanoğlu genellikle nesnelerin ve olayların kesin ve keskin (crisp) tanımlarının peşinden koşar. Bu kesinlik olgunun nümerik veya kategorik değerler ile ifade edilmesi yoluyla başılır. Fakat hepsinde olmasa da, gerçek dünya senaryolarının çoğunda tamamen kesin değerlere hiçbir zaman sahip olunamayacaktır. Her zaman bir düzensizlik derecesi olacaktır. Buna rağmen, klasik matematik bu bulanıklık nedeniyle önemli zorluklarla karşılaşacaktır. Çoğu gerçek dünya durumlarında, keskinlik (crispness) ve kesinlik basitleştirme ve idealleştirme iken bulanıklık gerçekliktir denebilir. Bulanıklık ile kesinlik arasındaki kutuplaşma modern bilgi işleme sistemlerinin gelişimindeki oldukça dikkat çekici bir çelişkidir. Zıtlığı çözmek için etkin bir araç bulanık küme teorisidir [20].

Bulanık mantık, temelde çok değerli mantık, olasılık kuramı ve yapay zeka alanları üzerine oturturulmuş bir kavramdır.

Bulanık mantık, incelenen olayın çok karmaşık olması ve bununla ilgili yeterli bilginin bulunmaması durumunda kişilerin görüş ve değer yargılarına yer verilmesine ihtiyaç gösteren durumlarda geçerli olmaktadır. Belirsizlik anlamına gelen bulanık mantık bu prensipten yola çıkarak kesin olmayan bir bilgiyi basit ve kullanışlı bir hale getirmektedir

Genel özellikleriyle bulanık mantık ifade edilmek istenirse,

- Bulanık mantıkta kesin değerlere dayanan düşünme yerine yaklaşık düşünme kullanılmaktadır.
- Bulanık mantıkta her şey $[0,1]$ aralığında belirli bir derece ile gösterilmektedir.

- Bulanık mantıkta bilgi büyük, küçük, çok az gibi dilsel ifadeler şeklindedir.
- Bulanık çıkarım işlemi dilsel ifadeler arasında tanımlanan kurallar ile yapılmaktadır.
- Her mantıksal sistem bulanık olarak ifade edilebilmektedir.
- Bulanık mantık matematiksel modeli çok zor elde edilen sistemler için oldukça uygun olmaktadır.
- Bulanık mantık tam olarak bilinmeyen veya eksik girilen bilgilere göre işlem yapma yeteneğine sahiptir.

Söz konusu kavramın temelini bulanık küme kuramı oluşturmaktadır. Bu kavrama hayattan kesin olmayan bir örnek verecek olursak, sıcak ve soğuk gibi kesin ifadelerin arasında kalan “az soğuk” veya “soğuğa yakın” gibi belirsizlikleri matematiksel olarak ifade etmede kullanılan bir teoremdir. Geleneksel mantık anlayışında olduğu gibi bulanık küme kuramında da (1) ve (0) değerleri vardır ancak bulanık mantık bu değerlerin yanı sıra, ara değerleri de kullanmaktadır. Bulanık küme kuramının temellerini 1965 yılında Azerbaycanlı matematikçi Prof Dr. Lotfi A.Zadeh oluşturmuştur.

Bulanık küme teorisinde, bulanık kümeleri içeren bir evrensel küme içerisindeki elemanların üyelik geçişi dereceli olmaktadır. Eğer bir eleman herhangi bir kümeye ait olacaksa, o elemanın o kümeye ait olma derecesi de söz konusu olmaktadır. Bu derecelendirme bulanık kümelerin sınırlarına belirsizlik özelliğini katmaktadır. Bu sebeple bir elemanın bu kümeye aitliği belirsizliğini ölçmeye yarayan bir fonksiyonla tanımlanabilmektedir. Söz konusu fonksiyon evrensel kümenin elemanlarını belirli bir aralıktaki reel sayılara karşılık getirerek elemanlar arasındaki derecelendirmeyi gerçekleştirmektedir. Küme içerisinde değişkenlerin aldığı yüksek değerler de üyelik derecesinin yüksekliğini göstermektedir. Buradaki fonksiyon üyelik fonksiyonu ve bu fonksiyonun oluşturduğu küme de “Bulanık Küme” olarak ifade edilebilmektedir. Bulanık bir A kümesini aşağıdaki şekilde ifade etmek mümkün olmaktadır:

X boş olmayan bir küme olmak üzere; X’ deki bir bulanık A kümesi

$$\forall x \in X \text{ için ; } \mu_A(x): X \rightarrow [0,1]$$

olarak ifade edilebilmektedir. Burada $\mu_A(x)$ 'e, bulanık kümeye karşılık gelen üyelik fonksiyonu adı verilmektedir. $\mu_A(x)$; A kümesinin elemanlarının istenilen özelliği hangi ölçüde sağladığının ifadesi olmaktadır.

Bulanık kümeler üzerine kurulan matematiksel yapı, klasik matematikten daha fazla açıklayıcı bir güce sahip olmasına karşın bulanık kümelerin kullanılabilirliği, uygulama alanlarında ortaya çıkan kavramlar için uygun üyelik fonksiyonlarının belirlenebilmesine bağlı olmaktadır.

Bulanık bir kümeyi ifade etmede üyelik fonksiyonları kullanılmaktadır. Bu sebeple üyelik fonksiyonlarının şekilleri ve bu fonksiyonların geliştirilmesi bulanık küme teorisi içinde önemli bir yer tutmaktadır.

Bulanık küme teorisinde üyelik fonksiyonlarının değer aralığı daha önce de belirtildiği üzere $[0,1]$ aralığı olmaktadır. İşte bir bulanık kümenin elemanlarını bu aralıktaki bir sayıya karşılık getiren fonksiyon da “üyelik fonksiyonu” olarak adlandırılmaktadır. Başka bir tanımla ifade etmek gerekirse, 0 ile 1 arasındaki değişimin her bir öge için değeri üyelik derecesi olarak adlandırılırken, üyelik derecelerinin bir alt küme içindeki değişimleri ise üyelik fonksiyonu olarak adlandırılmaktadır. Böylece üyelik fonksiyonu altında toplanan öğeler önem derecelerine göre birer üyelik derecesine sahip olmaktadır.

Küme kavramı kesikli ve sürekli kümeler olarak ikiye ayrılabilir. Gerek kesikli gerekse sürekli kümelerde bulunan elemanların o kümeye olan üyelik derecelerinin $[0,1]$ aralığında belirlenmesi ile de her iki kümede bulanık kesikli küme veya bulanık sürekli küme haline gelebilmektedir. Herhangi bir elemanın, hem kesikli bulanık kümeye hem de sürekli bir bulanık kümeye “üyelik derecesi” söz konusu olmaktadır. Ve bir x değeri için bir üyelik derecesi atanabilmektedir. Kesikli verilerde değerler kümesinin her bir elemanı için ayrı ayrı üyelik derecesi atanırken, değerler kümesi sürekli olduğunda ise üyelik dereceleri x değerlerine bağlı olarak $f(x)$ gibi bir fonksiyon halini alabilmektedirler.

Klasik kümelerin üyelik fonksiyonları ya bir nokta ya da bir doğru şeklinde iken bulanık kümelerin üyelik fonksiyonları nokta veya doğru olabildiği gibi doğrusal veya eğrisel bir fonksiyon şeklinde de olabilmektedir. Klasik

kümelerin üyelik fonksiyonları arasında bir geçiş bölgesi söz konusu değilken, bulanık kümelerin üyelik fonksiyonları içiçe geçebilmektedir. Uygulanacak üyelik fonksiyonlarının seçimi, uygulayacak olan kişi tarafından bulanık küme elemanlarının bulanık kümeye aitlik derecesine göre belirlenebilmektedir. Başka bir ifadeyle üyelik fonksiyonunun türünü uygulayacak olan kişinin görüşü belirleyebilmektedir.

Uygulamalarda üyelik fonksiyonları denetlenen sürecin özelliklerine göre en çok kullanılan fonksiyonlar olan, Üçgen üyelik fonksiyonu, Yamuk üyelik fonksiyonu, Gauss üyelik fonksiyonu ve Genelleştirilmiş Bell üyelik fonksiyonu olarak karşımıza çıkabilmektedir [43].

2.5.6.1. Üçgen Üyelik Fonksiyonu

Üçgen üyelik fonksiyonu, bulanık mantıkta hem giriş (input) hem de çıkış (output) parametrelerini tanımlamak için kullanılabilmektedir ve $\{a,b,c\}$ olmak üzere üç parametre ile özelleştirilmiştir. Söz konusu üyelik fonksiyonunun denklemi ise,

$$\mu_A(x) = \text{üçgen}(x; a, b, c) = \begin{cases} 0 & , x \leq a \\ \frac{x-a}{b-a} & , a \leq x \leq b. \\ \frac{c-x}{c-b} & , b \leq x \leq c. \\ 0 & , c \leq x. \end{cases} \quad (3.29)$$

şeklinde ifade edilmektedir [20, 43].

2.5.6.2. Yamuk Üyelik Fonksiyonu

Yamuk üyelik fonksiyonu, üçgen üyelik fonksiyonu gibi hem giriş hem de çıkış parametrelerini tanımlamak için kullanılabilmektedir. $\{a,b,c,d\}$ olmak üzere dört parametre ile özelleştirilmiştir. Yamuk üyelik fonksiyonunun denklemi ise,

$$\mu_A(x) = yamuk(x; a, b, c, d) = \begin{cases} 0 & , x \leq a. \\ \frac{x-a}{b-a} & , a \leq x \leq b. \\ 1 & , b \leq x \leq c. \\ \frac{d-x}{d-c} & , c \leq x \leq d \\ 0 & , d \leq x \end{cases} \quad (3.30)$$

şeklinde ifade edilmektedir [20, 43].

2.5.6.3. Gauss Üyelik Fonksiyonu

Gauss üyelik fonksiyonu, $\{c, \sigma\}$ gibi iki parametre ile özelleştirilmiştir. Denetlenecek olan süreçlerde hem giriş hem de çıkış parametrelerini tanımlamak için kullanılabilmektedir. Gauss üyelik fonksiyonunun ifadesi ise,

$$\mu_A(x) = gauss(x; c, \sigma) = e^{-\frac{1}{2} \left(\frac{x-c}{\sigma} \right)^2} \quad (3.31)$$

şeklindedir [20], [43].

2.5.6.4. Genelleştirilmiş Bell Üyelik Fonksiyonu

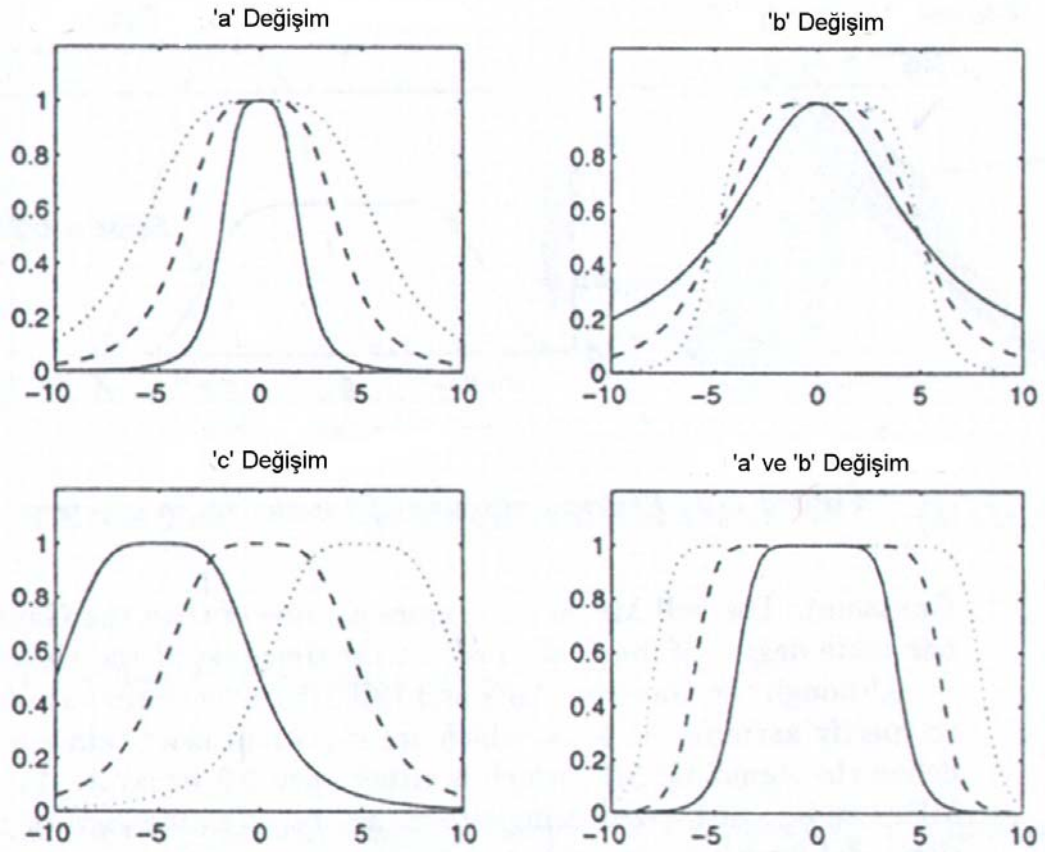
Genelleştirilmiş Bell üyelik fonksiyonu ise, bulanık mantık uygulamalarında hem giriş hem de çıkış parametrelerini tanımlamak için kullanılabilmektedir. $\{a, b, c\}$ olmak üzere üç parametre ile özelleştirilmiştir. Genelleştirilmiş Bell üyelik fonksiyonu “Cauchy dağılımının” genelleştirilmiş hali olduğundan “Cauchy üyelik fonksiyonu” olarak da adlandırılmaktadır. Bell üyelik fonksiyonunun denklemi ise,

$$\mu_A(x) = Bell(x; a, b, c) = \frac{1}{1 + \left| \frac{x-c}{a} \right|^{2b}} \quad (3.32)$$

şeklinde ifade edilmektedir [23, 43].

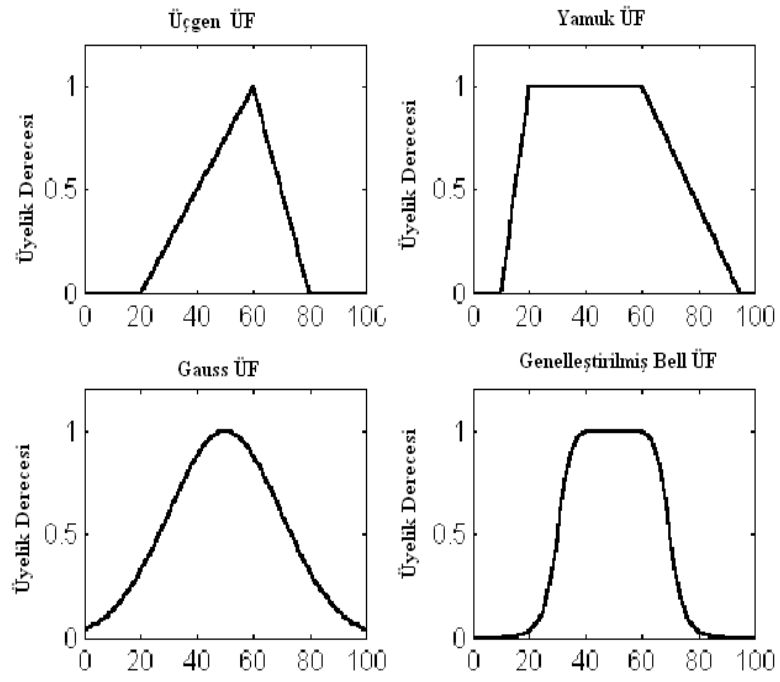
Söz konusu dağılımda a şekil parametresi olup fonksiyonun şeklinin değişmesini gerçekleştirirken, c konum parametresi olup fonksiyonun merkezini belirlemekte ve b parametresi ise a parametresi ile birlikte değişmektedir.

Dağılımda b parametresi geçiş noktalarını a parametresiyle birlikte kontrol etmektedir. Bu parametreler değiştiğinde dağılım da bu parametrelere bağlı olarak değişmektedir. Şekil 2.11’de $[a,b,c]$ parametrelerinin değişiminden fonksiyonun nasıl etkilendiğine ilişkin grafiksel gösterimlere yer verilmiştir:



Şekil 2.11 Genelleştirilmiş Bell üyelik fonksiyonunun parametre değişimleri

Üçgen üyelik fonksiyonun, yamuk üyelik fonksiyonun, Gauss üyelik fonksiyonun ve Genelleştirilmiş Bell üyelik fonksiyonunun şekilleri Şekil 2.12’de gösterildiği gibidir [20, 43].



Şekil 2.12. Üçgen, Yamuk, Gauss ve Genelleştirilmiş Bell Üyelik fonksiyonları

3. PORTFÖY, PÖRTFÖY YÖNETİMİ VE EŞ İŞLEM STRATEJİSİ

Finansal piyasalar bir ekonomide fon fazlası veren birimlerle, fon açığı veren birimleri bir araya getiren, köprü görevi gören piyasalardır. Finansal piyasalar, fon fazlası verenler yani tasarrufçular, fon açığı verenler yani yatırımcılar, finansal aracı kurumlar ile idari ve hukuki düzenden oluşur. Finansal piyasaları birçok açıdan sınıflandırmak mümkün olmakla birlikte en çok kullanılan sınıflandırma biçimi para ve sermaye piyasaları şeklinde olanıdır. Para piyasaları kısa vadeli fon arz ve talebinin bulunduğu piyasalardır.

Sermaye piyasaları ise uzun vadeli fon fazlası olanlarla yatırım projelerini gerçekleştirmek isteyen ve uzun vadeli fon açığı bulunanları bir araya getiren piyasalardır. Sermaye piyasaları tasarrufları etkin, verimli ve karlı yatırım alanlarına yönlendirdikleri için ekonomik kalkınmayı ya da büyümeyi hızlandırıcı bir işleve sahip olmaları yanında, sanayiye ucuz maliyetli fon sağlarken, tasarruf sahiplerine de yüksek kazanç sağlayabilmektedirler.

Tasarruf sahipleri ise sahip oldukları varlıkların toplam getirilerini, risk faktörlerini de dikkate alarak yüksek tutmayı amaçlar. Bu yolla tasarruf sahipleri, hisse senedi, tahvil gibi menkul kıymetlere yatırım yaparlar.

Tasarruf sahiplerinin birikimlerinin sermaye piyasalarında değerlendirilmesine başlanması ile birlikte, portföy ve portföy yönetimi ile ilgili konular önem kazanmaya başlamıştır [44, 45, 46].

3.1. Portföy Kavramı

Portföy, belirli amaçları gerçekleştirmek için yatırım yapılan, ağırlıklı olarak hisse senedi, tahvil gibi çeşitli menkul kıymetlerden oluşan ve bu menkul kıymetler arasında bir ilişki olduğundan, kendine öz, ölçülebilir nitelikleri olan finansal bir varlık grubudur.

Sermaye piyasalarında yatırım amacıyla, değişik menkul kıymetlerden veya yatırım araçlarından, çok sayıda portföy oluşturulabilir. Hisse senedi ve tahvil gibi geleneksel menkul kıymetler temel alındığında üç farklı portföyden söz edilebilir: Yalnız tahvillerden bir portföy oluşturulabileceği gibi, yalnız hisse senetlerinden bir portföy oluşturulabilir. Ancak, yatırımcıların en çok arzuladığı

portföyler karma portföylerdir. Bunun yanı sıra, diğer yatırım araçlarından oluşan portföylerden de söz etmek mümkündür [44].

Tamamen tahvillerden oluşan portföyler, anaparanın güvenliğini önde tutan, yani risk almayı sevmeyen, piyasayı izlemekte güçlük çeken tasarruf sahiplerinin tercih ettikleri portföylerdir. Değişik şirket ve devlet tahvilleri ile birlikte hazine bonolarından oluşturulan portföy, sahip olduğu düşük risk oranının yanında kısıtlı bir gelir sağlar. Sağlam bir yatırım türü olabileceği için, bu tür portföylerin ekonominin durgunluk dönemlerinde oluşturulmasında yarar vardır. Öte yandan, değişik vadelerden oluşan tahvil portföyü ile yatırımcı bazı risklerden de kendini korumuş olur.

Hisse senetleri, tahvil ve türev ürünlerden oluşan portföyler, en çok kullanılan portföy türüdür. Ekonominin içinde bulunduğu duruma göre, anapara, belli oranlarda hisse senedi, tahvil ve türev ürünler arasında paylaştırılır. Bu şekilde emniyet ve kârlılık unsurlarının birleştirilmesiyle, dengeli bir portföy oluşturulmaya çalışılır.

Bilindiği gibi, ekonomideki gelişmeler zaman içerisinde farklılık gösterir. Piyasadaki eğilim ve gelişmeler, zaman zaman hisse senetlerine veya tahvillere ağırlık verilmesini gerektirebilir. Sermaye piyasası, ekonominin aynası olduğundan, genelde ekonominin durgun olduğu dönemlerde, tahvil piyasasında bir canlanma, ekonominin canlandığı dönemlerde ise, hisse senedi piyasasında bir hareketlilik görülür. Bu özelliklerden dolayı, portföyü hisse senedi, tahvil ve türev ürünlerden oluşturan portföy sahipleri, portföylerini küçük operasyonlarla genel ekonomiye kolayca uyum sağlayacak şekilde düzenleyerek, zarar etmemeye çalışırlar.

Tamamı hisse senetlerinden oluşan portföylerin diğer portföy türlerinden ayrılan yanı, bir kaç farklı yatırım aracını değil de yalnızca hisse senetlerini kapsamasıdır. Hisse senetlerinden bir portföy oluşturmaya çalışan yatırımcının, elindeki seçeneklerinin sınırsız olduğu düşünülebilir.

Bu tür portföylerde her türlü risk düzeyine uygun yatırım yapmak mümkündür. Hisse senetlerinden oluşan portföylerde yatırımcı tipinin portföy oluşturma kararına büyük etkisi vardır [44].

3.2. Portföy Yönetimi

Dinamik ekonomilerde zamanla ekonomik koşullar değişebilmektedir; Menkul kıymetlere yatırım yapan yatırımcının portföyünden bazı kıymetleri çıkarıp satması, yerlerine yenilerini alması gerekmektedir. Portföy yönetimi ise yatırımcıların sahip oldukları fonları, mevcut menkul kıymetler arasında belirli bir risk düzeyinde ve en çok karlılığı veya belirli bir karlılığı en az riskle sağlayacak şekilde dağıtması faaliyetidir. Portföy yönetimi, yatırımcının sahip olduğu toplam menkul kıymetlerin seçimi ve her birinden ne miktarda portföye dahil edileceği konusunda belli yöntem ve teknikleri kapsar. Amaçlanan ise risk ve beklenen getiri karşılaştırması yaparak yatırımcının portföye yüksek getiri sağlayan menkul kıymetleri almak veya riskli olanları satmaktır [45, 46, 47].

Finansal açıdan risk, beklenen getirinin gerçekleşen getiriden sapma olasılığıdır. Başka bir deyişle, yatırımcının yapmış olduğu yatırımın sağlayacağı verimin, beklenen verimin altına düşme veya üstüne çıkma olasılığıdır. Bu olasılık, finansal açıdan, yatırımcının yapmış olduğu yatırımın riskini oluşturur.

Beklenen getiri ise, yatırım yapılan finansal varlıkların belli bir dönem getirileri ile bu getirilerin gerçekleşme olasılıkları çarpımının toplamıdır [44].

Portföy oluşturulması için piyasasının dikkatli bir şekilde izlenmesi ve istenildiği anda alım-satım yapabilme özeliğinin olması gerekir. Portföydeki senetlerden fiyatları düşen veya değişmeyenlerin satılmasında yarar vardır. Hisse senetlerinden oluşan portföyler, ekonominin istikrarlı olduğu dönemlerde başarıyla uygulanabilir. Bu portföy türünün kendine has özellikleri vardır. Bu tip portföye dahil edilecek hisse senetleri, *kısa vadede prim yapabilecek hisse senetleri* ve *uzun vadede prim yapabilecek hisse senetleri* olmak üzere iki grupta incelenebilir.

Bir portföy, hisse senedi ve tahvil gibi temel menkul kıymetler dışında, Varlığa Dayalı Menkul Kıymet, Finansman Bonoları, Hazine Bonosu, Gelir Ortaklığı Senetleri, Banka Bonoları, Mevduat, Repo, Döviz ve Döviz Tevdiat Hesapları gibi yatırım araçlarıyla da oluşturulabilir. Bu tür portföyler oluşturulurken, yatırım araçları arasında karşılaştırma yapılır. Yatırımın süresi boyunca hangi tür varlıkların daha verimli olacağı, beklenen değer, standart

sapma, korelasyon, kovaryans gibi çeşitli istatistikî tekniklerle hesaplanarak tahmin edilir [44]. Bu varlıklar seçilerek portföye dahil edilir.

Portföy yaklaşımları iki ana başlık altında toplanmaktadır: Geleneksel Portföy Yaklaşımı ve Modern Portföy Yaklaşımı.

Geleneksel portföy yaklaşımına göre, portföy yönetiminin kendine özgü kuralları vardır. Bunlar yatırımcı açısından önemlidir ve dikkatli bir çalışmayı gerektirir. Ancak, bu teorik araçları etkin bir biçimde kullanabilme yeteneği, sınırsız olarak kişiden kişiye değişen bilgi ve deneyimine bağlıdır. Bu nedenle, geleneksel portföy analizinin sezgi, içe doğuş gibi sübjektif yaklaşımlar içerdiği söylenebilir. Geleneksel yaklaşımın amacı, yatırımcının sağlayacağı faydayı maksimize etmektir. Başka bir deyişle, yatırımcının risk ve getiriye ilişkin fayda tercihlerini maksimize edecek bir portföyü seçtiği kabul edilmektedir. Diğer bir ifadeyle, ortaya çıkan risk düzeyine göre, yatırımcı belirlemiş olduğu faydayı maksimize etmeye çalışmaktadır.

Geleneksel portföy yaklaşımında, portföy getirisi, portföyü oluşturan menkul kıymetlerin temettü ve belli bir dönemdeki değer artışıdır. Bu nedenle, yatırımcıların portföyün etkinliğinin yükseltilebilmesi için gelecekteki menkul kıymet getirilerini tahmin etmeleri gerekmektedir. Öte yandan, çeşitli portföy getirilerine göre ortaya çıkabilecek riskler de hesaplanmalıdır. Portföy oluşturma'nın esas amacı, riskin dağıtılmasıdır. Portföyü oluşturan menkul kıymetlerin getirileri aynı yönde hareket etmeyeceğinden, portföyün riski tek bir menkul kıymetin riskinden küçük olacaktır.

Geleneksel portföy teorisi, riskin dağıtılması prensibinden hareket ederek, portföy içindeki menkul kıymet sayısının artırılmasına dayanır. Menkul kıymet portföylerinin yönetiminde geleneksel yaklaşım; yatırımcının amacının belirlenmesi, portföye alınacak menkul kıymetlerin seçimi ve portföyü yönetmek aşamalarını kapsar. Genel olarak yatırımcıların amaçları, sabit ve istikrarlı gelir elde etmek, sermayenin korunması veya sermaye kazancı elde etmek gibi ifade edilebilir.

Sermaye piyasasının gelişmiş olduğu ülkelerde, 1950'li yıllara gelinceye kadar, yatırımcılar, portföyde yer alan menkul kıymetlerin getirileri arasındaki

ilişkileri göz önünde bulundurmadan, sadece portföydeki menkul kıymetlerin sayılarını arttırarak riski azaltabileceklerini düşünmüşlerdir [44, 45].

Oysa, modern portföy yaklaşımında, sadece portföy çeşitlendirmesine gidilerek riskin azaltılamayacağı, çünkü, portföyde yer alan menkul kıymetlerin ya da menkul kıymet gruplarının, aynı ya da ters yönde hareket ettikleri ileri sürülmektedir.

Modern portföy teorisi ilk defa, Markowitz, 1952 yılında yayınladığı "portföy seçimi" başlıklı makalesinde, portföyde yer alan menkul kıymetlerin, belirli risk seviyelerinde mümkün olan maksimum getiri oranını nasıl sağlayabileceğini araştırması ile ortaya çıkmıştır. Markowitz, geleneksel portföy yönetimine üç önemli noktada katkıda bulunmuştur. Bunlardan birincisi ve en önemlisi, portföy yönetiminde, kısımların veya parçaların toplamının, bütüne eşit olmadığına ispatlanmasıdır. Markowitz, burada portföy riskinin portföyü oluşturan varlıkların riskinden daha az olabileceğini ve belirli koşullarda portföyün sistematik olmayan riskinin sıfır yapılabileceğini göstermiştir. İkincisi, yatırımcıların bazı portföyleri aynı getiriyi sağlamakla birlikte, daha riskli oldukları için, bazı portföyleri de aynı risk düzeyinde olmakla birlikte, daha az getiri sağladıkları için tercih etmeyeceklerini, dolayısıyla bazı portföylerin diğerlerine göre daha üstün olduklarını ve bu durumu üstünlük ilkesi olarak ileri sürmüştür. Markowitz'e göre, menkul kıymetlerin seçiminde etkin sınır söz konusudur. Üçüncü, önemli nokta, etkin sınırın kuadratik programlama yolu ile elde edilebileceğidir. Markowitz'in geliştirdiği yöntem, karmaşık bir takım hesaplamaları gerektirir [44, 46, 47].

1963 yılında Markowitz'in öğrencisi, William Sharpe, bu yöntemi geliştirerek, basit bir şekilde ortaya koymuştur. Tekli indeks modeli, olarak bilinen bu yöntem 1970'lerde paket program haline getirilerek, bilgisayar çözümlemelerinde kullanılmaya başlanmıştır. Tekli indeks modeli daha çok alternatif hisse senetlerine yapılan yatırımların getirilerinin maksimizasyonu için kullanılırken, Markowitz'in geliştirdiği model, tahvil, hisse senedi, gayrimenkul ve menkullere yapılan yatırımların analizinde kullanılmaktadır.

Portföy yönetimi konusunda, daha sonraları, Sharpe (1964), Lintner (1965) ve Mossin (1966), bütün tasarruf sahiplerinin modern portföy kuramına uygun

olarak, menkul kıymetlere ve özellikle hisse senetlerine yatırım yapmaları halinde, fiyatların ne yönde değişeceğini araştırmışlardır. Bu çalışmaların sonucunda (SVFM) Sermaye Varlıklarını Fiyatlama Modeli geliştirilmiştir. SVFM, yirmi yıla yakın bir süre, batıda portföy yönetiminde aracı kuruluşlar tarafından kullanıldıktan sonra, 1977'de Richard Roll modelin yetersizliğini ileri sürerek, portföy yönetiminde kullanılmamasını önermiştir.

Steve Ross, SVFM'e alternatif olarak Arbitraj Fiyatlama Kuramı (AFK) olarak bilinen modeli ortaya atmıştır. Bu modelde Ross, risk ve getiri arasında, hiç bir yatırımcıya arbitraj yolu ile sınırsız olarak servetini artırma olanağı vermeyecek bir ilişkiyi araştırmıştır. Günümüzde, Sermaye Varlıklarını Fiyatlandırma Modeli uygulamada geniş ölçüde kullanılırken, AFK henüz akademik çevrelerde tartışılmaktadır.

Portföy kuramında çok sayıda menkul kıymete yatırım yapılması durumunda, portföy riskinin ne olacağı hususu araştırılmaktadır. Portföy riski, getirilerdeki değişme yönünden ölçülür. Getirilerdeki değişme veya sapma, varyans veya standart sapma ile ifade edilir. Portföye hangi menkul kıymetlerin alınacağı konusu önem taşır. Portföye bir menkul kıymet eklendiğinde veya çıkarıldığında, portföy getirisi üzerindeki etkisi ne olur? Portföy riski artacak veya azalacak mıdır? Gelecek yıl portföy getirileri beklenen oranda gerçekleşecek mi? Bu ve benzeri soruların cevapları yatırımcıların bilmek istedikleri konulardır. Kısaca, yatırımcı bir grup, nasıl bir portföy oluşturmalıdır ki portföy getirisi en çok ve riski en az olsun. Portföy yatırımcıları ile ilgili söz konusu endişeler, modern portföy kuramı yardımıyla en aza indirilmeye çalışılmıştır [44].

Sermaye piyasalarında gerçekleştirilen portföy yatırımlarının yönetiminde, pasif portföy yönetim stratejisi ve aktif portföy yönetim stratejisi olmak üzere iki temel stratejiden bahsedilebilir. Portföyün yönetim stratejisinin belirlenmesinde, yatırımcının menkul kıymet fiyatlarına ilişkin düşüncelerine ve oluşturduğu portföyün getiri beklentisine büyük ölçüde bağlıdır. Seçilen yönetim stratejisi, portföy hakkında alınacak kararlar üzerinde etkili olacaktır [48].

3.2.1. Pasif Portföy Yönetim Stratejileri

Pasif yönetim stratejisi sermaye piyasalarının kısmen de olsa etkin olduğu varsayımına dayanır. Bu strateji piyasada düşük fiyatlandırılmış menkul

kıymetlerin belirlenerek kazanç elde edilmesinin mümkün olmadığı ve piyasa fiyatlarının gerçek değerlerine çok yakın olduğu düşüncesinden yola çıkar [48].

Pasif portföy yönetim stratejisi piyasada düşük fiyatlandırılmış menkul kıymetleri belirleyerek kazanç elde etmenin mümkün olmadığına ve piyasa fiyatlarının gerçek değerlere çok yakın olduğuna inanan yatırımcılar için uygun bir stratejidir. Pasif portföy yönetim stratejisi, seçilen bir endeksin performansını taklit etmek için portföy tasarlama girişimidir. Bir pasif portföy yöneticisi ücretini seçilen hisse senedi endeksinin performansını olabildiğince yakından izleyebilen bir portföyü oluşturarak kazanır. Hisse senedi endeksi müşterilerin ihtiyaçlarını, amaçlarını karşılamakta ve benchmark endeks olarak adlandırılmaktadır. Özellikle piyasa riski hesaplanırken hedef endeksin volatilitesi belirlendiğinden, pasif portföy yöneticilerinin risk yönetimi çerçevesinde gerçekleştirecekleri finansal risk ölçümlerinde daha az zorlukla karşılaşacağı söylenebilir.

Tek tek hisse senetlerinin fiyat hareketlerinden kazanç elde etmenin mümkün olmaması durumunda, piyasada işlem yapan yatırımcıların temel amacı bir bütün olarak piyasanın sağlayacağı kazanç seviyesinde bir getiri elde etmek olacaktır. Pasif yönetim stratejisi yoluyla gerçekleştirilmesi mümkün olan bu amaç, her bir hisse senedinin piyasa değeri oranında yer alacağı ve piyasa portföyünü yansıtır şekilde portföy oluşturmakla gerçekleştirilebilecektir [48].

3.2.2. Aktif Portföy Yönetim Stratejileri

Aktif yönetim stratejisi sermaye piyasalarının etkin olmadığı varsayımına dayanır. Bu strateji ise herhangi bir zamanda piyasada oluşmuş fiyatlardan hareketle ileride gerçekleşebilecek fiyat değişimlerinden kazanç elde edilebileceği düşüncesinden yola çıkar.

Aktif portföy yönetim stratejisinde, bir menkul kıymetin düşük fiyatlandırılmış olduğu beklentisi ve fiyat hareketlerinden kazanç elde edebileceğini düşüncesi söz konusudur. Bu nedenle, portföye ağırlıklı olarak düşük fiyatlandırıldığı beklentisine sahip hisse senetlerinin dahil edilmesi beklenir.

Aktif stratejilere göre yönetilen bir portföy, genellikle kısa dönemli piyasa hareketlerinin avantajlarından yararlanmak için oluşturulmaktadır. Bu nedenler varlıklar arasındaki göreceli değişikliklerden yararlanmak önemlidir [48].

Yatırımcının, bir menkul kıymetin düşük fiyatlandırılmış olduğuna inanması ve fiyat hareketlerinden kazanç elde edebileceğini düşünmesi halinde aktif portföy yönetimi söz konusudur. Aktif portföy yönelim stratejisini uygulayan yatırımcı için temel sorun, piyasada düşük fiyatlandırılmış olan hisse senetlerinin tespit edilmesidir. Bu ise fiyat/kazanç katsayısı, piyasa değeri/defter değeri oranı, süre, kâr payı verimi gibi farklı değerlendirme yöntemleri kullanılarak gerçekleştirilebilir.

Yatırımcının düşük fiyatlandırıldığına inandığı hisse senetlerini belirlemesi durumunda, oluşturacağı portföye ağırlıklı olarak bu hisse senetlerini dahil etmesi beklenir. Bu tür bir portföy oluşturma ise, piyasadan daha yüksek seviyede bir getiri elde etme amacındaki yatırımcının, pasif olarak yönetilen bir portföy yatırımcısına oranla daha fazla risk almasını gerektirir. Çünkü, portföye dahil edilen menkul kıymetlere ilişkin değerlendirme yatırımcının kendi subjektif değerlendirmesi olup, söz konusu değerlendirmelerin gerçekleşmeme olasılığı bulunmaktadır. Sonuçta, pasif olarak yönetilen portföylerin getirisinin üstünde bir getiriyi hedefleyen aktif portföy yöneticisinin üstlendiği risk seviyesi de, risk ve getiri arasındaki temel ilişkiye uygun olarak artmaktadır.

Yöneticinin piyasaya girişteki zamanlaması büyük ölçüde yöneticinin tecrübe ve yeteneklerine bağlı olup, başarıyı belirleyen temel faktör portföye dahil edilecek menkul kıymetlerin doğru olarak seçilmesidir. Bu ise, menkul kıymet seçimine ilişkin uygulanan tekniklerden bağımsız değildir.

Aktif yöneticiler iki zorluğun üstesinden gelmelidirler. Birincisi yüksek işlem maliyetlerini karşılamak, ikincisi ise benchmarka göre daha yüksek riski karşılayabilecek şekilde portföyü yönetebilmektir. Çünkü kısa süreli fiyat hareketleri, artan komisyon ödemeleri nedeniyle küçük kârlar ile sonuçlanacaktır. Bir portföyün aktif yönetimi önceden tanımlanan risk limitleri içinde uygun olmalıdır. Bunlar tamamen yönetilmeli ya da riske maruz kalan bazı finansal varlıklar ve getiri bölümleri yönetilebilmelidir. Menkul kıymet piyasalarında aktif olarak yönetilen fakat farklı seçim yöntemleri kullanılarak oluşturulan çok sayıda portföy mevcuttur.

Bu çalışmanın konusunu oluşturan uygulamalar, bir aktif portföy yönetim stratejisi olan ve ilgili bölümde ayrıntılarına yer verilen Eş İşlem Stratejisi olarak adlandırılmaktadır [48].

3.3. Eş İşlem Stratejisi

İstatistiksel arbitraj'ın özel bir formu olarak da kabul edilen eş işlem (pair trading) ise piyasadan bağımsız bir pozisyon sağlarken yatırımcının (trader) anormallikleri, nispi gücü veya iki hisse senedi veya sepet arasındaki temel farklılıkları bile yakalamasına izin veren bir stratejidir [49, 50].

Eş işlemin amacı, geçmiş dönemde fiyatları birlikte hareket eden, ilişkili hisse senetlerinin bulunması, söz konusu ilişkilerin sapma zamanlarından faydalanarak, bir hisse senedinin diğerine göre ucuz veya pahalı görüldüğü bir durumdan kar etmeye çalışmaktır. Uzun pozisyonun kısa pozisyona göre yükseleceğini tahmin ederek, diğerine göre ucuz olan hisse senedini alınır ve pahalı olan hisse senedi satılır [50, 51]. Sermaye piyasalarında kısa pozisyon, sahip olunmayan menkul kıymetin ödünç alınmak sureti ile satılması olarak tanımlanan açığa satış işlemini ifade etmektedir. Kısa pozisyonda, ödünç alınan menkul kıymetin fiyat azalışından kar elde edilmesi amaçlanır. Uzun pozisyon ise kısa pozisyonun karşılığıdır. Kısa pozisyonun aksine, uzun pozisyonda menkul kıymete sahip olunması söz konusudur. Bu durumda, uzun pozisyonda, menkul kıymetin fiyat artışından kar elde edilmesi amaçlanır [53, 54]. Piyasa aşağı giderse, kısa pozisyon kar yapmalıdır. Tersisi durumda ise uzun pozisyon kar etmelidir. Hisse senedi çiftini oluşturan iki hisse senedi yukarı doğru da hareket edebilir, fakat uzun pozisyondaki hisse senedi kısa pozisyondaki hisse senedinden daha yukarı ve daha hızlı hareket edecektir. İki hisse senedi aşağı doğru da hareket edebilir, bu durumda ise kısa pozisyondaki hisse senedi, uzun pozisyondaki hisse senedinden daha fazla ve hızlı düşecektir. Böylece hisse senedi çiftinin fiyat farkından kar edilecektir [49, 51, 52].

Eş işlem stratejisinin bir avantajı da başlangıç sermayesine ihtiyaç duyulmamasıdır. Uygulamada, eş işlem strateji kullanılarak oluşturulan ve yönetilen portföyün klasik anlayış ile getirisinin ifade edilebilirliği sınırlılık arz etmektedir. Geleneksel yatırım yönetiminde bir başlangıç sermayesi ve bunun bir yüzdesi olarak ifade edilen bir getiriden bahsetmek mümkündür. Oysa finansal

piyasaların gelişen mimarisi günümüz yatırımcılarına kaldıraç imkanı sunarak, başlangıç sermayesinin miktarını minimumda sabitleme olanağı tanımaktadır.

Uygulamada ele alınan eş işlem stratejisi, bu konuda yeni bir aşamayı temsil eden niteliklere sahiptir. Bu nitelikler aynı anda farklı hisse senetlerinde eşit tutarlı ters pozisyonlar oluşturarak sermaye gereksinimini ortadan kaldırmaktadır. Bu yapı sayesinde eş işlem stratejisini uygulayan bir yatırımcı sıfır başlangıç sermayesi ile yatırım yapabilme ayrıcalığına sahip olabilmektedir. Detaylarına ilgili bölümde yer verilen stratejinin çıktıları, diğer bir ifadeyle dönem sonu getirisini (overall), aynı zamanda yatırımın mutlak getirisini de temsil etmektedir.

4. İSTANBUL MENKUL KIYMETLER BORSASI'NDA EŞ İŞLEM STRATEJİSİ İLE VERİ MADENCİLİĞİ UYGULAMASI

Bu çalışmada İstanbul Menkul Kıymetler Borsasında işlem gören hisse senetleri için bir uygulama gerçekleştirilmiştir. Hisse senetleri iki sektör göz önüne alınarak veri madenciliği felsefesi yardımıyla eş işlem stratejisi kullanılarak çözümlenmiştir. Bu bölümde çalışmanın ana detayları ve sonuçlar sunulmaktadır.

4.1. Araştırmanın Amacı ve Önemi

Tasarruflarını sermaye piyasalarında değerlendirmek isteyen yatırımcılar, sahip oldukları menkul kıymetlerin getirilerinin, risk faktörünü de dikkate alarak yüksek olmasını amaçlar. Yatırım yapılan bu menkul kıymetlerden oluşan portföyün, yatırımcının yüksek getiri beklentisinin karşılanabilmesi için, dinamik ekonomilerde zamanla değişebilen ekonomik koşullara göre yönetilmesi gerekir. Portföy yönetimi ise, yatırımcıların sahip olduğu birikimleri, mevcut menkul kıymetler arasında belirli bir risk düzeyinde, en çok karlılığı sağlayacak şekilde dağıtması için, portföye hangi menkul kıymetlerin, ne kadar gireceğine ve zamanla değişen ekonomik koşullara göre hangi varlıkların çıkacağına karar vermeyi amaçlar.

Portföy yönetiminin amacına ulaşabilmesi için çeşitli kaynaklardan gelen verilerin bir araya getirilerek derlenmesi, analiz edilmesi ve elde edilen sonuçların karar vericilere sunulması gerekir. Ancak bu sürecin, sürekli değişen ekonomik koşullarda, zaman kaybetmeden hızlı bir şekilde gerçekleştirilmesi gerekir. Tersı durumda alınacak kararların gerekenden erken veya geç alınarak değerlerini yitirmelerine ve bunun sonucu olarak yatırımcıların elde edilmesi gerekenden daha az getiri elde etmelerinin yanında hiç getiri elde edememelerine de neden olabilir.

Bu çalışmada, eş işlem stratejisi kullanarak, İstanbul Menkul Kıymetler Borsası'nda işlem gören hisse senetlerinden, portföy oluşturmak amacıyla hisse senedi seçimi yapan, seçilen hisse senetleri üzerinde işlem yapmak amacıyla sinyal veren bir veri madenciliği sistemi oluşturmak amaçlanmıştır. Araştırmanın önemi; yatırımcıya yığın olarak gelen sektör verilerinin bir otomasyona bağlı olarak değerlendirilerek doğru zamanda, doğru hisse senedi seçimine yardımcı

olabilmesidir. Bu yapısı ile sektörel gelişmelerden haberdar olan bir yatırımcı hem elindeki bu bilgiler hem de bu çalışmada önerilen teknikleri kullanarak karlı bir duruma geçebilmektedir.

4.2. Yöntem

4.2.1. Araştırmanın Modeli

Bu araştırma, portföy oluşturmak ve işlem zamanının belirlenmesi için, eş işlem stratejisi kullanımına yönelik betimsel bir araştırmadır. Eş işlem stratejisinin uygulamasında izlenen adımlar aşağıdaki gibi sıralanabilir: İlk adımda aynı sektörde yer alan hisse senetlerinden uzun dönemde (altı ay – 1 yıl) yüksek ilişki gösteren, kısa dönemde (1 – 2 hafta) ise düşük ilişki gösterenler hisse senetleri bir çift olarak eşlenir ve hisse senetlerinin kapanış fiyat oranı hesaplanır. Burada uzun dönemde yüksek ilişki gösteren hisse senetlerinin seçilmesinin sebebi, bu ilişkinin uzun dönem ilişki çizgisinden uzaklaştığında bu durumdan faydalanılabilmesi için daha tutarlı bir ilişkiye sahip olmalarının beklenmesidir. Kısa dönemde düşük ilişki göstermeleri ise, çifti oluşturan hisse senetlerinin tamamen rassal olarak dalgalanmalarının beklenmesidir.

İkinci adımda, eşlenen hisse senetlerinin güncel fiyat oranlarının uzun dönem ortalamasından $\pm 2\sigma$ uzaklıkta olup olmadığı kontrol edilir. Fiyat oranları $\pm 2\sigma$ uzaklığa yaklaştığında, fiyat oranını oluşturan hisse senetlerinde işlem yapmak için harekete geçilir. Burada beklenti, geçmişte yüksek ilişki gösteren iki hisse senedinin gelecekte de aynı durumu devam ettirmesidir. Eşlerin kapanış fiyatındaki yayılmanın, yani fiyat farkının artmasından bir süre sonra azalması beklentisinden dolayı iyi performans gösteren hisse senedinde kısa pozisyon almak (açığa satış) ve düşük performans gösteren hisse senedinde uzun pozisyon almaktır (satın alma) [51, 55].

Bu çalışmada, niceliksel metot olarak, ikinci el veri derleme yönteminden yararlanılmıştır.

4.2.2. Araştırmanın Evreni ve Örneklemi

Bu araştırmada evren; İstanbul Menkul Kıymetler Borsası'nda işlem gören hisse senetleri olarak ele alınmıştır. İstanbul Menkul Kıymetler Borsasında Ekim

2007 tarihi ile yaklaşık 300 adet hisse senedi bulunmaktadır. Bu evren sektörel bazda alt gruplara bölünmüştür. Bu guruplardan Bankacılık ve Tekstil sektöründe yer alan hisse senetleri örneklem olarak ele alınmıştır.

Bu çalışmaya Bankacılık ve Tekstil sektörlerinin seçilmesinin sebebi bazı özelliklere sahip olmalıdır. Bu özelliklerin arasında, bu sektörlerde diğer sektörlerle göre daha fazla işletmenin yer alması, bu işletmelerin veri setinin derinliğe, başka bir ifadeyle geçmişe dönük sürekliliğe sahip olmalarının yanında İMKB'nin toplam piyasa kapitalizasyonunda büyük paya sahip olmaları ön plana çıkmaktadır.

Örneklem bu çalışmada Bankacılık sektörü için 12 ve Tekstil sektörü için 31 hisse senedi olmak üzere toplam 43 hisse senedini içermektedir. Sektörlere göre bu çalışmaya dahil edilen hisse senetleri sırasıyla AKBNK, ALNTF, DENIZ, FINBN, GARANTI, ISBTR, ISCTR, SKBNK, TEBNK, TEKST, TKBNK, YKBNK, AKALT, AKIPD, ALTIN, ARAT, ARSAN, ATEKS, BERDN, BISAS, BOSSA, CEYLN, CYTAS, DERIM, DESA, EDIP, ESEMS, GEDIZ, IDAS, KORDS, KOTKS, KRTEK, LUKSK, MEMSA, MNDRS, MTEKS, OKANT, SKTAS, SONME, UKIM, VAKKO, YATAS ve YUNSA hisse senetleridir.

4.2.3. Araştırmanın Değişkenleri ve Veri Derleme Yöntemi

Bu çalışmada veri derleme için yayınlanmış istatistiklerin kullanılması yöntemi izlenmiştir. Veri, İstanbul Menkul Kıymetler Borsası tarafından veri dağıtım için yetkili kılınan PLATO veri dağıtım hizmetleri A.Ş.'in internet sitesinden elde edilmiştir [56]. Verinin internet ortamına bahsi geçen firma tarafından eksiksiz ve doğru olarak sunulduğu varsayılmaktadır. Veri için kullanılan değişken hisse senetlerinin gün sonu kapanış fiyatlarıdır.

Çalışmanın istatistiksel analizinde kullanılan veri için tanımlanan zaman aralığı 3 Ocak 2005 ile 26 Eylül 2006 tarihleri olarak tespit edilmiştir. Veri setinin 3 Ocak 2005 ile 22 Şubat 2006 tarihleri arasında bulunan kısmı eğitim verisi olarak kullanılmıştır. 23 Şubat 2006 ile 29 Eylül 2006 tarihleri arasında yer alan veri ise test verisi olarak kullanılmıştır.

4.3. Uygulama

Çalışmanın bu aşamasında, veriler Plato Veri Dağıtım Hizmetleri A.Ş.'in internet sitesinden elde edilmiş ve veri madenciliğinde kullanılabilecek şekilde veri matrisine dönüştürülmüştür. Veri matrisleri, Çizelge 4.1'deki gibi, işlem tarihi ve ilgili sektörde yer alan hisse senetlerinin işlem tarihine ait gün sonu kapanış fiyatlarını içermektedir.

Çizelge 4.1 Bankacılık ve Tekstil Sektörü Veri Matrisleri

TARİH	AKBNK	ALNTF	DENİZ	FINBN	GARAN	ISBTR	ISCTR	SKBNK	TEBNK	TEKST	TKBNK	YKBNK
01.03.2005	8,40	1,00	3,22	2,52	4,34	1290,00	7,65	3,04	8,10	0,66	4,58	4,30
01.04.2005	8,05	1,02	3,38	2,56	4,46	1270,00	7,45	2,98	7,95	0,66	4,52	4,24
01.05.2005	7,70	0,98	3,22	2,46	4,46	1260,00	7,25	2,86	7,70	0,71	4,50	4,16
01.06.2005	7,80	0,99	3,26	2,43	4,42	1260,00	7,30	2,84	7,75	0,70	4,58	4,16
01.07.2005	8,20	1,03	3,26	2,50	4,72	1280,00	7,60	3,20	7,80	0,72	4,92	4,34

a) Bankacılık Sektörü Veri Matrisi Örneği

TARİH	AKALT	AKIPD	ALTIN	ARAT	ARSAN	...	SKTAS	SONMEZ	UKIM	VAKKO	YATAS	YUNSA
01.03.2005	6,35	4,3	1,25	0,72	1,63	...	16,9	1,27	3,84	2,52	1,21	1,47
01.04.2005	6,3	4,36	1,27	0,72	1,63	...	17,1	1,25	3,82	2,46	1,22	1,47
01.05.2005	6,25	4,24	1,26	0,71	1,6	...	16,7	1,26	3,78	2,43	1,18	1,46
01.06.2005	6,65	4,22	1,26	0,76	1,63	...	16,9	1,29	3,78	2,45	1,19	1,46
01.07.2005	6,45	4,14	1,26	0,75	1,62	...	16,9	1,27	3,82	2,46	1,19	Na

b) Tekstil Sektörü Veri Matrisi Örneği

Sektörlerden oluşturulacak hisse senedi çiftlerinin belirlenmesinde kullanılan ilişki matrisinin ve fiyat oranlarının hesaplanabilmesi için, hisse senedi fiyatlarında var olan kayıp verinin giderilmesi gereklidir. Kayıp veri, zaman serileri gibi, kesikli noktalar kümesinden örneklenmiş verinin interpolasyonu ve yaklaşımı için geniş ölçüde kullanılmakta olan spline yöntemi kullanılarak doldurulmuştur [57]. Doldurma işlemi için, R for Windows uygulamasının *spline()* fonksiyonu kullanılarak, kübik spline yönteminden faydalanılmıştır [58].

Bu aşamanın sonunda, veri madenciliği uygulaması için, elde edilen veri, uygun bir şekilde derlenmiş, var olan kayıp veri uygun yöntemle doldurulmuş ve analize hazır hale gelmiştir.

Veri madenciliğinin bir sonraki aşaması olan analiz aşamasında, ilişki katsayısı yardımıyla, belirlenen ölçütlere göre hisse senedi çiftleri oluşturulacaktır. Elde edilen hisse senedi çiftlerinin kapanış fiyat oranları, oran

ortalamaları ve sapma limitleri hesaplanacak ve kapanış fiyat oranının limitleri aşp aşmadığı kontrol edilecektir.

4.3.1. Bankacılık Sektörü için Hisse Senedi Çiftlerinin Belirlenmesi

İstanbul Menkul Kıymetler Borsası'nda işlem gören bankacılık sektörüne ait hisse senetlerinden AKBNK, ALNTF, DENİZ, FINBN, GARAN, ISBTR, ISCTR, SKBNK, TEBNK, TEKST, TKBNK ve YKBK için uzun dönemde yüksek ilişki, kısa dönemde ise düşük ilişki gösteren hisse senedi çiftlerini belirlemek amacıyla, kısa dönem ve uzun dönem kapanış fiyatları ilişki matrisi Çizelge 4.2 ve Çizelge 4.3'de sırasıyla verilmiştir.

Çizelge 4.2 Bankacılık Sektörü Kısa Dönem Kapanış Fiyatı İlişki Matrisi

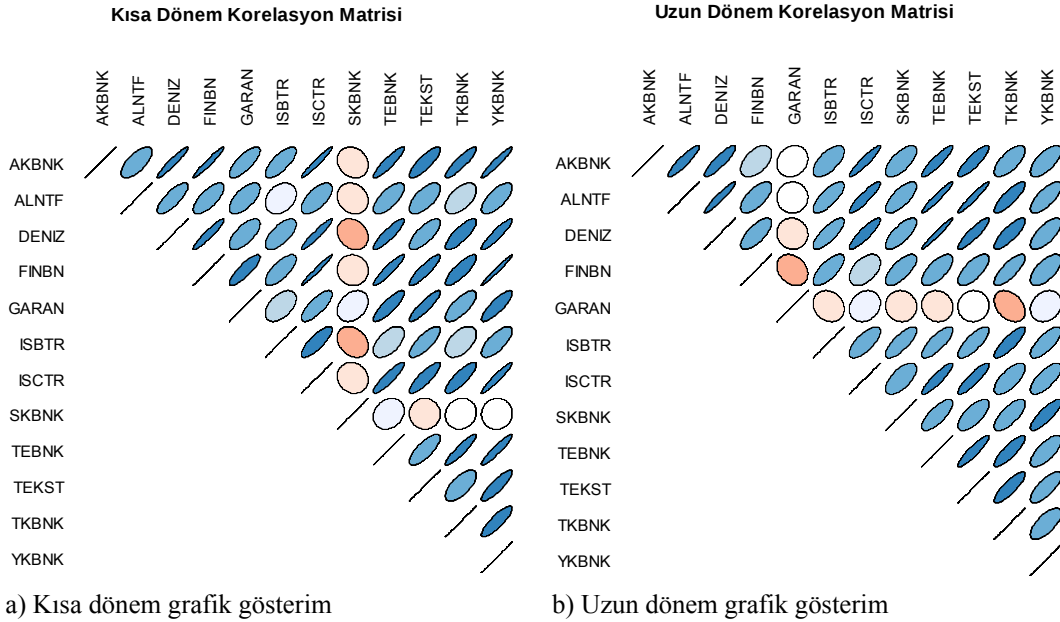
	AKBNK	ALNTF	DENİZ	FINBN	GARAN	ISBTR	ISCTR	SKBNK	TEBNK	TEKST	TKBNK	YKBK
AKBNK	1,000	0,688	0,951	0,965	0,768	0,745	0,967	-0,178	0,906	0,828	0,888	0,954
ALNTF	0,688	1,000	0,793	0,715	0,651	0,296	0,649	-0,172	0,700	0,648	0,555	0,705
DENİZ	0,951	0,793	1,000	0,940	0,678	0,675	0,949	-0,217	0,892	0,751	0,811	0,915
FINBN	0,965	0,715	0,940	1,000	0,851	0,739	0,980	-0,038	0,926	0,888	0,830	0,983
GARAN	0,768	0,651	0,678	0,851	1,000	0,494	0,773	0,313	0,850	0,891	0,737	0,848
ISBTR	0,745	0,296	0,675	0,739	0,494	1,000	0,813	-0,282	0,550	0,786	0,513	0,724
ISCTR	0,967	0,649	0,949	0,980	0,773	0,813	1,000	-0,112	0,897	0,862	0,833	0,954
SKBNK	-0,178	-0,172	-0,217	-0,038	0,313	-0,282	-0,112	1,000	0,209	-0,011	0,101	0,031
TEBNK	0,906	0,700	0,892	0,926	0,850	0,550	0,897	0,209	1,000	0,770	0,920	0,937
TEKST	0,828	0,648	0,751	0,888	0,891	0,786	0,862	-0,011	0,770	1,000	0,668	0,870
TKBNK	0,888	0,555	0,811	0,830	0,737	0,513	0,833	0,101	0,920	0,668	1,000	0,853
YKBK	0,954	0,705	0,915	0,983	0,848	0,724	0,954	0,031	0,937	0,870	0,853	1,000

Çizelge 4.3 Bankacılık Sektörü Uzun Dönem Kapanış Fiyatı İlişki Matrisi

	AKBNK	ALNTF	DENİZ	FINBN	GARAN	ISBTR	ISCTR	SKBNK	TEBNK	TEKST	TKBNK	YKBK
AKBNK	1,000	0,909	0,882	0,508	0,167	0,651	0,933	0,677	0,904	0,862	0,695	0,682
ALNTF	0,909	1,000	0,949	0,674	0,008	0,760	0,903	0,730	0,957	0,961	0,843	0,733
DENİZ	0,882	0,949	1,000	0,714	-0,047	0,788	0,890	0,749	0,976	0,895	0,820	0,707
FINBN	0,508	0,674	0,714	1,000	-0,259	0,718	0,473	0,700	0,682	0,703	0,788	0,644
GARAN	0,167	0,008	-0,047	-0,259	1,000	-0,177	0,201	-0,046	-0,084	0,035	-0,366	0,204
ISBTR	0,651	0,760	0,788	0,718	-0,177	1,000	0,715	0,779	0,793	0,788	0,852	0,746
ISCTR	0,933	0,903	0,890	0,473	0,201	0,715	1,000	0,637	0,903	0,880	0,706	0,722
SKBNK	0,677	0,730	0,749	0,700	-0,046	0,779	0,637	1,000	0,756	0,698	0,724	0,835
TEBNK	0,904	0,957	0,976	0,682	-0,084	0,793	0,903	0,756	1,000	0,912	0,855	0,717
TEKST	0,862	0,961	0,895	0,703	0,035	0,788	0,880	0,698	0,912	1,000	0,847	0,760
TKBNK	0,695	0,843	0,820	0,788	-0,366	0,852	0,706	0,724	0,855	0,847	1,000	0,682
YKBK	0,682	0,733	0,707	0,644	0,204	0,746	0,722	0,835	0,717	0,760	0,682	1,000

İlişki matrislerinin daha rahat takip edilebilmesi amacıyla aşağıda bunların grafiksel gösterimleri de sunulmuştur. Grafiksel gösterim için R paket programının *ellipse* paketinde yer alan *plotcorr* fonksiyonundan yararlanılmıştır. Bu fonksiyon, ilişki matrisinde yer alan her bir ilişki katsayısını, sağlanması durumunda renk kodlamasını da kullanarak, elips biçimli şekiller kullanarak ilişki

matrisinin grafiksel gösterimini sağlar. Grafiksel gösterimde her bir ilişki katsayısını temsil eden şekil daireden elipse doğru değişim gösterdikçe ilişkinin derecesini, elipsin yönü ise ilişkinin yönünü gösterir. Grafiksel gösterimde, sola yatık elips negatif ilişki katsayısını, daire sıfıra çok yakın bir ilişki katsayısını ve sağa yatık bir elips ise pozitif bir ilişki katsayısını temsil eder. Bu gösterimde, sola veya sağa yatık doğru ise tam ilişkiyi yani -1 veya +1 değerini almış bir ilişki katsayısını temsil eder. Yine bu fonksiyon, grafiksel gösteriminin daha kolay okunabilirliğinin sağlamak amacıyla, ilişki matrisinin öğelerini en yakın tam sayıya yuvarlayarak sayısal olarak gösterimini de sağlar [59]. Bankacılık sektörüne ait ilişki matrislerinin grafiksel gösterimleri Şekil 4.1a ve b ve Şekil 4.2a ve b de verilmiştir.



Şekil 4.1 Bankacılık sektörü için a) Kısa dönem ve b) Uzun dönem ilişki matrislerinin grafiksel gösterimler

Kısa Dönem Korelasyon Matrisi													Uzun Dönem Korelasyon Matrisi												
	AKBNK	ALNTF	DENİZ	FINBN	GARAN	ISBTR	ISCTR	SKBNK	TEBNK	TEKST	TKBNK	YKBNK		AKBNK	ALNTF	DENİZ	FINBN	GARAN	ISBTR	ISCTR	SKBNK	TEBNK	TEKST	TKBNK	YKBNK
AKBNK	10	7	10	10	8	7	10	-2	9	8	9	10	AKBNK	10	9	9	5	2	7	9	7	9	9	7	7
ALNTF		10	8	7	7	3	6	-2	7	6	6	7	ALNTF		10	9	7	0	8	9	7	10	10	8	7
DENİZ			10	9	7	7	9	-2	9	8	8	9	DENİZ			10	7	0	8	9	7	10	9	8	7
FINBN				10	9	7	10	0	9	9	8	10	FINBN				10	-3	7	5	7	7	7	8	6
GARAN					10	5	8	3	9	9	7	8	GARAN					10	-2	2	0	-1	0	-4	2
ISBTR						10	8	-3	6	8	5	7	ISBTR						10	7	8	8	8	9	7
ISCTR							10	-1	9	9	8	10	ISCTR							10	6	9	9	7	7
SKBNK								10	2	0	1	0	SKBNK								10	8	7	7	8
TEBNK									10	8	9	9	TEBNK									10	9	9	7
TEKST										10	7	9	TEKST										10	8	8
TKBNK											10	9	TKBNK											10	7
YKBNK												10	YKBNK												10

a) Kısa dönem sayısal grafik gösterim

b) Uzun dönem sayısal grafik gösterim

Şekil 4.2 Bankacılık Sektörü için a) Kısa dönem ve b) Uzun dönem ilişki matrislerinin sayısal grafik gösterimler

Şekil 4.2a da kısa dönemde düşük ilişki gösteren ve Şekil 4.2b de uzun dönem yüksek ilişki gösteren hisse senetleri eşleştirilmiştir. Elde edilen hisse senedi çiftleri portföye alınmış ve belirlenen $\mp 2\sigma$ kriterine göre karşılıklı olarak işlem yapılmıştır. Bankacılık sektöründe oluşan hisse senedi çiftleri Çizelge 4.4'teki gibidir.

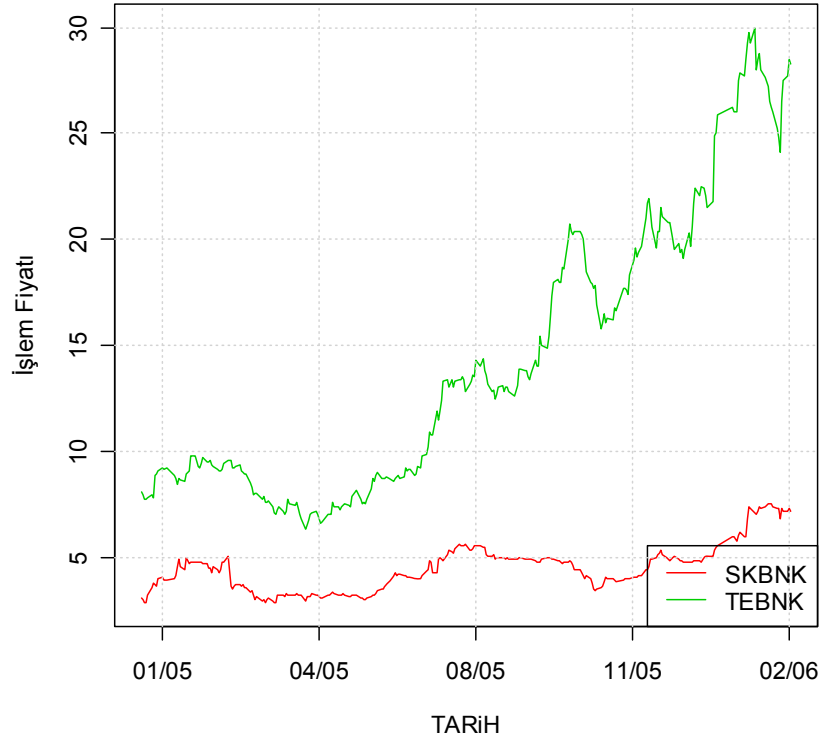
Çizelge 4.4. Bankacılık Sektörü Uzun Dönem – Kısa Dönem Kapanış Fiyatı İlişkilerine Göre Eşleşmeler

	Kısa Dönem İlişkisi	Uzun Dönem İlişkisi
	TEBNK	
SKBNK	0.209009482290152	0.756111183306442
	YKBNK	
SKBNK	0.0306167083003155	0.835317636652844

Oluşan hisse senedi çiftlerinden SKBNK/YKBNK kapanış fiyat oran grafiği ve kapanış fiyat grafiği Şekil 4.3 ve Şekil 4.4'teki gibidir. Eş işlem stratejisinin uygulama şekline göre bankacılık sektöründen elde edilen ilk hisse senedi çifti olan SKBNK/TEBNK hisse senedinin kapanış fiyatı oranı 27.01.2005 tarihinde $+2\sigma$ limitini aşmıştır.



Şekil 4.3. SKBNK/TEBNK hisse senedi çiftinin kapanış fiyat oran grafiği



Şekil 4.4. SKBNK/TEBNK hisse senedi çiftinin kapanış fiyatları grafiği

Bu noktadan itibaren, TEBNK hisse senedinin kapanış fiyatı SKBNK hisse senedinin kapanış fiyatından daha hızlı düşüş göstermiş ve fiyat aralığını

daraltmıştır. Kapanış fiyatları arasındaki bu daralma, kapanış fiyat oranının uzun dönem oran ortalamasına yaklaşmasına kadar sürmüştür. Kapanış fiyatı oranının, oran ortalamasını kestiği 05.09.2005 tarihinde, iki hisse senedi arasındaki fiyat farkı daralmıştır. Bu hisse senedi çiftinin kapanış fiyat oranı ortalama oranı keserek -2σ limitine doğru hareketlendiğinde de, iki hisse senedinin kapanış fiyatları yukarı doğru hareketlenmiş, ancak TEBNK hisse senedinin kapanış fiyatı SKBNK hisse senedinin kapanış fiyatından daha hızlı artış göstermiştir. Bu nedenle yine kapanış fiyatları arasındaki fark büyümüştür. Eş işlem stratejisinin amacı, eşleştirilen bu hisse senedi çiftlerinin fiyat hareketlerinden doğan fiyat farklarından yararlanmaktır. Fiyat farklarını yakalamak amacıyla da $\mp 2\sigma$ limitlerinin aşılması ölçütü dikkate alınmaktadır.

27.01.2005 tarihinde $+2\sigma$ limitini aşan SKBNK/TEBNK hisse senedi çiftinde, TEBNK hisse senedinin performansının, SKBNK hisse senedinden kötü olması nedeniyle, TEBNK hisse senedinde, yükseleceği tahminiyle uzun pozisyon, TEBNK hisse senedine göre daha iyi performans göstermesinden dolayı, düşeceği tahminiyle, SKBNK hisse senedinde kısa pozisyon alınacaktır. Kapanış fiyatı oranı, 05.09.2005 tarihinde oran ortalamasına döndüğünden hisse senedi çiftinde açılmış olan pozisyon kapatılacaktır.

Bu hisse senedi çiftinin yapılan alım – satım işlemleri sonucunda edinilen getirisi aşağıdaki çizelgedeki gibidir.

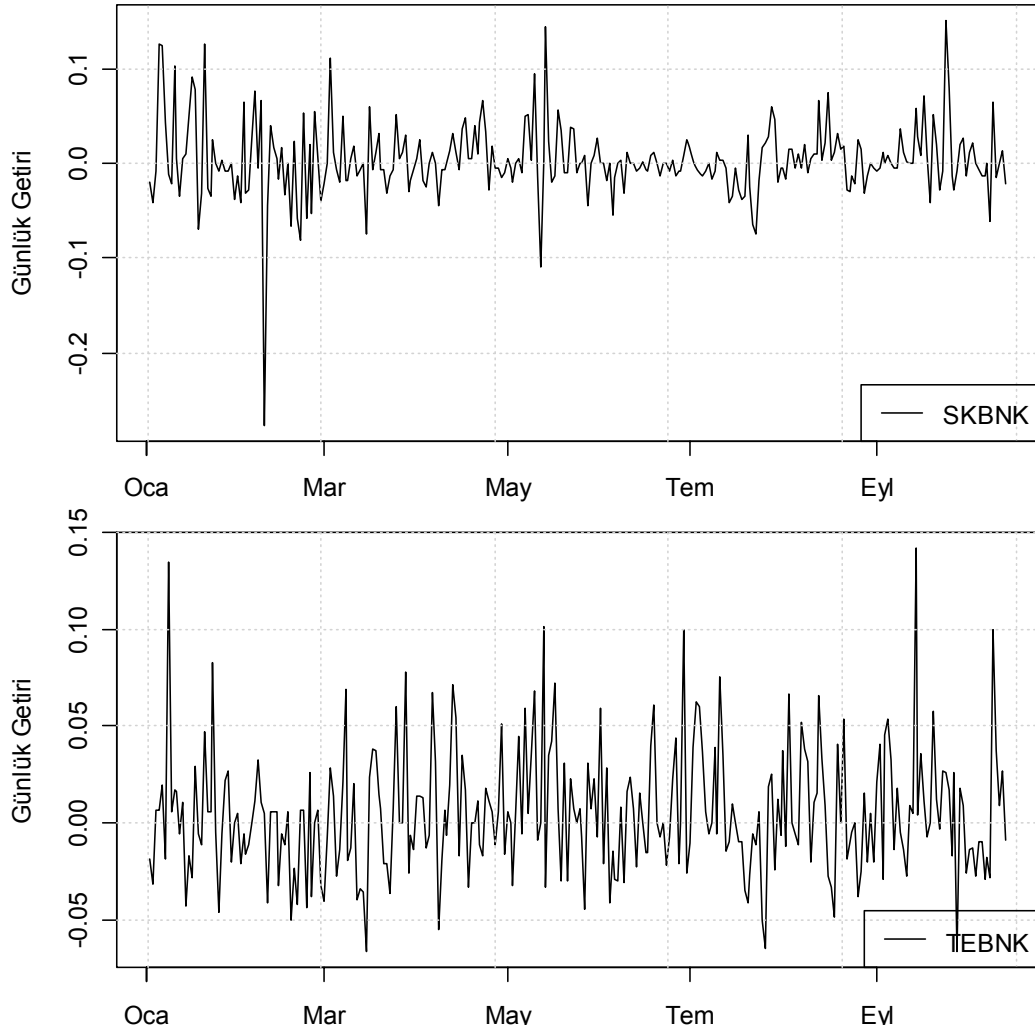
Çizelge 4.5. SKBNK/TEBNK çiftinin getiri çizelgesi

İşlem Tarihi		Hisse Senetleri	Tutar (YTL)		Adet		Getiri (YTL)	
Pozisyon Açma	27.01.2005	SKBNK/TEBNK	1.000	1.000	-204	+114		
Pozisyon Kapatma	05.09.2005		1.004	1.586			-4	+586

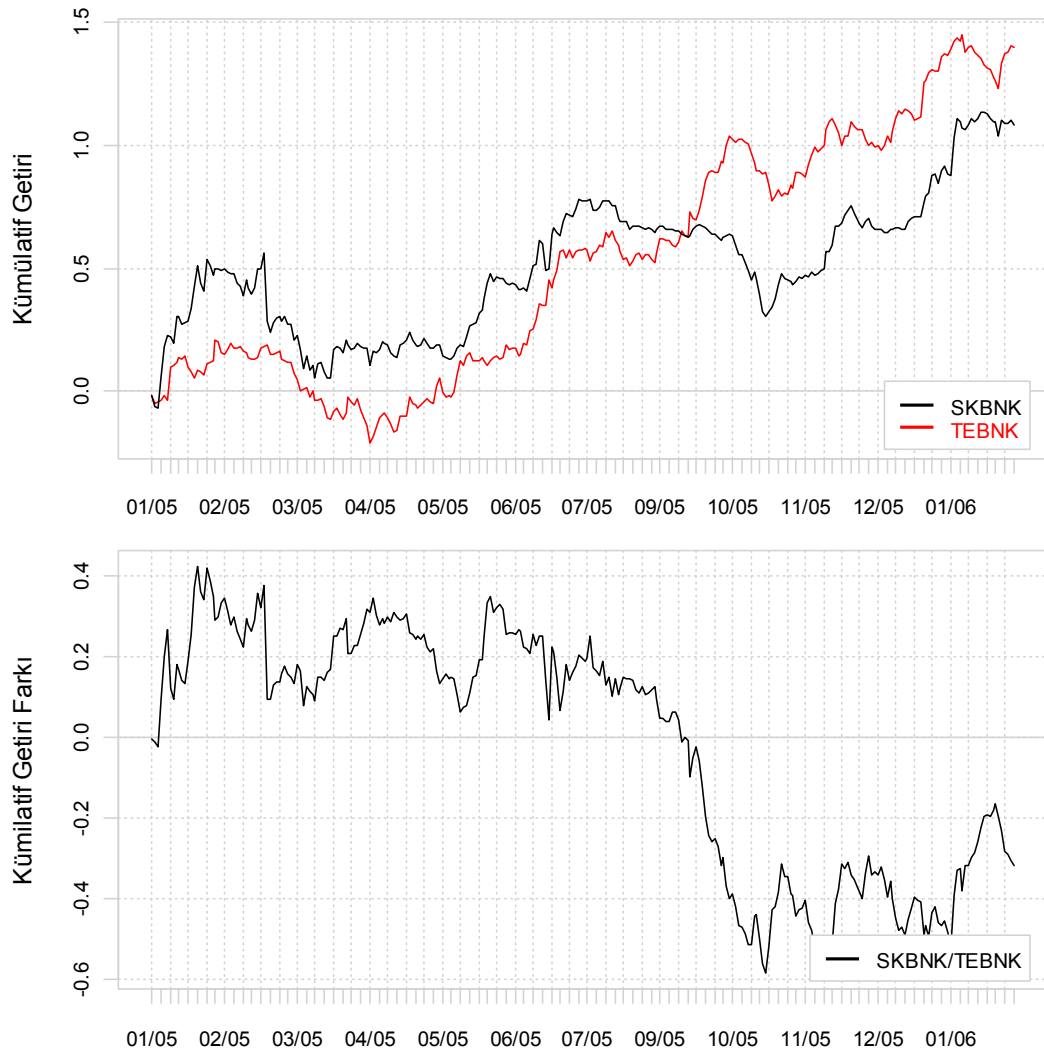
Çizelge 4.5'e göre, 27.01.2005 tarihinde SKBNK hisse senedinde kısa pozisyon alınmıştır. Buna göre, 1.000 YTL karşılığı 204 adet hisse senedinin açığa satışı yapılmış, yine 1.000 YTL karşılığı 114 adet TEBNK hisse senedi satın alınmıştır. Bu pozisyon 05.09.2005 tarihinde, açığa satışı yapılan 204 adet SKBNK hisse senedinin ilgili tarihteki kapanış fiyatından geri satın alınmış, satın alınan 114 adet TEBNK hisse senedi ilgili tarihteki kapanış fiyatından satılmıştır.

Yapılan işlem sonucunda, açılan pozisyonun kapatıldığı tarihte, açığa satışı yapılan SKBNK hisse senedinin geri alım fiyatı satış fiyatının üstünde olması

nedeni ile 4YTL zarar edilmiştir. TEBNK hisse senedinin ilgili tarihteki kapanış fiyatının da, alış fiyatının üstünde olması nedeni ile 586 YTL kar elde edilmiştir. Pozisyonun toplam getirisi 582 YTL olmuştur. SKBNK/TEBNK çiftinin günlük ve birikimli getiri grafikleri sırasıyla Şekil 4.5 ve Şekil 4.6’da ki gibidir.



Şekil 4.5. SKBNK/TEBNK hisse senedi çiftinin günlük getiri grafikleri



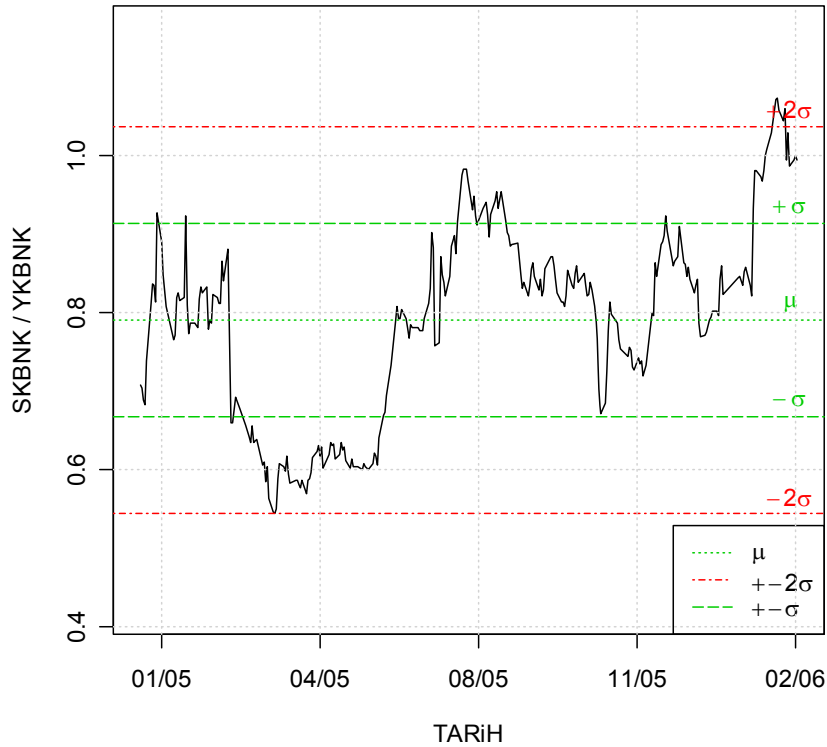
Şekil 4.6. SKBNK/TEBNK hisse senedi çiftinin birikimli getiri ve kapanış fiyatları grafikleri

İkinci hisse senedi çifti olan SKBNK/YKBNK kapanış fiyatı oranı ise, Şekil 4.7’de görüldüğü gibi, -2σ ve $+2\sigma$ limitlerini geçmiştir. Bu noktada, Şekil 4.8’de görüldüğü gibi, bu iki hisse senedi arasındaki kapanış fiyat farkı oldukça geniştir. Kapanış fiyat oran grafiği ve kapanış fiyatları grafiği birlikte incelendiğinde, iki hisse senedinin kapanış fiyatları farkı fiyat oranı uzun dönem ortalamasına döndüğünde kapanmakta ve -2σ veya $+2\sigma$ limitlerine doğru hareketlendiğinde ise açılmaktadır. Bu noktalarda, yine iyi performans gösteren hisse senedi satılacak, kötü performans gösteren hisse senedi alınacaktır.

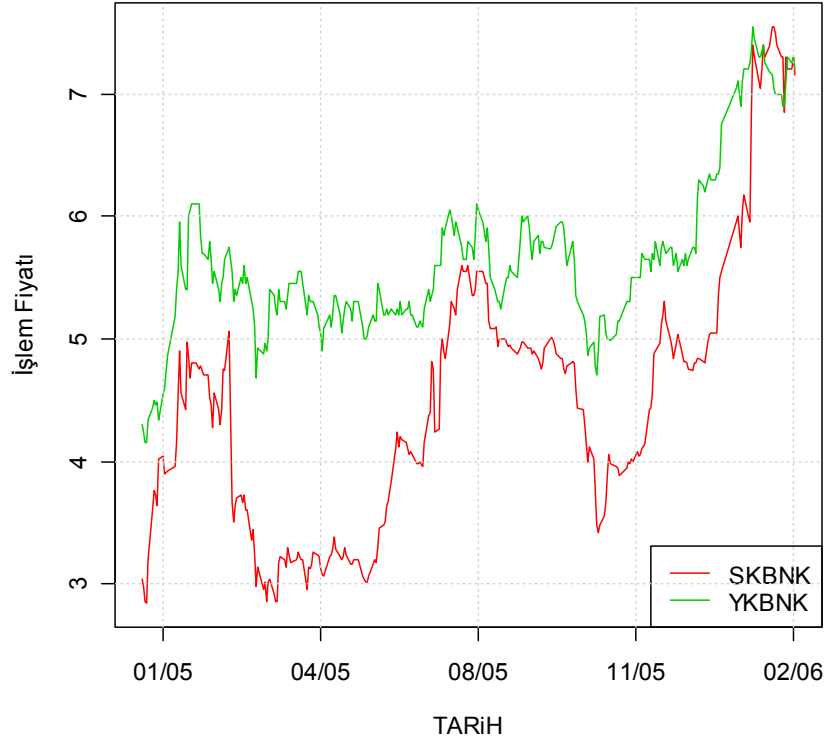
SKBNK/YKBNK hisse senedi çiftinin, Şekil 4.7’de ki kapanış fiyatı oran grafiğine bakıldığında 28.02.2005 tarihinde -2σ limitine çok yaklaştığı, 07.02.2005 tarihinde ise $+2\sigma$ limitini aştığı görülmektedir. Bu tarihler, hisse

senedi çifti için, pozisyon alma sinyalleridir. İlk pozisyonda, Şekil 4.8’de görüldüğü gibi, SKBNK hisse senedi YKBNK hisse senedine göre daha iyi performans, YKBNK hisse senedi ise diğer hisse senedine göre daha iyi performans göstermektedir. Buna göre, SKBNK hisse senedinde uzun pozisyon, YKBNK hisse senedin de kısa pozisyon alınmıştır. Bu pozisyonun kapatılma sinyali 14.06.2005 tarihinde, kapanış fiyatı oran grafiği oran ortalamasına döndüğünde alınmıştır.

Bu hisse senedi çiftinin kapanış fiyatı oranı, yine 07.02.2005 tarihinde $+2\sigma$ limitini geçmiş ve pozisyon alma sinyali vermiştir. Bu tarihte, SKBNK hisse senedi YKBNK hisse senedine göre daha iyi performans, YKBNK hisse senedi ise diğer hisse senedine göre kötü bir performans göstermiştir. Buna göre, SKBNK hisse senedin de kısa pozisyon, YKBNK hisse senedin de ise uzun pozisyon alınacaktır. Bu hisse senedi çiftinin açılan ikinci pozisyonu için henüz kapatılma sinyali alınmamıştır.



Şekil 4.7. SKBNK/YKBNK hisse senedi çiftinin kapanış fiyat oran grafiği



Şekil 4.8. SKBNK/YKBNK hisse senedi çiftinin kapanış fiyatları grafiği

SKBNK/YKBNK hisse senedi çiftinin yapılan alım – satım işlemleri sonucunda elde edilen getiri aşağıdaki çizelgede ki gibidir.

Çizelge 4.6. SKBNK/YKBNK çiftinin getiri çizelgesi

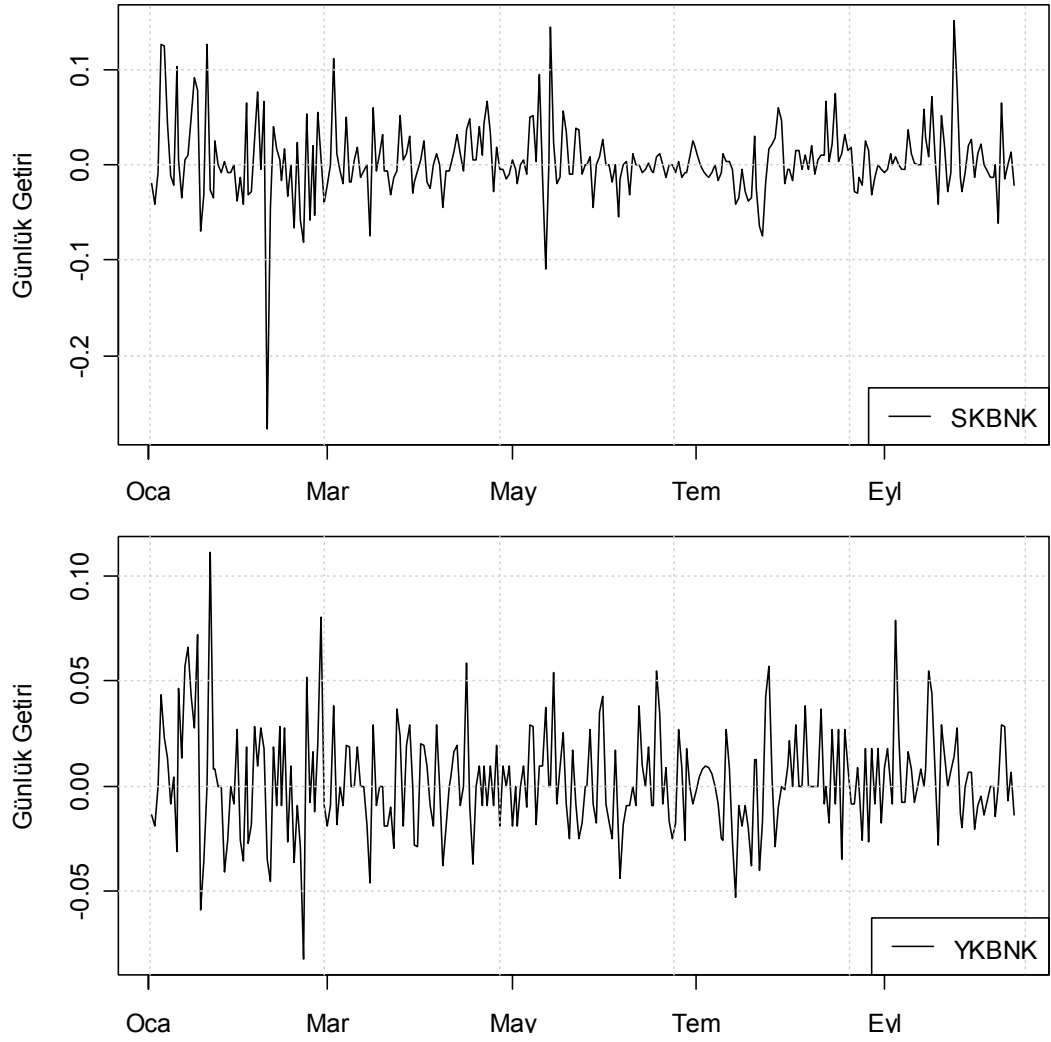
İşlem Tarihi		Uzun/Kısa Pozisyon	Tutar (YTL)		Adet		Getiri (YTL)	
Pozisyon Açma	28.03.2005	SKBNK/YKBNK	1.000	1000	+342	-186		
Pozisyon Kapatma	14.06.2005		1.452	981			+452	+19

Çizelge 4.6'a göre, 28.03.2005 tarihinde SKBNK hisse senedinde uzun pozisyon alınmıştır. Buna göre, 1.000 YTL karşılığı 342 adet SKBNK hisse senedi satın alınmış, yine 1.000 YTL karşılığı 186 adet YKBNK hisse senedinin açığa satışı yapılmıştır. Bu pozisyon 14.06.2005 tarihinde, satın alınan 342 adet SKBNK hisse senedi ilgili tarihteki kapanış fiyatından satılarak ve yine açığa satışı yapılan 186 adet YKBNK hisse senedi geri alınarak kapatılmıştır.

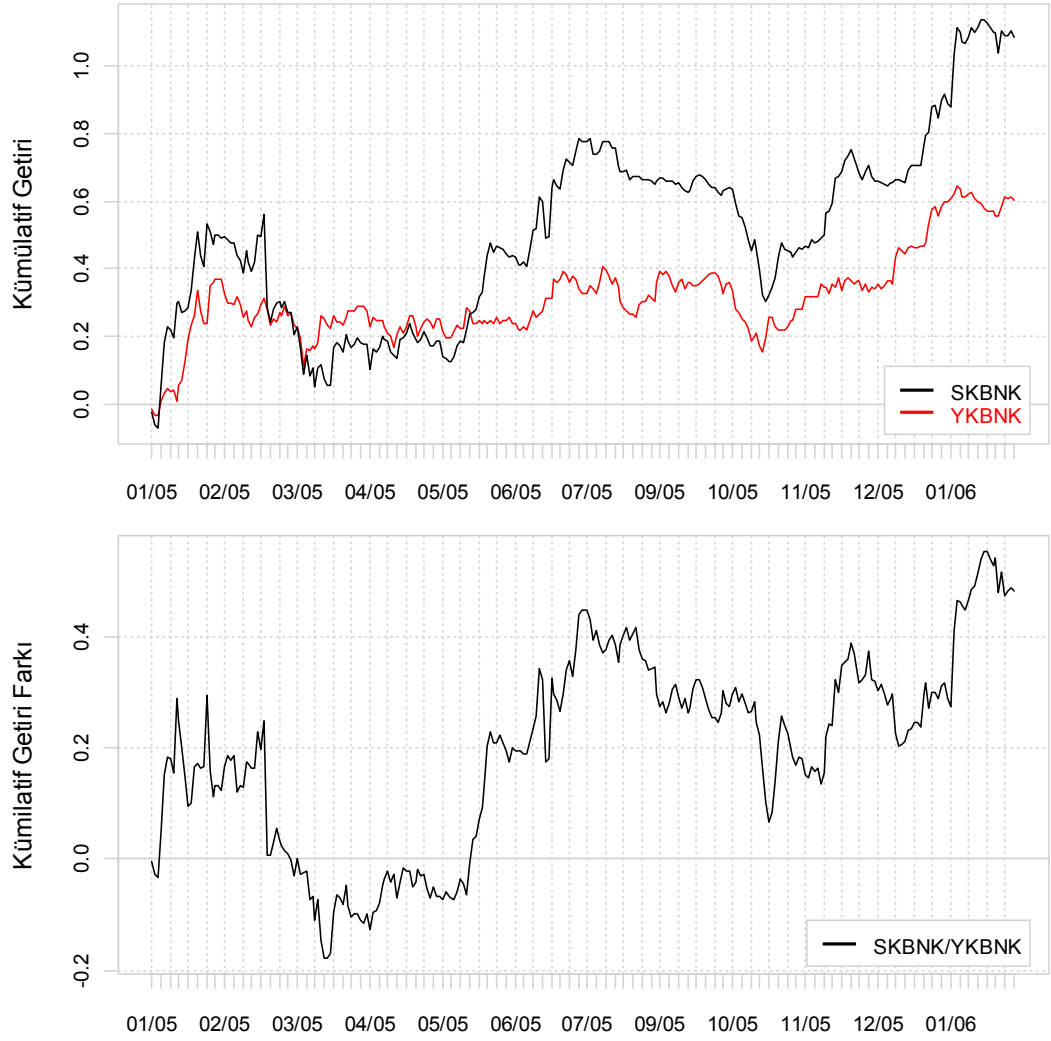
Yapılan işlemler sonucunda, pozisyon kapatıldığı tarihte, satın alınan SKBNK hisse senedinin satış fiyatının alış fiyatının üstünde olması nedeni ile 453 YTL kar elde edilmiştir. Açığa satışı yapılan YKBNK hisse senedinin ise aynı

tarihteki geri alış fiyatının satış fiyatının altında olması nedeni ile 19 YTL kar elde edilmiştir.

Bu hisse senedi çiftinin toplam getirisi 471 YTL olmuştur. SKBNK/YKBNK çiftinin günlük ve birikimli getiri grafikleri sırasıyla Şekil 4.9 ve Şekil 4.10'da ki gibidir.

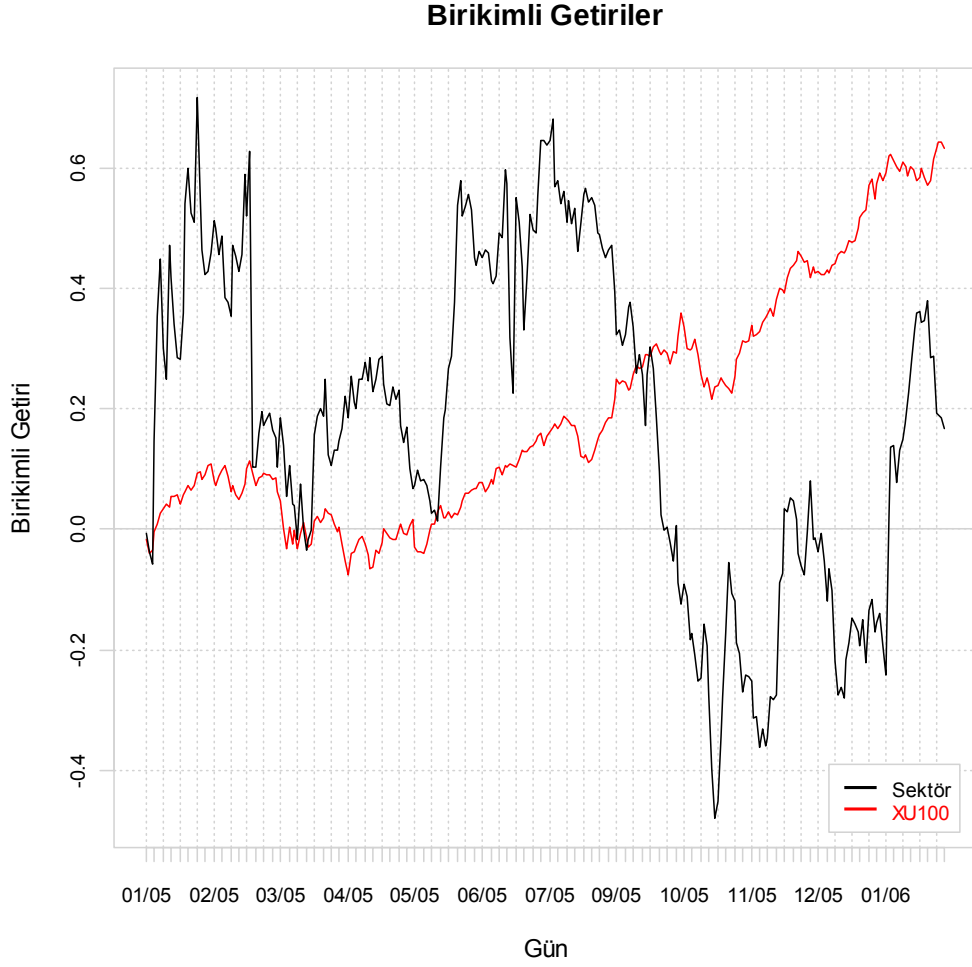


Şekil 4.9. SKBNK/YKBNK hisse senedi çiftinin günlük getiri grafikleri



Şekil 4.10. SKBNK/YKBNK hisse senedi çiftinin birikimli getiri ve kapanış fiyatları grafikleri

Bankacılık sektöründen elde edilen hisse senedi çiftlerinin toplam getirisinin Ulusal Tüm Endeks getirisi ile karşılaştırma grafiği Şekil 4.11'deki gibidir.

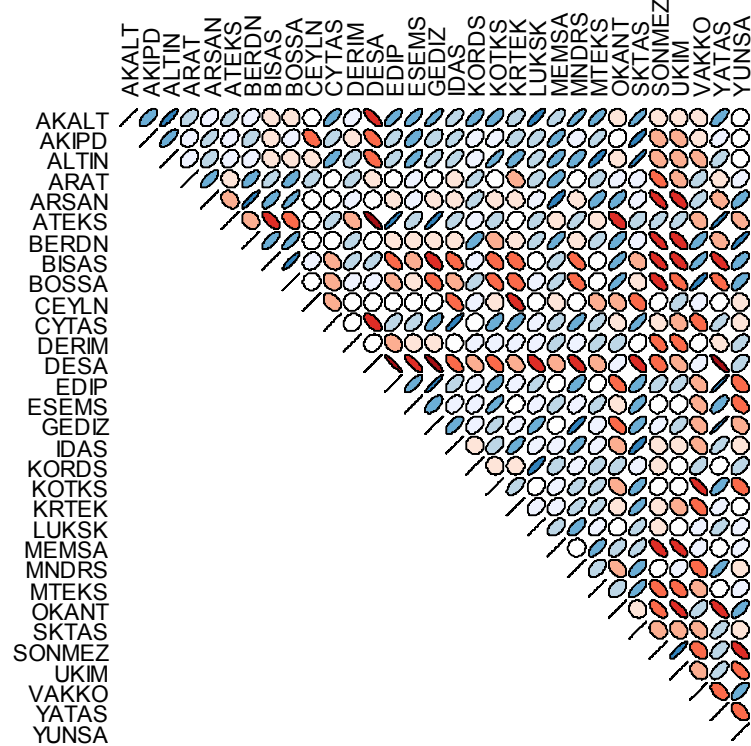


Şekil 4.11. Bankacılık Sektörü getirisinin Ulusal 100 Endeks getirisi ile karşılaştırması

4.3.2. Tekstil Sektörü için Hisse Senedi Çiftlerinin Belirlenmesi

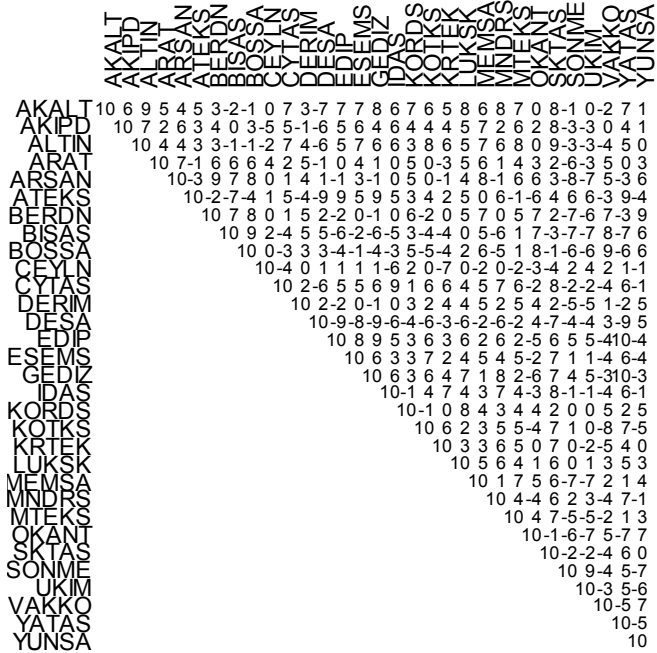
Tekstil sektöründe yer alan hisse senetleri AKALT, AKIPD, ALTIN, ARAT, ARSAN, ATEKS, BERDN, BISAS, BOSSA, CEYLN, CYTAS, DERIM, DESA, EDIP, ESEMS, GEDIZ, IDAS, KORDS, KOTKS, KRTEK, LUKSK, MEMSA, MNDRS, MTEKS, OKANT, SKTAS, SONME, UKIM, VAKKO, YATAS, YUNSA hisse senetlerinden oluşmaktadır. Tekstil sektörü için kısa dönem ve uzun dönem kapanış fiyatları ilişki matrisleri, çok fazla değişken olması nedeniyle görsel olarak Şekil 4.12 ve Şekil 4.13’de sırasıyla verilmiştir. İlgili ilişki matrisi EK – I ve EK – II de verilmiştir.

Kısa Dönem Korelasyon Matrisi



a) Kısa Dönem grafik gösterim

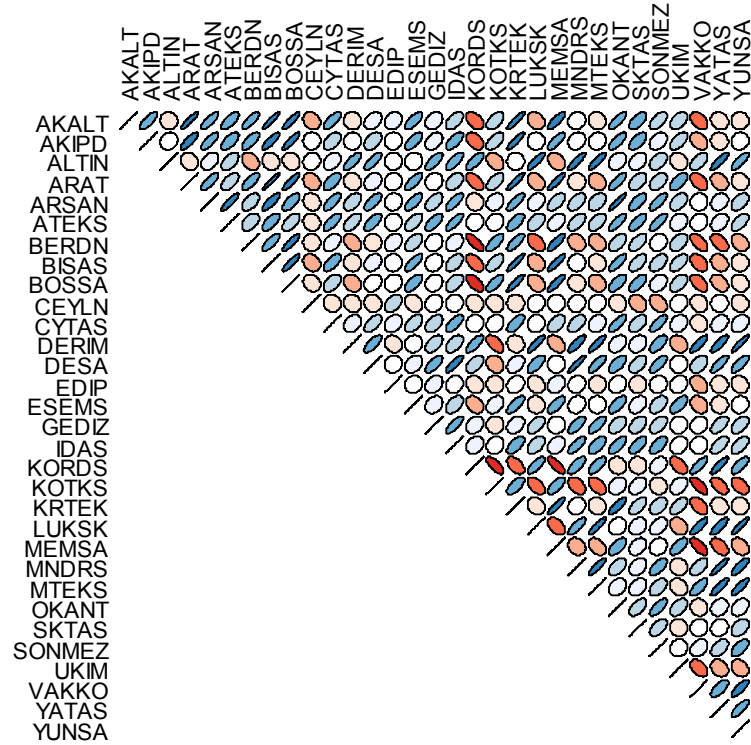
Kısa Dönem Korelasyon Matrisi



b) Kısa Dönem sayısal grafik gösterim

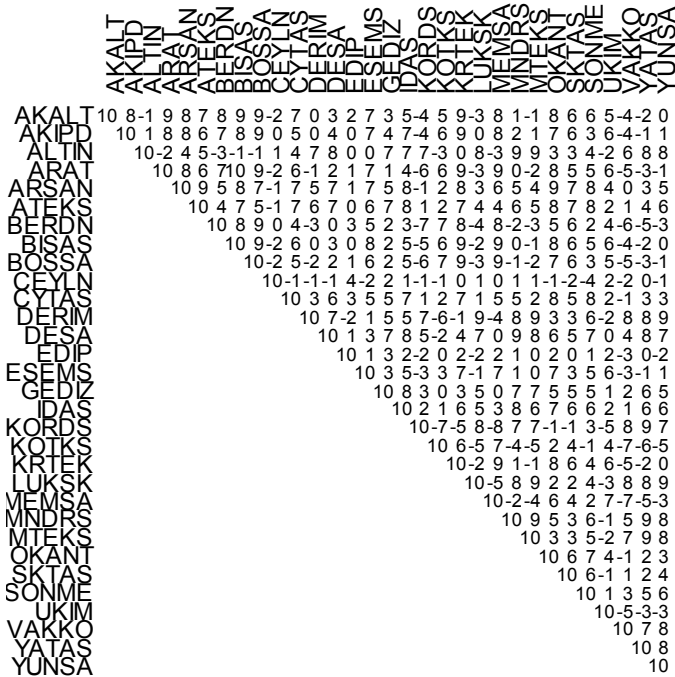
Şekil 4.12. Tekstil Sektörü Kısa Dönem İlişki Matrisi a) Grafik ve b) Sayısal grafik gösterimi

Uzun Dönem Korelasyon Matrisi



a) Uzun dönem grafik gösterim

Uzun Dönem Korelasyon Matrisi



b) Uzun dönem sayısal grafik gösterim

Şekil 4.13. Tekstil Sektörü Uzun Dönem İlişki Matrisi a) Grafik ve b) Sayısal grafik gösterimleri

Şekil 4.12 de kısa dönem ve Şekil 4.13 de uzun dönem ilişki matrislerine göre, kısa dönemde düşük ilişki ve uzun dönemde yüksek ilişki gösteren hisse senetleri eşleştirilmiştir. Elde edilen hisse senedi çiftleri portföye alınmış ve belirlenen $\mp 2\sigma$ kriterine göre karşılıklı olarak işlem yapılacaktır. Tekstil sektöründe oluşan hisse senedi çiftleri Çizelge 4.7’deki gibidir.

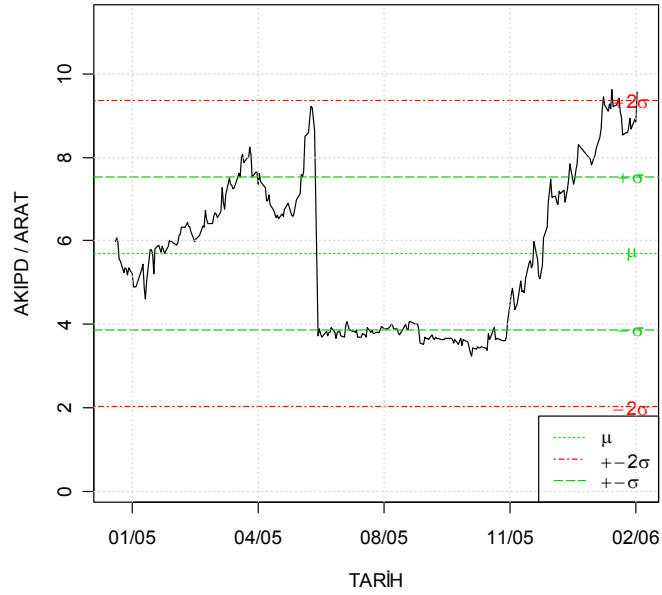
Çizelge 4.7. Uzun Dönem – Kısa Dönem Kapanış Fiyatı İlişkilerine Göre Eşleşmeler

	Kısa Dönem İlişkisi	Uzun Dönem İlişkisi
	ARAT	
AKIPD	0.242030692583632	0.816179581793994
	YUNSA	
ALTIN	0.00155142091642792	0.790668906693817
	MNDRS	
DERIM	0.241767310434274	0.807541482509538
	VAKKO	
DERIM	0.138740891355359	0.821311443325013
	YATAS	
KORDS	0.169561403214872	0.862891491686397
	YATAS	
MTEKS	0.104024064583081	0.850566976406115

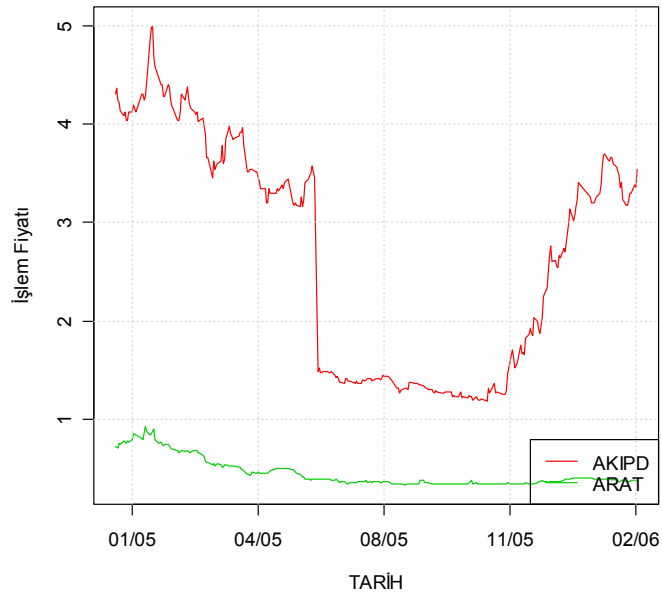
Oluşan hisse senedi çiftleri AKIPD/ARAT, ALTIN/YUNSA, DERIM/MNDRS, DERIM/VAKKO, KORDS/YATAS, MTEKS/YATAS çiftleridir.

AKIPD/ARAT hisse senedi çiftinin, Şekil 4.14’deki kapanış fiyat oran grafiğinde görüldüğü gibi, $+2\sigma$ limitine doğru hareket etmiş ancak limiti aşmamıştır. Bu noktada, Şekil 4.15’te görüldüğü gibi kötü performans gösteren AKIPD ve ARAT hisse senetlerinin her ikisi de aşağı yönlü hareket etmektedir. Ancak, AKIP hisse senedi diğer hisse senedine göre daha hızlı hareket etmektedir. Kapanış fiyatı oranı, ikinci defa $+2\sigma$ limitine doğru yönelmiş ve bu limiti aşmıştır. Hisse senedi çiftinin fiyat aralığı yine artmaktadır. $+2\sigma$ limitinin aşıldığı noktada pozisyon almak amacıyla, iyi performans gösteren AKIPD hisse senedinin aşağı doğru hareket edeceği tahminiyle açığa satışı yapılırken, kötü performans gösteren ARAT hisse senedinin yukarı doğru hareket edeceği tahminiyle nakit sermaye karşılığı satın alınacaktır. Bu hisse senedi çiftindeki pozisyonunun kapatılması için gerekli sinyal, kapanış fiyatı oranının uzun dönem kapanış fiyatı oranı

ortalamasına dönmesi durumunda verilmiş olacaktır. Bu anda, açığa satışı gerçekleştirilmiş olan AKIPD hisse senedinin satış zamanı fiyatından daha düşük bir fiyata, ARAT hisse senedinin ise alış zamanı fiyatından daha yüksek bir fiyata sahip olması beklenir. Bu beklenti gerçekleştiğinde, hisse senedi çiftinin her ikisinden de kar edilmiş olunacaktır. AKIPD/ARAT hisse senedi çifti henüz pozisyon kapatma sinyali vermemiştir.

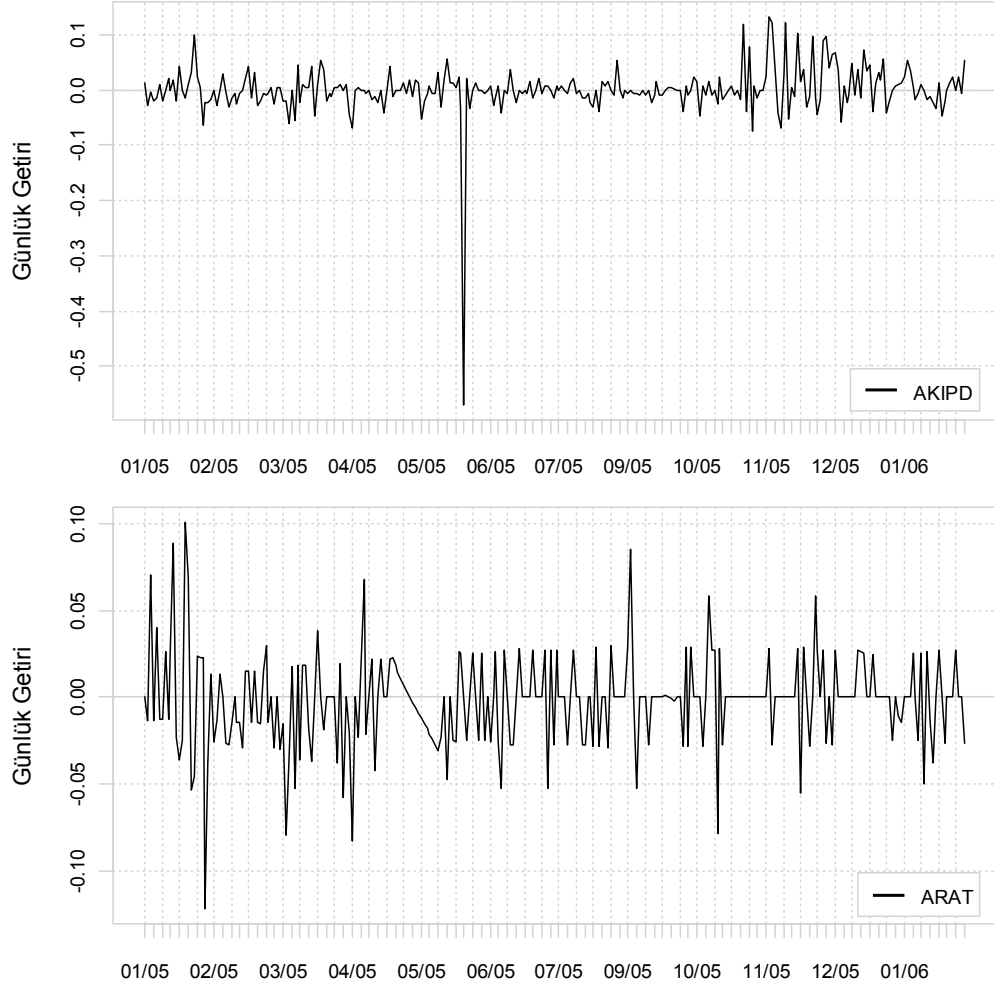


Şekil 4.14. AKIPD/ ARAT hisse senedi çiftinin kapanış fiyat oran grafiği

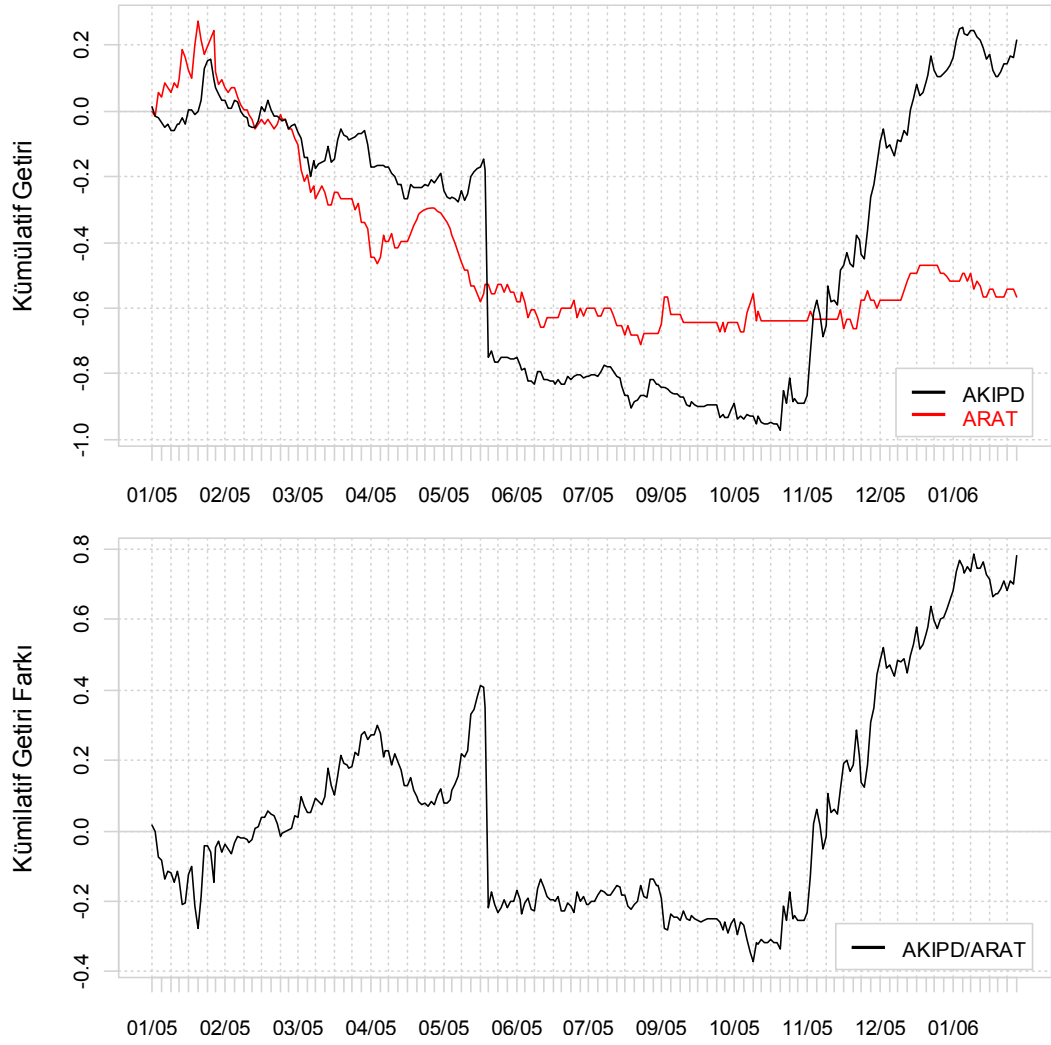


Şekil 4.15. AKIPD/ARAT hisse senedi çiftinin kapanış fiyatları grafiği

Bu hisse senedi çiftinin günlük ve kümülatif getiri grafikleri sırasıyla Şekil 4.16 ve Şekil 4.17’de görüldüğü gibidir.



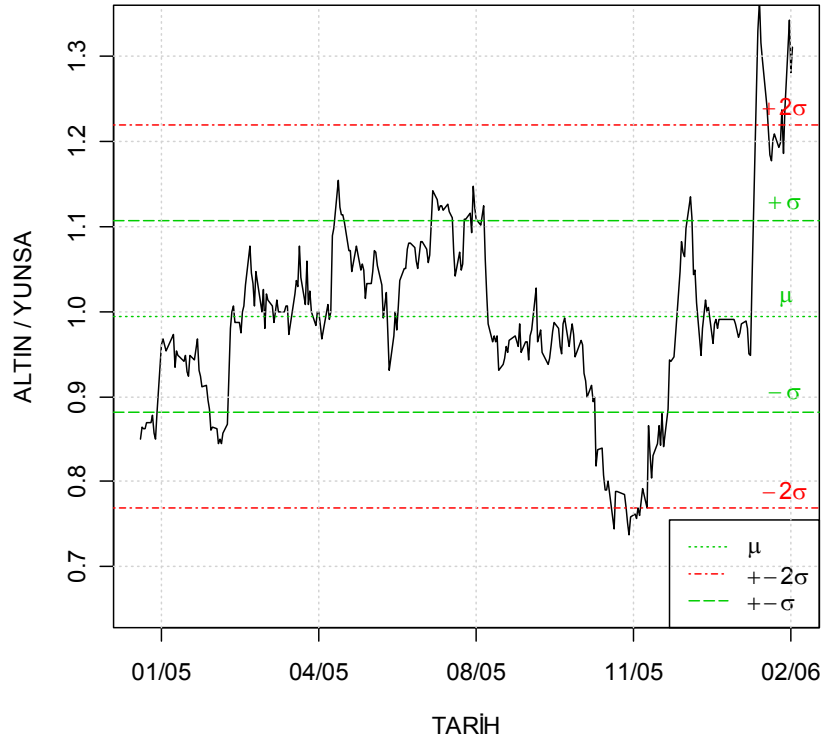
Şekil 4.16. AKIPD/ARAT hisse senedi çiftinin günlük getiri grafiği



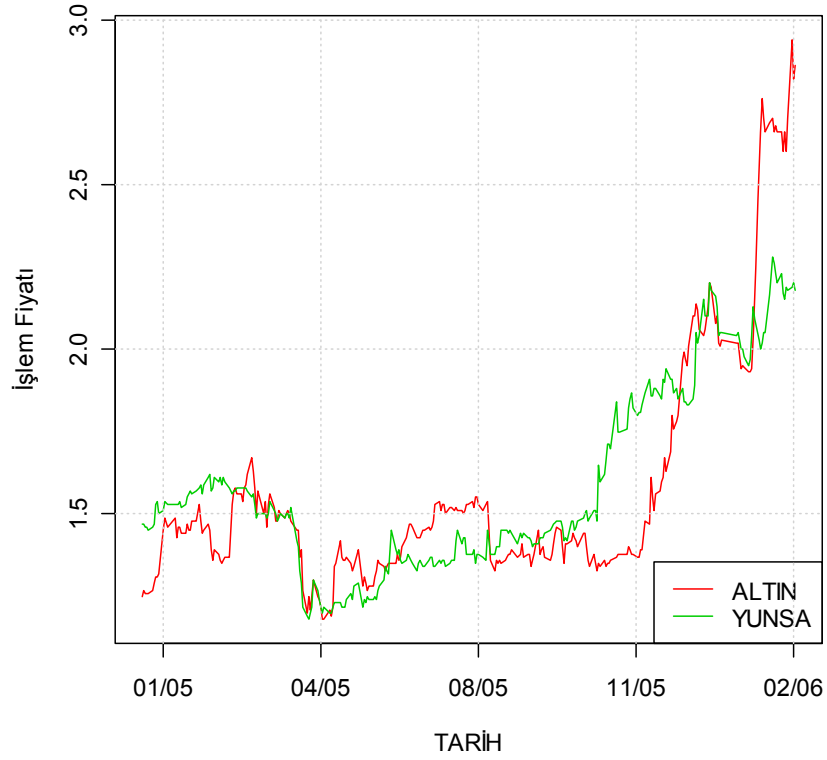
Şekil 4.17. AKIPD/ARAT hisse senedi çiftinin kümülatif getiri grafikleri

ALTIN/YUNSA hisse senedi çiftinin Şekil 4.18'deki kapanış fiyatı oran grafiğine bakıldığında, 31.10.2005 tarihinde -2σ limitini aştığı görülmektedir. Bu noktada, YUNSA hisse senedi diğer hisse senedine göre daha iyi performans, ALTIN hisse senedi ise diğer hisse senedine göre kötü performans gösterdiğinden ALTIN hisse senedinde uzun pozisyon alınacak, YUNSA hisse senedinde ise kısa pozisyon alınacaktır. 21.12.2005 tarihinde, kapanış fiyatı oranı uzun dönem oran ortalamasına döndüğünden, Şekil 4.19'da görüldüğü gibi, kapanış fiyatları arasındaki farkta kapanmıştır, dolayısıyla hisse çiftinde alınan pozisyon için kapatılma sinyali alınmıştır. $+2\sigma$ limiti kırıldığında ise, ALTIN hisse senedi iyi performans, YUNSA hisse senedi kötü performans gösterdiğinden, YUNSA hisse senedi satın alınacak, ALTIN hisse senedinin açığa satışı yapılacaktır. Bu

pozisyon, kapanış fiyatı oranı, oran ortalamasına döndüğünde ters işlem yapılarak kapatılacaktır. ALTIN/YUNSA hisse senedi çifti, yine 30.01.2005 tarihinde $+2\sigma$ limitini aşarak pozisyon alma sinyali vermiştir. Bu tarihte, ALTIN hisse senedi YUNSA hisse senedine göre iyi performans, YUNSA hisse senedi ise diğer hisse senedine göre kötü performans göstermiştir. Buna göre, ALTIN hisse senedinde kısa pozisyon, YUNSA hisse senedinde uzun pozisyon alınacaktır. Ancak, ikinci pozisyon için henüz kapatma sinyali vermemiştir.



Şekil 4.18. ALTIN/YUNSA hisse senedi çiftinin kapanış fiyat oran grafiği



Şekil 4.19. ALTIN/YUNSA hisse senedi çiftinin kapanış fiyatları grafiği

Bu hisse senedi çiftinin yapılan alım – satım işlemler sonucunda elde edilen getirisi aşağıdaki çizelgedeki gibidir.

Çizelge 4.8. ALTIN/YUNSA çiftinin getiri çizelgesi

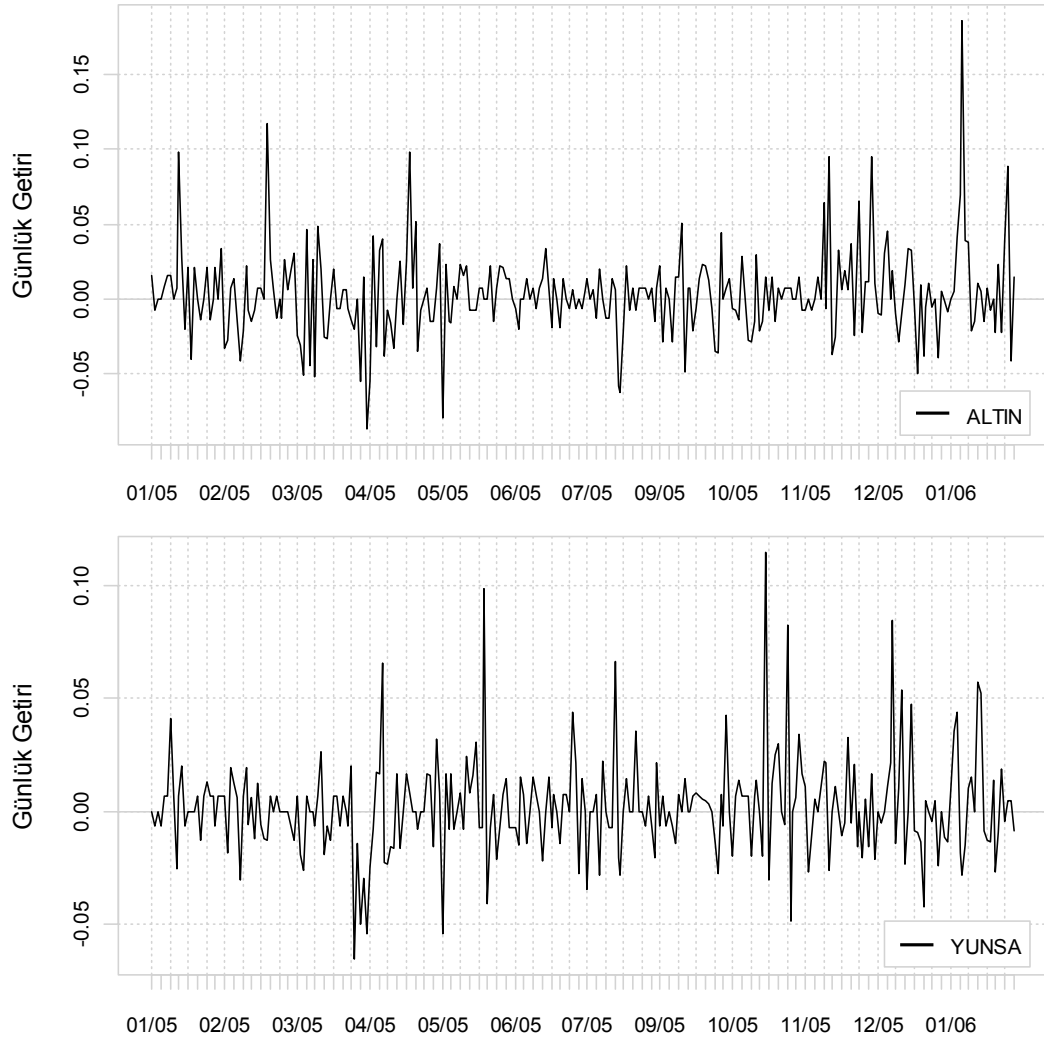
İşlem Tarihi		Uzun/Kısa Pozisyon	Tutar (YTL)		Adet		Getiri (YTL)	
Pozisyon Açma	31.10.2005	ALTIN/YUNSA	1.000	1.000	+730	-543		
Pozisyon Kapatma	12.12.2005		1.438	1.022			+438	-22

Çizelge 4.8'e göre, 31.10.2005 tarihinde ALTIN hisse senedine uzun pozisyon, YUNSA hisse senedinde kısa pozisyon alınmıştır. Buna göre, 1.000 YTL karşılığı 730 adet hisse senedi satın alınmış, yine 1.000 YTL karşılığı 543 adet YUNSA hisse senedinin açığa satışı yapılmıştır. Bu pozisyon, 12.12.2005 tarihinde, satın alınan 730 adet ALTIN hisse senedi ilgili tarihteki kapanış fiyatından satılmış, açığa satışı yapılan 543 adet YUNSA hisse senedi ilgili tarihteki kapanış fiyatından geri satın alınmıştır.

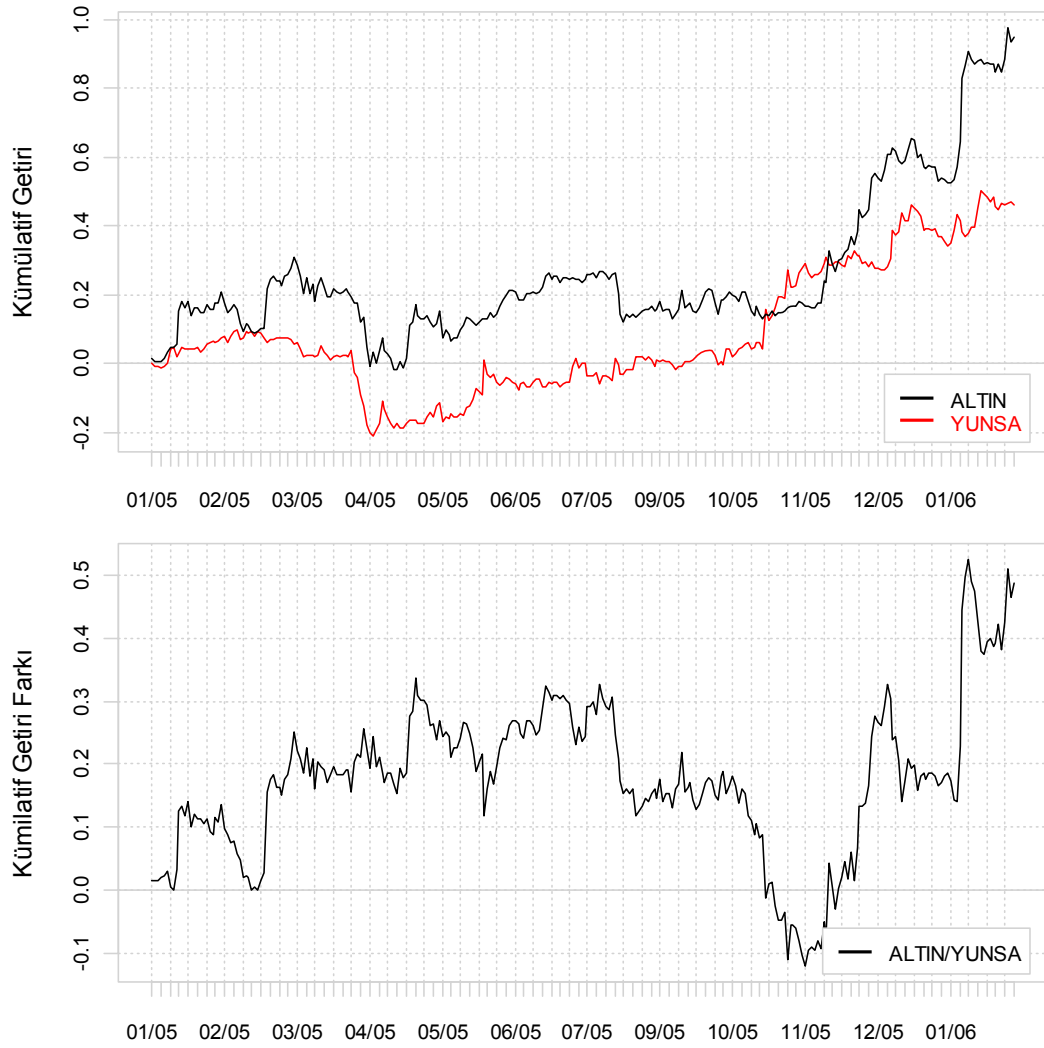
Yapılan işlemler sonucunda, ALTIN hisse senedinin satış fiyatının alış fiyatının üstünde olması nedeni ile 438 YTL kar elde edilmiştir. YUNSA hisse

senedinin geri satın alma fiyatı ise, açığa satış fiyatının üstünde olması nedeniyle 22 YTL zarar edilmiştir. Pozisyonun toplam getirisi 416 YTL olmuştur.

Bu hisse senedi çiftinin günlük ve kümülatif getiri grafikleri sırasıyla Şekil 4.20 ve Şekil 4.21’de ki gibidir.



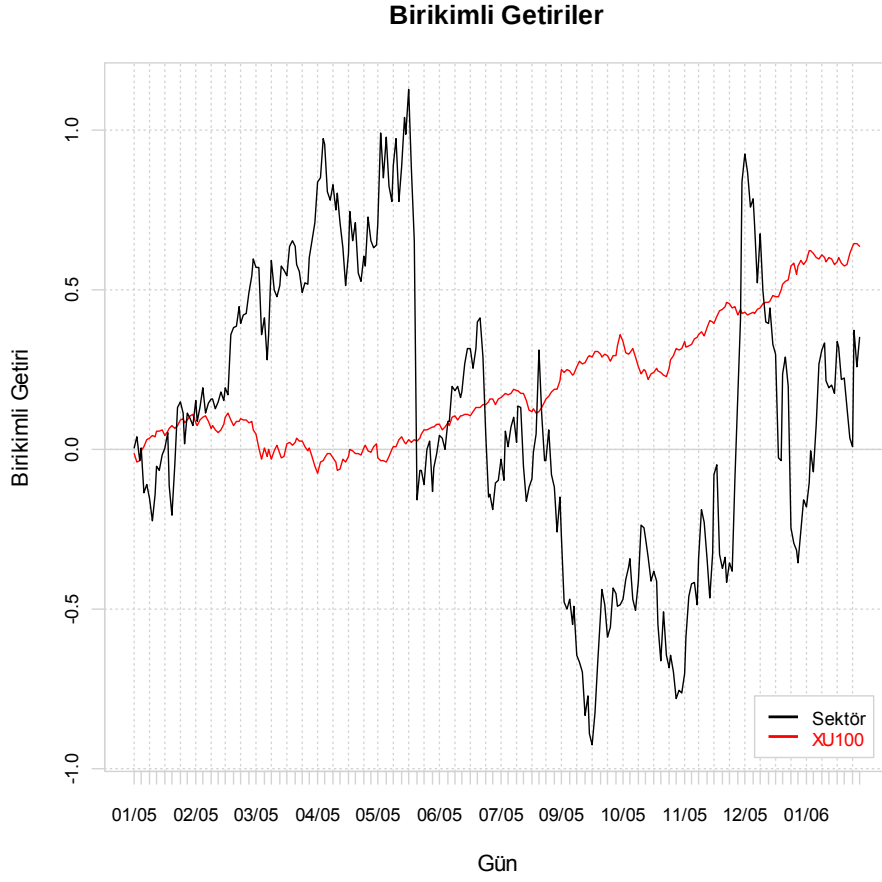
Şekil 4.20. ALTIN/YUNSA hisse senedi çiftinin günlük getiri grafiği



Şekil 4.21. ALTIN/YUNSA hisse senedi çiftinin kümülatif getiri grafikleri

Portföye dahil olan hisse senedi çiftleri olan DERİM/MNDRS, DERİM/VAKKO, KORDS/YATAS ve MYEKS/YATAS çiftlerinin, kapanış fiyat oranı, kapanış fiyatları, günlük getiri ve kümülatif getiri grafikleri ve getiri çizelgeleri EK – III de verilmiştir.

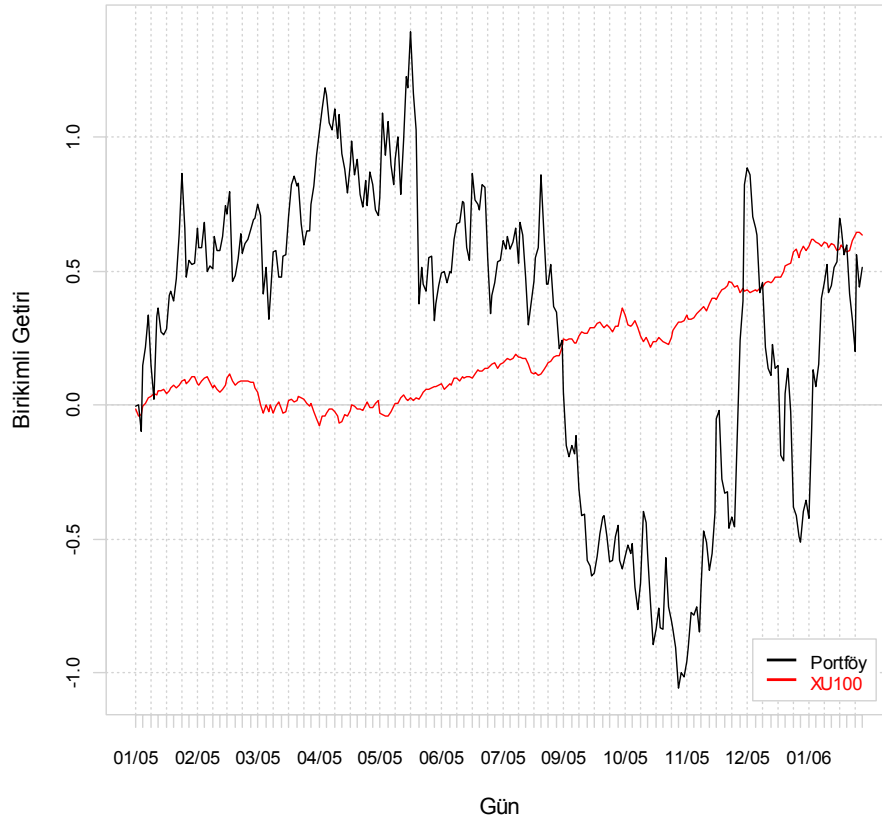
Tekstil sektöründen elde edilen hisse senedi çiftlerinin toplam getirilerinin Ulusal Tüm Endeks getirisi ile karşılaştırması Şekil 4.22'deki gibidir. Sektör getirisinde, pozisyon alma sinyali vermeyen hisse senedi çiftleri de dahil edilmiştir.



Şekil 4.22. Tekstil Sektörü getirisinin Ulusal 100 Endeks getirisi ile karşılaştırması

Eş işlem stratejisi kullanılarak oluşturulan portföyün toplam getirisinin Ulusal Tüm Endeks getirisi ile karşılaştırılması Şekil 4.23’da ki gibidir. Toplam portföy getirisinde, pozisyon alma sinyali vermeyen hisse senedi çiftleri de dahil edilmiştir.

Portföy ve XU100 Birikimli Getirileri



Şekil 4.23. Portföy getirisinin Ulusal Tüm Endeks getirisi ile karşılaştırması

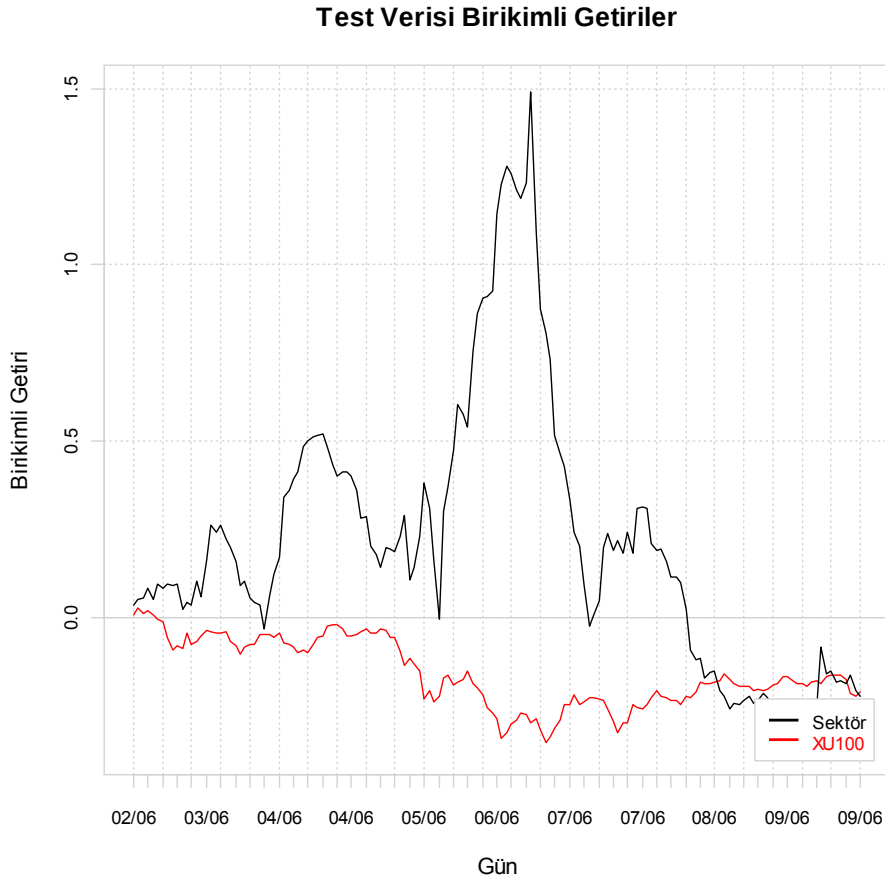
Eş işlem stratejisi sonunda portföyü oluşturan hisse senetlerinin getirileri ve portföyün toplam getirisi Çizelge 4.9’de ki gibi elde edilmiştir.

Çizelge 4.9. Portföy Nakit Getirisi

İşlem Tarihi		Uzun/Kısa Pozisyon	Tutar (YTL)		Adet		Getiri (YTL)	
Pozisyon Açma	28.03.2005	SKNBNK/YKBNK	1.000	1000	+342	-186		
Pozisyon Kapatma	14.06.2005		1.452	981			+452	+19
Pozisyon Açma	27.01.2005	SKBNK/TEBNK	1.000	1.000	-204	+114		
Pozisyon Kapatma	01.09.2005		1.004	1.586			-4	+586
Pozisyon Açma	31.10.2005	ALTIN/YUNSA	1.000	1.000	+730	-543		
Pozisyon Kapatma	12.12.2005		1.438	1.022			+438	-22
Pozisyon Açma	16.11.2005	DERIM/MNDRS	1.000	1.000	-405	+1.471		
Pozisyon Kapatma	28.12.2005		1.190	1.529			-190	+529
Pozisyon Açma	11.07.2005	KORDS/YATAS	1.000	1.000	-272	+892		
Pozisyon Kapatma	08.08.2005		1.130	1.491			-130	+491
Pozisyon Açma	18.10.2005	KORD/YATAS	1.000	1.000	-240	+800		
Pozisyon Kapatma	05.12.2005		1.034	1.392			-34	+392
Pozisyon Açma	23.05.2005	MTEKS/YATAS	1.000	1.000	-1.389	+1099		
Pozisyon Kapatma	13.06.2005		944	1.418			+54	+418
Pozisyon Açma	12.12.2005	MTEKS/YATAS	1.000	1.000	-752	+599		
Pozisyon Kapatma	16.01.2006		1.135	1.512			-135	+512
TOPLAM								2.212

4.3.3. Eş İşlem Stratejisinin Test Verisi İle Sınanması

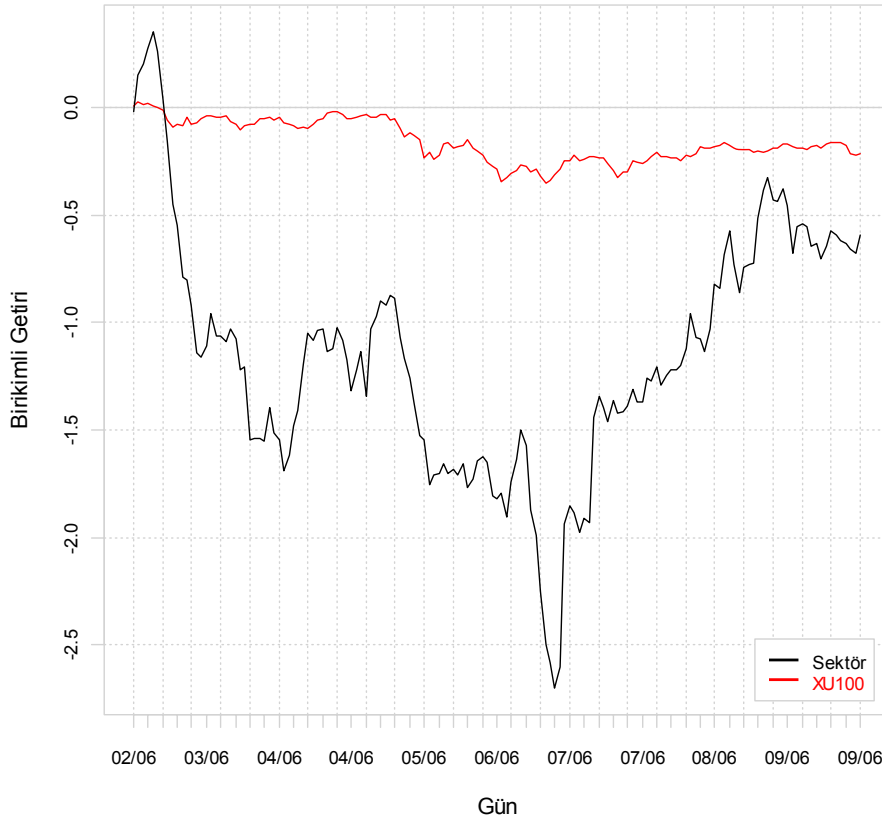
Bu kesimde, İMKB’de, Bankacılık ve Tekstil sektörleri temel alınarak, işlem gören hisse senetleriyle, eş işlem stratejisi kullanılarak oluşturulan portföye dahil edilen hisse senetlerinin test verisi (gelecek dönem kapanış fiyatları) ile kümülatif getirileri hesaplanarak Ulusal 100 Endeksi ile genel bir karşılaştırması yapılmıştır. Bankacılık sektöründen oluşturulan hisse senedi çiftlerinin kümülatif getirisinin Ulusal 100 endeksi ile karşılaştırılması Şekil 4.24’ta ki gibidir.



Şekil 4.24. Bankacılık Sektörü Gelecek Dönem Getirisi - Ulusal 100 Endeksi Karşılaştırması

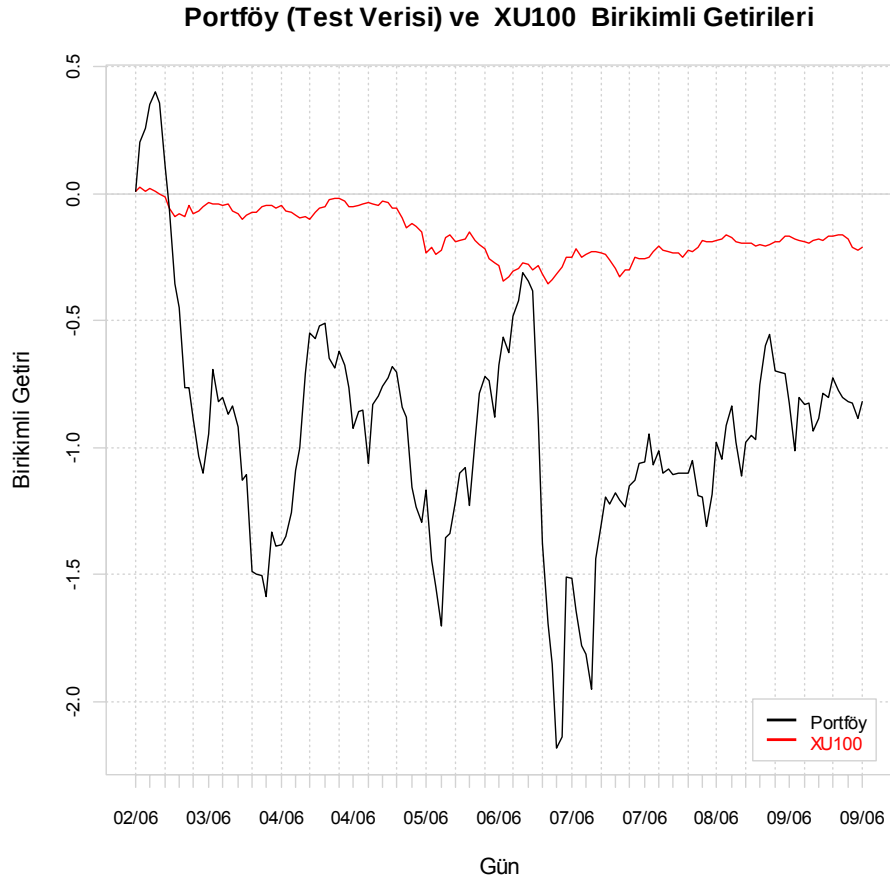
Tekstil sektöründen oluşturulan hisse senedi çiftlerinin kümülatif getirilerinin Ulusal 100 endeksi ile karşılaştırması Şekil 4.25’de ki gibidir.

Test Verisi Birikimli Getiriler



Şekil 4.25. Tekstil Sektörü Gelecek Dönem Getirisi - Ulusal 100 Endeksi Karşılaştırması

Eş işlem stratejisi kullanılarak elde edilen hisse senedi çiftleriyle gelecek dönem kapanış fiyatlarının kümülatif getirileri hesaplanarak, oluşturulan portföyün getirisinin Ulusal 100 Endeksi ile karşılaştırması Şekil 4.26'daki gibidir.



Şekil 4.26. Portföy Gelecek Dönem Getirisi - Ulusal 100 Endeksi karşılaştırması

Gelecek dönem getirilerinin karşılaştırması sonucunda Bankacılık sektöründen elde edilen hisse senedi çiftlerinin getirisinin Ulusal 100 Endeksinden daha iyi olduğu gözükmemektedir. Ancak, Tekstil sektöründen elde edilen hisse senedi çiftlerinin getirisinin ise Ulusal 100 Endeksinden kötü olduğu görülmektedir.

5. SONUÇ VE ÖNERİLER

Günümüz teknolojisindeki gelişmeler, veri kaynaklarının çeşitlenmesine neden olduğu gibi, veri üretimini de hızlandırmıştır. Çeşitli veri kaynaklarından, hiç ara vermeksizin elde edilen veri, veriyi değerlendirmek isteyen karar vericinin çeşitli zorluklarla karşılaşmasına neden olmaktadır. Örneğin, bir süpermarket zinciri yöneticisi, zincir mağazalarından alışveriş yapan müşterilerinin profilini ve alış –veriş alışkanlıklarını öğrenerek, ürün gamını ve raf düzenlerini müşteri profili ve alış –veriş alışkanlıklarına göre düzenlemek isteyebilir. Ancak, teknolojinin getirmekte olduğu yenilikler, her bir mağazanın herhangi bir ödeme kasasında yapılan alış –veriş listelerinin, kalem kullanılarak kesilen makbuzlardan her ürüne özel hazırlanan barkodların barkod okuyucularla saniyelerle ölçülen sürelerde veriye dönüştürülmesine neden olmaktadır. Karar vericilerin bu kadar büyük ve bu kadar hızlı güncellenen bir veri yığını içinde kaybolması ve verdiği kararların geçerliliğini yitirmesi işten bile değildir. Bununla birlikte, büyük miktarda ve hızlı bir şekilde akan veri içinde fark edilememiş gizli bilgilerin var olması da yadsınamaz.

Rüzgar gülü, barometre ve termometre gibi ölçüm cihazlarının yanında Dopler radarı, uzaktan algılama özelliklerine sahip uydu ve benzeri gelişmiş cihazlardan gelen verilerle yapılan meteorolojik analizler de başka bir örneği teşkil edebilir. Burada da benzer sorunlar yanında başka sorunlarda vardır. Meteoroloji uzmanının yaptığı analizin ve elde ettiği sonuçların anlam ve önemini kaybetmemesi için, farklı bilgi kaynaklarından, farklı veri yapılarında gelen bilgiyi önce anlaşılabilir ortak bir veri yapısına dönüştürmesi ve hızlı bir şekilde yorumlaması gereklidir.

Buna benzer durumlar finansal piyasalarda yatırım yapan yatırımcılar için de geçerli olabilmektedir. Yatırımcı yapmış olduğu yatırımların fiyat değişikliklerini, yatırımlarını ilgilendiren ulusal veya uluslararası haber kaynaklarında çıkan haberler gibi çeşitli bilgileri basılı, görsel veya dijital ortamdan takip edebilmektedir. Ancak, veri akışını kontrol altında tutması ve değerlendirmesi, hızlı ve farklı kaynaklardan gelen veri nedeniyle zorlaşmakta ve yatırımları hakkında vermesi gereken kararları erken ve geç almasına neden olabilmektedir.

Daha öncede değinildiği gibi, veri madenciliği, otomatik veya yarı otomatik anlamda, birçok veri kaynağından hızlı ve bazen farklı yapılarda gelen büyük miktardaki statik veya dinamik veriyi derleyerek, görselleştirmek, yorumlamaya yardımcı olmak, işlerlik kazandırmak ve karar sürecine uyarlamak için çeşitli modelleri bir araya getirerek bir sistem oluşturulmasını hedefleyen bir süreçtir. Bununla birlikte veri içerisinde var olabilecek gizli kalmış yeni ilişkileri, eğilimleri ve örüntüleri ortaya çıkarmakta hedeflerinden bir tanesidir. Veri, bilgi, özbilgi, kavrama ve bilgelik adımlarından oluşan bu süreç, istatistiksel ve matematiksel tekniklerle, örüntü tanımlama, yapay zeka ve veritabanı gibi teknolojileri bir araya getirir. Bu sürecin adımlarını, amaçların belirlenmesi, verinin derlenmesi, veri madenciliği modelinin belirlenmesi, veri analiz ve modelin değerlendirilmesi ve son adımda karar sürecine uyarlaması oluşturur.

Sermaye piyasalarında yatırım yapan tasarruf sahipleri, çeşitli menkul kıymetlere yatırım yaparak, yatırımlarından minimum riskle maksimum getiriye elde etmeyi amaçlar. Yatırımcının, bu amaca ulaşabilmek için, önünde hisse senedi ve tahvil gibi çeşitli menkul kıymet seçenekleri mevcuttur ve seçeneklerden bir veya birkaçını seçerek oluşturduğu portföyünü ekonomik koşullara göre etkin bir şekilde yönetmesi gereklidir. Bu noktada yatırımcı hangi menkul kıymeti, ne zaman, ne miktarda bulundurması gerektiği gibi sorularla karşılaşır. İşte bu soruları cevaplayabilmek için, yapacağı yatırımlar ile ilgili veriyi çeşitli kaynaklardan derlemesi, daha anlaşılır hale getirmesi, analiz etmesi ve yapmak istediği yatırımlar hakkında alacağı kararlara uyarlaması gerekir. Bu noktada veri madenciliği yatırımcıya yardımcı olarak devreye girmektedir.

Bu çalışmada, İstanbul Menkul Kıymetler Borsası'nda işlem gören bankacılık ve tekstil sektörlerine dahil hisse senetlerinin, eş işlem stratejisi kullanılarak, hisse senedi seçimi yaparak bir portföy oluşturan ve bu hisse senetleri ile işlem zamanı için sinyal veren bir veri madenciliği sistemi oluşturulmuştur. Eş işlem stratejisi, sermaye piyasasında işlem gören hisse senetlerinin fiyatlarını dikkate alarak, uzun dönemde yüksek ilişki gösteren, kısa dönemde ise düşük ilişki gösteren hisse senetlerini eşleyerek bir portföy oluşturur. Stratejinin işlem sinyalini vermesi, elde edilen hisse senedi çiftlerinin uzun dönem fiyat oranlarının $\pm 2\sigma$ limitlerini geçmesi ve tekrar uzun dönem fiyat oran

ortalamasına dönmesi ilkesine dayanır. Hisse senedi kapanış fiyatları verisi Ocak 2005 – Eylül 2006 tarihleri arasında resmi iş günlerindeki kapanış fiyatlarından oluşmaktadır. Bu kapanış fiyatları verisinin Ocak 2005 – Şubat 2006 tarihleri arasındaki kısmı eğitim verisi, geriye kalan kısmı ise test verisi olarak kullanılmıştır. Oluşturulan sistem, eğitim verisinde yer alan kapanış fiyatlarını gözden geçirerek, resmi iş günlerinde kapanış fiyatlarında eksik veri olup olmadığını incelemekte, eksik verinin olması durumunda kübik spline yöntemini kullanarak bu değerleri tahmin ederek gidermektedir.

Eş işlem stratejisinin dayandığı nokta, uzun dönem fiyat oranının $\mp 2\sigma$ limitlerine yaklaştıkça hisse senedi çiftlerinin fiyatları arasındaki farkın artmasıdır. Limit aşıldığı andaki fiyat performansına göre çifti oluşturan hisse senetleri üzerinde uzun ve kısa pozisyon alınacaktır. Alınan pozisyonlardaki beklenti, uzun pozisyon alınan hisse senedinin fiyatının artacağı, kısa pozisyon alınan hisse senedinin fiyatının ise azalacağı yönündedir. Hisse senedi çiftinin uzun dönem fiyat oranı uzun dönem oran ortalamasına düştüğünde, uzun pozisyonun ilk fiyatın üzerinde kapanması ve kısa pozisyonun ise ilk fiyatın altında kapanması durumunda iki pozisyondan pozitif getiri elde edilecektir. Oluşturulan sistemle, bankacılık ve tekstil sektörlerinin her biri için ilgili işlemler yapılmıştır. Eş işlem stratejisinin hisse senedi çiftlerini belirleme şekline göre, bankacılık sektöründen SKBNK/YKBNK, SKBNK/TEBNK ve tekstil sektöründen ise AKIPD/ARAT, ALTIN/YUNSA, DERIM/MNDRS, DERIM/VAKKO, KORDS/YATAS ve MTEKS/YATAS olmak üzere toplam sekiz hisse senedi çifti elde edilmiştir.

Hisse senedi çiftleri oluşturulurken, uzun dönem ilişkisi en az 0.75 düzeyinde, kısa dönem ilişkisi ise en fazla 0.25 düzeyinde olan hisse senetleri eşleştirilmiştir. Buna göre bankacılık sektöründen seçilen hisse senetleri, uzun ve kısa dönemde sırasıyla 0.835 ve 0.031, 0.756 ve 0.209 düzeyindeki ilişki katsayıları ile eşleşmiştir.

Tekstil sektöründen seçilen hisse senetleri ise, uzun ve kısa dönemde sırasıyla 0.816 ve 0.242, 0.791 ve 0.002, 0.808 ve 0.242, 0.821 ve 0.139, 0.863 ve 0.170, 0.850 ve 0.104 düzeyindeki ilişki katsayıları ile eşleşmiştir.

Elde edilen hisse senedi çiftlerinde işlem sinyalinin belirlenebilmesi için fiyat oran dönüşümünün yapılması gerekmektedir. Oluşturulan sistem, bu nedenle, elde edilen hisse senedi çiftleri için kapanış fiyat oran dönüşümlerini yaptıktan sonra, kapanış fiyat oranları için uzun dönem oran ortalaması ile $\mp \sigma$ ve $\mp 2\sigma$ limitlerini de hesaplayarak işlem sinyali için $\mp 2\sigma$ limitlerinin aşılp aşılmadığı incelenmiştir.

Sistemin yapmış olduğu inceleme sonucunda, bankacılık sektöründen elde edilen iki hisse senedi çifti de işlem sinyali vermiştir. Belirlenen $\mp 2\sigma$ limitlerinin aşıldığı noktada, bu noktaya kadar olan zaman diliminde iyi performans gösteren hisse senetlerinde fiyatının düşeceği beklentisiyle kısa pozisyon, kötü performans gösteren hisse senetlerinde ise fiyatının yükseleceği beklentisiyle uzun pozisyon alınarak, hisse senetlerinin kapanış fiyatı oranının uzun dönem kapanış fiyatı oran ortalamasına dönmesi beklenmiştir. Bu beklenti gerçekleştiğinde, ilgili hisse senedi çiftlerinde alınan pozisyonlar kapatılmıştır. Böylece, hisse senedi çiftlerinde alınan her iki pozisyondan da elde edilen getiriler hesaplanmıştır.

Bankacılık sektöründe yer alan SKBNK/YKBNK çiftinde, eğitim verisi kullanılarak elde edilen sinyallerden bir tanesinde, pozisyon açıldıktan sonra kapatılma sinyali alınmıştır. İkinci pozisyon açma sinyali alındıktan sonra pozisyonun kapatılma sinyali henüz alınmış olduğundan kapatılmamıştır. Bu hisse senedi çiftinde, açılıp – kapatılan pozisyondan 471 YTL getiri elde edilmiştir. SKBNK/TEBNK çifti ise, tek bir pozisyon açma sinyali alınmıştır. Bu pozisyonun kapatılma sinyali ile birlikte elde edilen getiri 582 YTL olmuştur. Pozisyon kapatma sinyali alındığında, fiyatının düşeceği beklentisi ile kısa pozisyon alınan SKBNK hisse senedi, beklentinin aksine fiyatı artış göstermiş ve eksi getiri sağlamıştır. Ancak, çifti oluşturan TEBNK hisse senedinde, fiyatının artacağı beklentisi ile uzun pozisyon alınmış olduğundan, fiyat artışı gerçekleştiği için pozitif getiri sağlamıştır.

Tekstil sektöründeki yer alan AKIPD/ARAT çiftinde pozisyon açma sinyali alınmıştır, ancak hisse senedi çifti için henüz pozisyonu kapatma sinyali alınmamıştır. ALTIN/YUNSA çiftinde, iki kez pozisyon açma sinyali alınmıştır. Birinci pozisyonda -2σ limiti aşılmıştır. Bu pozisyonda, iyi performans gösteren YUNSA hisse senedinde kısa pozisyon alınmış, ALTIN hisse senedinde ise uzun

pozisyon alınmıştır. Pozisyonun kapatılma sinyali alındığında ise pozisyon kapatılmıştır, elde edilen getiri 416 YTL olmuştur. İkinci işlem sinyali ise $+2\sigma$ limiti aşarak alınmıştır. Ancak, henüz kapatılma sinyali vermemiştir. Bu sektörden elde edilen üçüncü hisse senedi çifti DERİM/VAKKO eğitim verisinde $\mp 2\sigma$ limitlerini hiç aşmamış, işlem sinyali vermemiştir. Bu hisse senedi çiftinin, EK–III de verilen Şekil EK – III.5 den de görülebileceği gibi, $\mp \sigma$ limitlerinde hareket ettiği görülmektedir. Dolayısıyla, bu hisse senedi çiftinden bir getiri elde edilemediğinden, portföy yönetiminin gereği olarak, portföyden çıkartılabilir. Ancak sistem, veri madenciliğinin veri içinde gizli kalmış yeni ilişkileri, eğilimleri ve örüntüleri ortaya çıkarma hedefinde beklenenden de fazla başarılı olabilir. Bu hisse senedi çifti için $\mp \sigma$ limitleri, çift için özel bir duruma işaret ediyor olabilir. Diğer tekstil sektörü hisse senedi çiftleri de, pozisyonlarını pozitif getiri ile kapatmıştır.

Oluşturulan veri madenciliği sisteminin elde ettiği portföy, pozitif getiri ile sonuçlanmıştır. Sistem uygun biçimde verilen veri içindeki muhtemel eksik veriyi giderdikten sonra, belirlenen ölçüte göre, dikkate alınan iki sektörden hisse senetlerini sektör içinde eşleyerek, gerekli dönüşümleri ve sinyal ölçütlerinin uygulanabilmesi için gerekli kapanış fiyatı oranlarını ile $\mp \sigma$ ve $\mp 2\sigma$ limitlerini hesaplamıştır. Ve son olarak, karar vericinin işlem zamanlamasını yapabilmesi için, hisse senedi çiftlerinin görsel sunumlarını yaparak karar sürecine uyarlamıştır.

Sistemi uygulamak için sadece bankacılık ve tekstil olmak üzere iki sektör dikkate alınmıştır. Burada diğer sektörlerde veriye dahil edilerek, sektörsel çeşitlendirme yapılabilir ve böylece daha iyi getiri sağlanabilir. Bunun yanında, sektör dahilindeki hisse senetlerinde de bir ayırım gözetilmemiştir, dolayısıyla finansal kriterler kullanılarak, çiftleri oluşturacak hisse senetleri de bir ön elemeden geçirilebilir. Ayrıca, portföye dahil edilen hisse senedi çiftlerinin miktarları sabit bir yatırım tutarı karşılığı olarak belirlenmiştir. Bu nedenle, portföyü oluşturan hisse senedi çiftlerinin miktarlarında da iyileştirmeye gidilebilir.

KAYNAKLAR

- [1] Zamar, R. H., *Introduction to Data Mining*, Ders Notları, <http://ugrad.stat.ubc.ca/~stat447j/resources/chapter1.pdf>, 2005.
- [2] Eker, H., *Veri Madenciliği veya Bilgi Keşfi*, http://www.bilgiyonetimi.org/cm/pages/mkl_gos.php?nt=538, 2005.
- [3] Wang, J., *Data Mining and Decision Support for Business and Science*, Encyclopedia Of Data Warehousing and Mining, Idea Group Reference, 2005.
- [4] McCUE, C., *Data Mining and Value-added Analysis*, FBI Law Enforcement Bulletin, http://www.findarticles.com/p/articles/mi_m2194/is_11_72/ai_111496582, 2005.
- [5] Brown, D. E., *The Regional Crime Analysis Program (RECAP): A Framework for Data Mining to Catch Criminals*, IEEE, 1998.
- [6] Luan, J., *Data Mining Applications in Higher Education*, SPSS White Papers, <http://www.spss.com>, 2005.
- [7] Boginski, V., Butenko, S. ve Pardalos, P. M., *Statistical Analysis of Financial Networks*, Article In Pres, Computational Statistical and Data Analysis, 2004.
- [8] Mittermayer, M. A., *Forecasting Intraday Stock Prize Trends with Text Mining Techniques*, Proceedings of the 37th Hawaii International Conference on System Sciences, 2004.
- [9] Sheng, Y. P., Hou, W. ve Chen, Z., *Mining for Profitable Patterns in the Stock Market*, Encyclopedia of Data Warehousing and Mining, Idea Group, 2006.
- [10] Wuthrich, B., Cho, V., Leung, S., Permuntilleke, D., Sankaran, K., Zhang, J. ve Lam, W., *Daily Stock Market Forecast from Textual Web Data*, IEEE International Conference on Systems, Man, and Cybernetics, **3**, P.2720 – 2725, 1998.
- [11] Fung, G. P. C., Yu, J. X. ve Lam, W., *Stock Predicting: Text Mining Approach using Real-Time News*, CIFE'03, Hong Kong.
- [12] Bellinger, G., Castro D. ve Mills, A., *Data, Information, Knowledge and Wisdom*, <http://www.systems-thinking.org/dikw/dikw.htm>, 2006.
- [13] Hey, J., *The Data, Information, Knowledge, Wisdom Chain: The Metaphorical Link*, Intergovernmental Oceanographic Commission, UNESCO, 2004.
- [14] Cabena, P., Hadjinian, P., Stadler, R., Verhees, J. ve Zanasi, A., *Discovering Data Mining*, Prentice Hall PTR, 1998.
- [15] Two Crows Corporation, *Introduction to Data Mining and Knowledge Discovery*, 3rd Edition, 1999.
- [16] Dunham, H. M., *Data Mining: Introductory and Advanced Topics*, Pearson Education Inc., 2003.

- [17] Friedman, J. H., *Data Mining and Statistics: What's the Connection?*, <http://www-stat.stanford.edu/~jhf/ftp/dm-stat.ps>, 2005.
- [18] IBM Red Boks, *Enhance Your Business Applications: Simple Integration of Advanced Data Mining Functions*, Aralık 2002.
- [19] Kudyba, S., Hoptroff, R., *Data Mining and Business Intelligence: A Guide To Productivity*, Idea Group Publishing, 2001.
- [20] Kantardzic, M., *Data Mining: Concepts, Models, Methods, And Algorithms*, John Wiley & Sons, Inc., 2003.
- [21] Berry, J. A. M., *Mastering Data Mining*, John Wiley & Sons Inc., 2000.
- [22] Giudici, P., *Applied Data Mining – Statistical Methods for Business and Industry*, John Wiley & Sons Ltd., 2003.
- [23] Benjamin, W., *Data Mining: Techniques, Issues and ROI*, Info-Tech White Paper, Info-Tech Research Group.
- [24] STATSOFT, <http://www.statsoft.com/textbook/stdatmin.html>, 2005.
- [25] Er, F., *Açıklayıcı Veri Analizi*, Kaan Kitapevi, 2003.
- [26] Han, J., Kamber, M., *Data Mining Concepts and Techniques*, Morgan Kaufmann Publishers, 2001.
- [27] Brown, M. L. ve Kros, J. F., *Data Mining and the Impact of Missing Data*, *Industrial Management & Data Systems*, **103** (8), P. 611–621, 2003.
- [28] Bal, C., *Çok Gruplu Veri Setlerinde Eksik Gözlem Sorununun Çözümlemesi ve Sağlık Alanında Bir Uygulama*, Doktora Tezi, Osmangazi Üniversitesi, Sağlık Bilimleri Enstitüsü, 2003.
- [29] Roderick, J. A. L. ve Rubin, D. B., *Statistical Analysis With Missing Data*, Second Edition, John Wiley & Sons Inc., 2002.
- [30] Witten, I. H. ve Frank, E., *Data Mining: Practical Machine Learning Tools and Techniques*, Second Edition, Elsevier, 2005.
- [31] Akkaya, Ş. ve Pazarlıoğlu, M. V., *Ekonometri II*, 2. Baskı, Erkam Matbaacılık, 1998.
- [32] Apaydın, A., Kutsal, A. ve Atakan, C., *Uygulamalı İstatistik*, Ankara, 1994.
- [33] Çömlekçi, N., *Deney Tasarımı ve Çözümlemesi*, TC. Anadolu Üniversitesi Eğitim, Sağlık ve Bilimsel Araştırma Çalışmaları Vakfı Yayınları, No: 58, Eskişehir, 1988.
- [34] Montgomery, D.C., *Design and Analysis of Experiments*, 5th. Edition, John Wiley & Sons Inc., 2001.
- [35] Muluk, Z., Kurt, S., Toktamış, Ö. ve Karaağaoğlu, E., *Deney Düzenlemede İstatistiksel Yöntemler*, Ege Üniversitesi Fen Fakültesi Yayınları, No: 146, Ege Üniversitesi Basım Evi, İzmir, 1994.
- [36] Şıklar, E., *Regresyon Analizine Giriş*, TC. Anadolu Üniversitesi Yayınları, No: 1255, 2000.

- [37] Özdamar, K., *Paket Programlar İle İstatistiksel Veri Analizi I*, 5. Baskı, Kaan Kitabevi, 2004.
- [38] Kalaycı, Ş., *SPSS Uygulamalı Çok Değişkenli İstatistik Teknikleri* (Ed: Kalaycı, Ş.), Asil Yayın Dağıtım Ltd. Şti., 2005.
- [39] Tatlıdıl, H., *Uygulamalı Çok Değişkenli İstatistiksel Analiz*, Akademi Matbaası, 1996.
- [40] Tan, P., Steinbach, M. ve Kumar, V., *Introduction to Data Mining*, Addison–Wesley, 2006.
- [41] Alataş, A. ve Akın, E., *Veri Madenciliğinde Yeni Yaklaşımlar*, Yöneylem Araştırması/Endüstri Mühendisliği – XXIV Ulusal Kongresi, 15–18 Haziran 2004.
- [42] Kurt, M., Semetay, C., *Genetik Algoritma ve Uygulama Alanları*, Mühendis ve Makine Dergisi, Sayı 501, Ekim, 2001.
- [43] Şentürk, S., *Deney Planlamasında Bulanık Mantık Yaklaşımı*, Doktora Tezi, Anadolu Üniversitesi, Fen Bilimleri Enstitüsü, 2007
- [44] Ceylan, A. ve Korkmaz, T., *Borsada Uygulamalı Portföy Yönetimi*, 3. Baskı, Ekin Kitabevi Yayınları, 1998.
- [45] Akay, D., Çetinyokuş, T. ve Dağdeviren, M., *Portföy Seçimi Problemi İçin KDS/GA Yaklaşımı*, Gazi Üniversitesi Mühendislik Mimarlık Fakültesi Dergisi, **17** (4), 125–138, 2002.
- [46] Atan, M., *Karesel Programlama İle Portföy Optimizasyonu*, VII. Ulusal Ekonometri ve İstatistik Kongresi, 26–27 Mayıs 2005.
- [47] Korkmaz, T., Pekkaya, M., *Excel Uygulamalı Finans Matematiği*, Erkin Kitabevi, 2005.
- [48] Sevil, G., *Finansal Risk Yönetimi Çerçevesinde Piyasa Volatilitésinin Tahmini ve Portföy VaR Hesaplamaları*, T.C. Anadolu Üniversitesi Yayınları, No. 1323, 2001.
- [49] Robert, J. E., Van Der Hoek, J. ve Malcolm, W. P., *Pairs Trading*, Quantitative Finance, **5** (3), 271–276, 2005.
- [50] Gatev, E., Goetzman, W. N. ve Rouwenhorst, K. G., *Pairs Trading: Performance of Relative –Value Arbitrage Rule*, The Review of Financial Studies, **19** (3), 2006.
- [51] Preston, T., *Pairs Trading*, http://www.traders-mag.co.uk/traders_team_aut.php?ID=42, 2005.
- [52] Anonim, *What is Pair Trading?*, <http://www.pairtrader.com/subject2.html>, 2005.
- [53] İMKB Yayınları, *Borsa Terimleri Sözlüğü*, 2000.
- [54] Wikipedia, http://en.wikipedia.org/wiki/Short_%28finance%29, 2006.
- [55] Block, F., *Pair Trading and Portfolio Construction*, http://www.insightful.com/insightful_doclib/document.asp?id=89, 2005.

- [56] Plato Veri Dağıtım Hizmetleri A.Ş., *İstanbul Menkul Kıymetler Borsası Veri Sağlama Hizmetleri*, <http://web.platodata.com.tr/msdata.php>, 2006.
- [57] Statistical Glossary, *Spline*, <http://www.statistics.com/resources/glossary/s/spline.php>, 2007.
- [58] R devolopement Core Team, *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, ISBN 3-900051-07-0, <http://www.R-project.org>, 2007.
- [59] Duncan, M. ve Chow, E. D., *ellipse: Funcitons for Drawing Ellipses and Ellipse-like Confidence Regions*, (porting to R by FRIAS, J. M. C.), R package version 0.3–4, 2006.

EK – I Tekstil Sektörü İçin Uzun Dönem İlişki Matrisi

	KORDS	IDAS	GEDİZ	ESEMS	EDİP	DESA	DERİM	CYTAS	CEYLN	BOSSA	BİSAS	BERDN	ATEKS	ARSAN	ARAT	ALTIN	AKİPD	AKALT
AKALT	-0,445	0,534	0,276	0,701	0,220	0,340	-0,022	0,715	-0,221	0,878	0,924	0,786	0,688	0,798	0,909	-0,063	0,755	1,000
AKİPD	-0,435	0,661	0,376	0,678	0,037	0,426	0,013	0,469	0,035	0,874	0,832	0,696	0,646	0,765	0,816	0,136	1,000	0,755
ALTIN	0,729	0,724	0,722	0,046	0,014	0,767	0,745	0,350	0,123	-0,147	-0,112	-0,296	0,526	0,366	-0,171	1,000	0,136	-0,063
ARAT	-0,552	0,418	0,132	0,741	0,072	0,242	-0,114	0,616	-0,235	0,876	0,955	0,740	0,569	0,769	1,000	-0,171	0,816	0,909
ARSAN	-0,056	0,751	0,540	0,706	0,090	0,683	0,463	0,727	-0,136	0,662	0,808	0,525	0,878	1,000	0,769	0,366	0,765	0,798
ATEKS	0,074	0,797	0,670	0,594	0,040	0,677	0,559	0,708	-0,124	0,547	0,652	0,447	1,000	0,878	0,569	0,526	0,646	0,688
BERDN	-0,737	0,282	0,159	0,505	0,292	-0,004	-0,336	0,371	-0,010	0,861	0,764	1,000	0,447	0,525	0,740	-0,296	0,696	0,786
BİSAS	-0,504	0,488	0,154	0,759	0,023	0,282	-0,043	0,641	-0,225	0,882	1,000	0,764	0,652	0,808	0,955	-0,112	0,832	0,924
BOSSA	-0,617	0,455	0,184	0,640	0,128	0,198	-0,224	0,455	-0,189	1,000	0,882	0,861	0,547	0,662	0,876	-0,147	0,874	0,878
CEYLN	-0,106	0,065	0,154	-0,191	0,424	-0,074	-0,115	1,000	-0,096	-0,189	-0,225	-0,010	-0,124	-0,136	-0,235	0,123	0,035	-0,221
CYTAS	0,065	0,682	0,453	0,501	0,273	0,588	0,292	1,000	-0,096	0,455	0,641	0,371	0,708	0,727	0,616	0,350	0,469	0,715
DERİM	0,744	0,497	0,518	0,074	-0,162	0,733	1,000	0,292	-0,115	-0,224	-0,043	-0,336	0,559	0,463	-0,114	0,745	0,013	-0,022
DESA	0,520	0,843	0,694	0,321	0,085	1,000	0,733	0,588	-0,074	0,198	0,282	-0,004	0,677	0,683	0,242	0,767	0,426	0,340
EDİP	-0,162	0,150	0,298	0,093	1,000	0,085	-0,162	0,273	0,424	0,128	0,023	0,292	0,040	0,090	0,072	0,014	0,037	0,220
ESEMS	-0,307	0,477	0,275	1,000	0,093	0,321	0,074	0,501	-0,191	0,640	0,759	0,505	0,594	0,706	0,741	0,046	0,678	0,701
GEDİZ	0,283	0,782	1,000	0,275	0,298	0,694	0,518	0,453	0,154	0,184	0,154	0,159	0,670	0,540	0,132	0,722	0,376	0,276
IDAS	0,240	1,000	0,782	0,477	0,150	0,843	0,497	0,682	0,065	0,455	0,488	0,282	0,797	0,751	0,418	0,724	0,661	0,534
KORDS	1,000	0,240	0,283	-0,307	-0,162	0,520	0,744	0,065	-0,106	-0,617	-0,504	-0,737	0,074	-0,056	-0,552	0,729	-0,435	-0,445
KOTKS	-0,724	0,145	-0,002	0,333	-0,025	-0,202	-0,561	0,191	-0,123	0,714	0,586	0,709	0,154	0,210	0,582	-0,348	0,574	0,528
KRTEK	-0,506	0,613	0,323	0,725	0,197	0,361	-0,095	0,660	-0,028	0,886	0,921	0,771	0,651	0,785	0,903	0,012	0,887	0,893
LUKSK	0,754	0,499	0,545	-0,059	-0,197	0,663	0,869	0,087	0,087	-0,311	-0,222	-0,422	0,430	0,286	-0,303	0,843	0,009	-0,255
MEMSA	-0,765	0,298	0,048	0,709	0,227	0,023	-0,385	0,456	0,028	0,858	0,870	0,821	0,397	0,567	0,888	-0,317	0,784	0,805
MNDRS	0,697	0,752	0,703	0,143	0,131	0,911	0,808	0,499	0,090	-0,085	0,029	-0,230	0,565	0,528	-0,016	0,859	0,195	0,112
MTEKS	0,704	0,610	0,716	0,013	-0,014	0,769	0,875	0,213	0,114	-0,207	-0,139	-0,274	0,509	0,416	-0,205	0,852	0,094	-0,094
OKANT	-0,113	0,686	0,472	0,660	0,160	0,639	0,345	0,756	-0,103	0,652	0,797	0,524	0,785	0,924	0,782	0,305	0,748	0,796
SKTAS	-0,061	0,633	0,465	0,342	-0,001	0,504	0,280	0,487	-0,246	0,643	0,583	0,584	0,710	0,666	0,468	0,320	0,605	0,606
SONMEZ	0,314	0,618	0,500	0,462	0,070	0,718	0,644	0,758	-0,350	0,315	0,502	0,185	0,761	0,769	0,451	0,443	0,341	0,576
UKİM	-0,509	0,199	0,088	0,624	0,176	0,030	-0,220	0,235	0,150	0,492	0,567	0,428	0,234	0,433	0,618	-0,180	0,570	0,459
VAKKO	0,795	0,070	0,175	-0,266	-0,312	0,404	0,821	-0,141	-0,152	-0,528	-0,410	-0,587	0,148	0,033	-0,472	0,551	-0,356	-0,431
YATAS	0,863	0,595	0,573	-0,062	-0,042	0,800	0,811	0,307	0,024	-0,349	-0,225	-0,481	0,364	0,258	-0,285	0,845	-0,078	-0,158
YUNSA	0,727	0,565	0,528	0,073	-0,163	0,743	0,914	0,280	-0,108	-0,146	-0,018	-0,293	0,586	0,467	-0,113	0,791	0,072	-0,010

EK - I Tekstil Sektörü İçin Uzun Dönem İlişki Matrisi (Devam)

	YUNSA	YATAS	VAKKO	UKIM	SONMEZ	SKTAS	OKANT	MTEKS	MNDRS	MEMSA	LUKSK	KRTEK	KOTKS	
AKALT	-0,010	-0,158	-0,431	0,459	0,576	0,606	0,796	-0,094	0,112	0,805	-0,255	0,893	0,528	AKALT
AKIPD	0,072	-0,078	-0,356	0,570	0,341	0,605	0,748	0,094	0,195	0,784	0,009	0,887	0,574	AKIPD
ALTIN	0,791	0,845	0,551	-0,180	0,443	0,320	0,305	0,852	0,859	-0,317	0,843	0,012	-0,348	ALTIN
ARAT	-0,113	-0,285	-0,472	0,618	0,451	0,468	0,782	-0,205	-0,016	0,888	-0,303	0,903	0,582	ARAT
ARSAN	0,467	0,258	0,033	0,433	0,769	0,666	0,924	0,416	0,528	0,567	0,286	0,785	0,210	ARSAN
ATEKS	0,586	0,364	0,148	0,234	0,761	0,710	0,785	0,509	0,565	0,397	0,430	0,651	0,154	ATEKS
BERDN	-0,293	-0,481	-0,587	0,428	0,185	0,584	0,524	-0,274	-0,230	0,821	-0,422	0,771	0,709	BERDN
BISAS	-0,018	-0,225	-0,410	0,567	0,502	0,583	0,797	-0,139	0,029	0,870	-0,222	0,921	0,586	BISAS
BOSSA	-0,146	-0,349	-0,528	0,492	0,315	0,643	0,652	-0,207	-0,085	0,858	-0,311	0,886	0,714	BOSSA
CEYLN	-0,108	0,024	-0,152	0,150	-0,350	-0,246	-0,103	0,114	0,090	0,028	0,087	-0,028	-0,123	CEYLN
CYTAS	0,280	0,307	-0,141	0,235	0,758	0,487	0,756	0,213	0,499	0,456	0,087	0,660	0,191	CYTAS
DERIM	0,914	0,811	0,821	-0,220	0,644	0,280	0,345	0,875	0,808	-0,385	0,869	-0,095	-0,561	DERIM
DESA	0,743	0,800	0,404	0,030	0,718	0,504	0,639	0,769	0,911	0,023	0,663	0,361	-0,202	DESA
EDIP	-0,163	-0,042	-0,312	0,176	0,070	-0,001	0,160	-0,014	0,131	0,227	-0,197	0,197	-0,025	EDIP
ESEMS	0,073	-0,062	-0,266	0,624	0,462	0,342	0,660	0,013	0,143	0,709	-0,059	0,725	0,333	ESEMS
GEDIZ	0,528	0,573	0,175	0,088	0,500	0,465	0,472	0,716	0,703	0,048	0,545	0,323	-0,002	GEDIZ
IDAS	0,565	0,595	0,070	0,199	0,618	0,633	0,686	0,610	0,752	0,298	0,499	0,613	0,145	IDAS
KORDS	0,727	0,863	0,795	-0,509	0,314	-0,061	-0,113	0,704	0,697	-0,765	0,754	-0,506	-0,724	KORDS
KOTKS	-0,466	-0,557	-0,711	0,371	-0,088	0,373	0,244	-0,490	-0,412	0,692	-0,515	0,636	1,000	KOTKS
KRTEK	-0,018	-0,157	-0,498	0,571	0,431	0,583	0,803	-0,069	0,131	0,890	-0,182	1,000	0,636	KRTEK
LUKSK	0,913	0,808	0,826	-0,261	0,397	0,243	0,162	0,887	0,770	-0,469	1,000	-0,182	-0,515	LUKSK
MEMSA	-0,341	-0,488	-0,687	0,650	0,171	0,356	0,615	-0,389	-0,219	1,000	-0,469	0,890	0,692	MEMSA
MNDRS	0,803	0,913	0,532	-0,089	0,631	0,327	0,501	0,875	1,000	-0,219	0,770	0,131	-0,412	MNDRS
MTEKS	0,849	0,851	0,700	-0,199	0,483	0,339	0,304	1,000	0,875	-0,389	0,887	-0,069	-0,490	MTEKS
OKANT	0,341	0,205	-0,081	0,449	0,714	0,577	1,000	0,304	0,501	0,615	0,162	0,803	0,244	OKANT
SKTAS	0,401	0,155	0,070	-0,063	0,556	1,000	0,577	0,339	0,327	0,356	0,243	0,583	0,373	SKTAS
SONMEZ	0,621	0,485	0,322	0,084	1,000	0,556	0,714	0,483	0,631	0,171	0,397	0,431	-0,088	SONMEZ
UKIM	-0,276	-0,265	-0,474	1,000	0,084	-0,063	0,449	-0,199	-0,089	0,650	-0,261	0,571	0,371	UKIM
VAKKO	0,807	0,661	1,000	-0,474	0,322	0,070	-0,081	0,700	0,532	-0,687	0,826	-0,498	-0,711	VAKKO
YATAS	0,788	1,000	0,661	-0,265	0,485	0,155	0,205	0,851	0,913	-0,488	0,808	-0,157	-0,557	YATAS
YUNSA	1,000	0,788	0,807	-0,276	0,621	0,401	0,341	0,849	0,803	-0,341	0,913	-0,018	-0,466	YUNSA

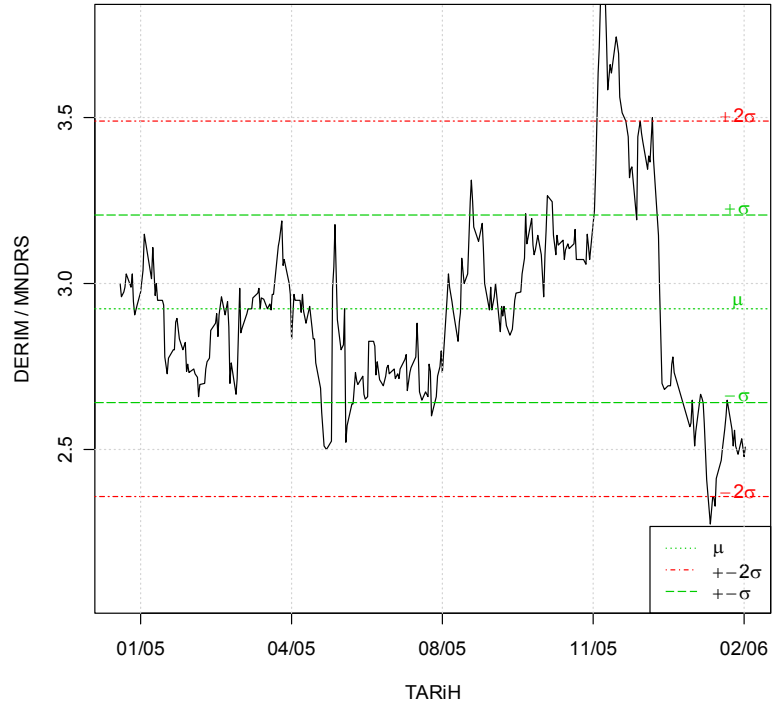
EK – II Tekstil Sektörü İçin Kısa Dönem İlişki Matrisi

	KORDS	IDAS	GEDİZ	ESEMS	EDİP	DESA	DERİM	CYTAS	CEYLN	BOSSA	BİSAS	BERDN	ATEKS	ARSAN	ARAT	ALTIN	AKİPD	AKALT
AKALT	0,651	0,553	0,781	0,727	0,715	-0,746	0,347	0,705	0,027	-0,052	-0,177	0,349	0,546	0,385	0,486	0,853	0,620	1,000
AKİPD	0,351	0,579	0,419	0,631	0,544	-0,564	-0,053	0,523	-0,469	0,255	-0,043	0,414	0,330	0,583	0,242	0,710	1,000	0,620
ALTIN	0,296	0,586	0,585	0,716	0,547	-0,550	0,418	0,671	-0,179	-0,088	-0,130	0,288	0,261	0,443	0,384	1,000	0,710	0,853
ARAT	0,539	0,000	0,119	0,373	0,040	-0,140	0,469	0,170	0,415	0,616	0,581	0,635	-0,096	0,716	1,000	0,384	0,242	0,486
ARSAN	0,535	-0,039	-0,124	0,279	-0,088	0,075	0,353	0,073	0,037	0,792	0,699	0,892	-0,282	1,000	0,716	0,443	0,583	0,385
ATEKS	0,258	0,517	0,884	0,535	0,918	-0,852	-0,379	0,529	0,085	-0,445	-0,651	-0,250	1,000	-0,282	-0,096	0,261	0,330	0,546
BERDN	0,618	-0,002	-0,107	-0,026	-0,174	0,176	0,488	0,070	0,021	0,786	0,722	1,000	-0,250	0,892	0,635	0,288	0,414	0,349
BİSAS	0,274	-0,514	-0,607	-0,217	-0,596	0,546	0,508	-0,388	0,207	0,892	1,000	0,892	-0,445	0,792	0,616	-0,130	-0,043	-0,177
BOSSA	0,453	-0,324	-0,448	-0,083	-0,381	0,289	0,263	-0,291	0,039	1,000	0,892	0,786	-0,445	0,792	0,616	-0,088	0,255	-0,052
CEYLN	0,240	-0,559	0,106	0,088	0,054	0,077	0,046	-0,373	1,000	0,039	0,207	0,021	0,085	0,037	0,415	-0,179	-0,469	0,027
CYTAS	0,092	0,876	0,619	0,521	0,536	0,172	1,000	1,000	-0,373	-0,291	-0,388	0,070	0,529	0,073	0,170	0,671	0,523	0,705
DERİM	0,281	0,003	-0,059	-0,039	-0,249	0,172	1,000	0,154	0,046	0,263	0,508	0,488	-0,379	0,353	0,469	0,418	-0,053	0,347
DESA	-0,391	-0,590	-0,875	-0,789	-0,940	1,000	0,172	-0,616	0,077	0,289	0,546	0,176	-0,852	0,075	-0,140	-0,550	-0,564	-0,746
EDİP	0,334	0,536	0,936	0,754	1,000	-0,940	-0,249	0,536	0,054	-0,381	-0,596	-0,174	0,918	-0,088	0,040	0,547	0,544	0,715
ESEMS	0,327	0,323	0,608	1,000	0,754	-0,789	-0,039	0,521	0,088	-0,083	-0,217	-0,026	0,535	0,279	0,373	0,716	0,631	0,727
GEDİZ	0,344	0,638	1,000	0,608	0,936	-0,875	-0,059	0,619	0,106	-0,448	-0,607	-0,107	0,884	-0,124	0,119	0,585	0,419	0,781
IDAS	-0,052	1,000	0,638	0,323	0,536	-0,590	0,003	0,876	-0,559	-0,324	-0,514	-0,002	0,517	-0,039	0,000	0,586	0,579	0,553
KORDS	1,000	-0,052	0,344	0,327	0,334	-0,391	0,281	0,092	0,240	0,453	0,274	0,618	0,258	0,535	0,539	0,296	0,351	0,651
KOTKS	-0,054	0,431	0,580	0,746	0,604	-0,560	0,200	0,611	-0,047	-0,535	-0,437	-0,248	0,406	-0,047	0,000	0,788	0,373	0,626
KRTEK	0,000	0,687	0,388	0,177	0,306	-0,349	0,355	0,636	-0,684	-0,404	-0,413	-0,021	0,199	-0,101	-0,296	0,623	0,425	0,485
LUKSK	0,812	0,394	0,680	0,385	0,596	-0,619	0,370	0,375	0,034	0,221	0,027	0,518	0,489	0,380	0,459	0,550	0,523	0,807
MEMSA	0,407	0,289	0,123	0,506	0,165	-0,211	0,502	0,473	-0,179	0,552	0,508	0,723	-0,014	0,812	0,603	0,673	0,702	0,580
MNDRS	0,337	0,662	0,841	0,372	0,639	-0,644	0,242	0,679	0,015	-0,506	-0,563	-0,032	0,588	-0,131	0,150	0,624	0,228	0,790
MTEKS	0,414	0,363	0,212	0,511	0,170	-0,246	0,478	0,553	-0,237	0,128	0,103	0,478	-0,079	0,616	0,429	0,834	0,593	0,723
OKANT	0,426	-0,293	-0,591	-0,177	-0,544	0,379	0,412	-0,162	-0,281	0,767	0,735	0,669	-0,606	0,640	0,316	-0,029	0,177	-0,030
SKTAS	0,221	0,808	0,655	0,656	0,639	-0,660	0,180	0,765	-0,430	-0,122	-0,300	0,221	0,410	0,343	0,225	0,920	0,848	0,795
SONMEZ	-0,044	-0,120	0,380	0,126	0,483	-0,419	-0,540	-0,152	0,191	-0,644	-0,698	-0,745	0,592	-0,754	-0,581	-0,296	-0,280	-0,076
UKİM	0,000	-0,091	0,494	0,081	0,501	-0,395	-0,523	-0,211	0,440	-0,591	-0,662	-0,637	0,583	-0,654	-0,326	-0,257	-0,321	-0,020
VAKKO	0,459	-0,366	-0,343	-0,370	-0,350	0,329	0,139	-0,433	0,202	0,866	0,755	0,696	-0,274	0,524	0,454	-0,376	-0,005	-0,184
YATAS	0,170	0,596	0,964	0,645	0,957	-0,871	-0,193	0,598	0,092	-0,571	-0,698	-0,289	0,927	-0,259	-0,044	0,517	0,373	0,666
YUNSA	0,457	-0,084	-0,293	-0,443	-0,440	0,458	0,468	-0,064	-0,050	0,616	0,626	0,879	-0,398	0,626	0,339	0,002	0,087	0,086

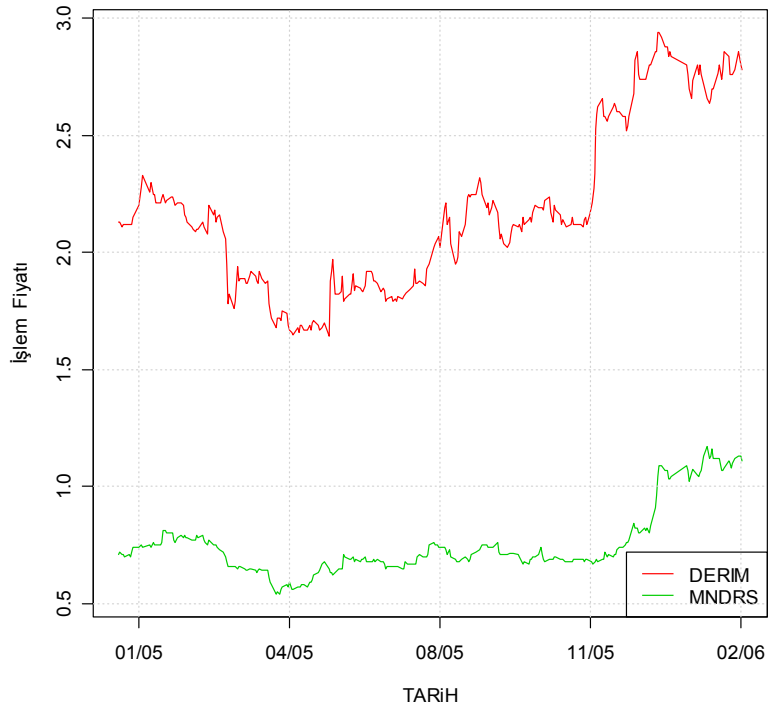
EK – II Tekstil Sektörü İçin Kısa Dönem İlişki Matrisi (Devam)

	YUNSA	YATAS	VAKKO	UKIM	SONMEZ	SKTAS	OKANT	MTEKS	MNDRS	MEMSA	LUKSK	KRTEK	KOTKS	
AKALT	0,086	0,666	-0,184	-0,020	-0,076	0,795	-0,030	0,723	0,790	0,580	0,807	0,485	0,626	AKALT
AKIPD	0,087	0,373	-0,005	-0,321	-0,280	0,848	0,177	0,593	0,228	0,702	0,523	0,425	0,373	AKIPD
ALTIN	0,002	0,517	-0,376	-0,257	-0,296	0,920	-0,029	0,834	0,624	0,673	0,550	0,623	0,788	ALTIN
ARAT	0,339	-0,044	0,454	-0,326	-0,581	0,225	0,316	0,429	0,150	0,603	0,459	-0,296	0,000	ARAT
ARSAN	0,626	-0,259	0,524	-0,654	-0,754	0,343	0,640	0,616	-0,131	0,812	0,380	-0,101	-0,047	ARSAN
ATEKS	-0,398	0,927	-0,274	0,583	0,592	0,410	-0,606	-0,079	0,588	-0,014	0,489	0,199	0,406	ATEKS
BERDN	0,879	-0,289	0,696	-0,637	-0,745	0,221	0,669	0,478	-0,032	0,723	0,518	-0,021	-0,248	BERDN
BISAS	0,626	-0,698	0,755	-0,662	-0,698	-0,300	0,735	0,103	-0,563	0,508	0,027	-0,413	-0,437	BISAS
BOSSA	0,616	-0,571	0,866	-0,591	-0,644	-0,122	0,767	0,128	-0,506	0,552	0,221	-0,404	-0,535	BOSSA
CEYLN	-0,050	0,092	0,202	0,440	0,191	-0,430	-0,281	-0,237	0,015	-0,179	0,034	-0,684	-0,047	CEYLN
CYTAS	-0,064	0,598	-0,433	-0,211	-0,152	0,765	-0,162	0,553	0,679	0,473	0,375	0,636	0,611	CYTAS
DERIM	0,468	-0,193	0,139	-0,523	-0,540	0,180	0,412	0,478	0,242	0,502	0,370	0,355	0,200	DERIM
DESA	0,458	-0,871	0,329	-0,395	-0,419	-0,660	0,379	-0,246	-0,644	-0,211	-0,619	-0,349	-0,560	DESA
EDIP	-0,440	0,957	-0,350	0,501	0,483	0,639	-0,544	0,170	0,639	0,165	0,596	0,306	0,604	EDIP
ESEMS	-0,443	0,645	-0,370	0,081	0,126	0,656	-0,177	0,511	0,372	0,506	0,385	0,177	0,746	ESEMS
GEDIZ	-0,293	0,964	-0,343	0,494	0,380	0,655	-0,591	0,212	0,841	0,123	0,680	0,388	0,580	GEDIZ
IDAS	-0,084	0,596	-0,366	-0,091	-0,120	0,808	-0,293	0,363	0,662	0,289	0,394	0,687	0,431	IDAS
KORDS	0,457	0,170	0,459	0,000	-0,044	0,221	0,426	0,414	0,337	0,407	0,812	0,000	-0,054	KORDS
KOTKS	-0,462	0,661	-0,776	0,046	0,150	0,668	-0,380	0,542	0,544	0,338	0,174	0,569	1,000	KOTKS
KRTEK	0,025	0,364	-0,521	-0,225	-0,038	0,697	0,000	0,533	0,577	0,268	0,340	1,000	0,569	KRTEK
LUKSK	0,321	0,504	0,283	0,058	-0,048	0,568	0,091	0,387	0,624	0,467	1,000	0,340	0,174	LUKSK
MEMSA	0,396	0,054	0,231	-0,721	-0,662	0,598	0,477	0,672	0,060	1,000	0,467	0,268	0,338	MEMSA
MNDRS	-0,055	0,734	-0,428	0,310	0,174	0,626	-0,395	0,445	1,000	0,060	0,624	0,577	0,544	MNDRS
MTEKS	0,307	0,104	-0,234	-0,516	-0,491	0,708	0,407	1,000	0,445	0,672	0,387	0,533	0,542	MTEKS
OKANT	0,671	-0,702	0,538	-0,745	-0,587	-0,097	1,000	0,407	-0,395	0,477	0,091	0,000	-0,380	OKANT
SKTAS	-0,042	0,593	-0,352	-0,205	-0,229	1,000	-0,097	0,708	0,626	0,598	0,568	0,697	0,668	SKTAS
SONMEZ	-0,661	0,508	-0,404	0,876	1,000	-0,229	-0,587	-0,491	0,174	-0,662	-0,048	-0,038	0,150	SONMEZ
UKIM	-0,578	0,549	-0,269	1,000	0,876	-0,205	-0,745	-0,516	0,310	-0,721	0,058	-0,225	0,046	UKIM
VAKKO	0,664	-0,481	1,000	-0,269	-0,404	-0,352	0,538	-0,234	-0,428	0,231	0,283	-0,521	-0,776	VAKKO
YATAS	-0,476	1,000	-0,481	0,549	0,508	0,593	-0,702	0,104	0,734	0,054	0,504	0,364	0,661	YATAS
YUNSA	1,000	-0,476	0,664	-0,578	-0,661	-0,042	0,671	0,307	-0,055	0,396	0,321	0,025	-0,462	YUNSA

EK – III Eş İşlem Strateji ile elde edilen Tekstil Sektörü Hisse Senedi Çiftleri



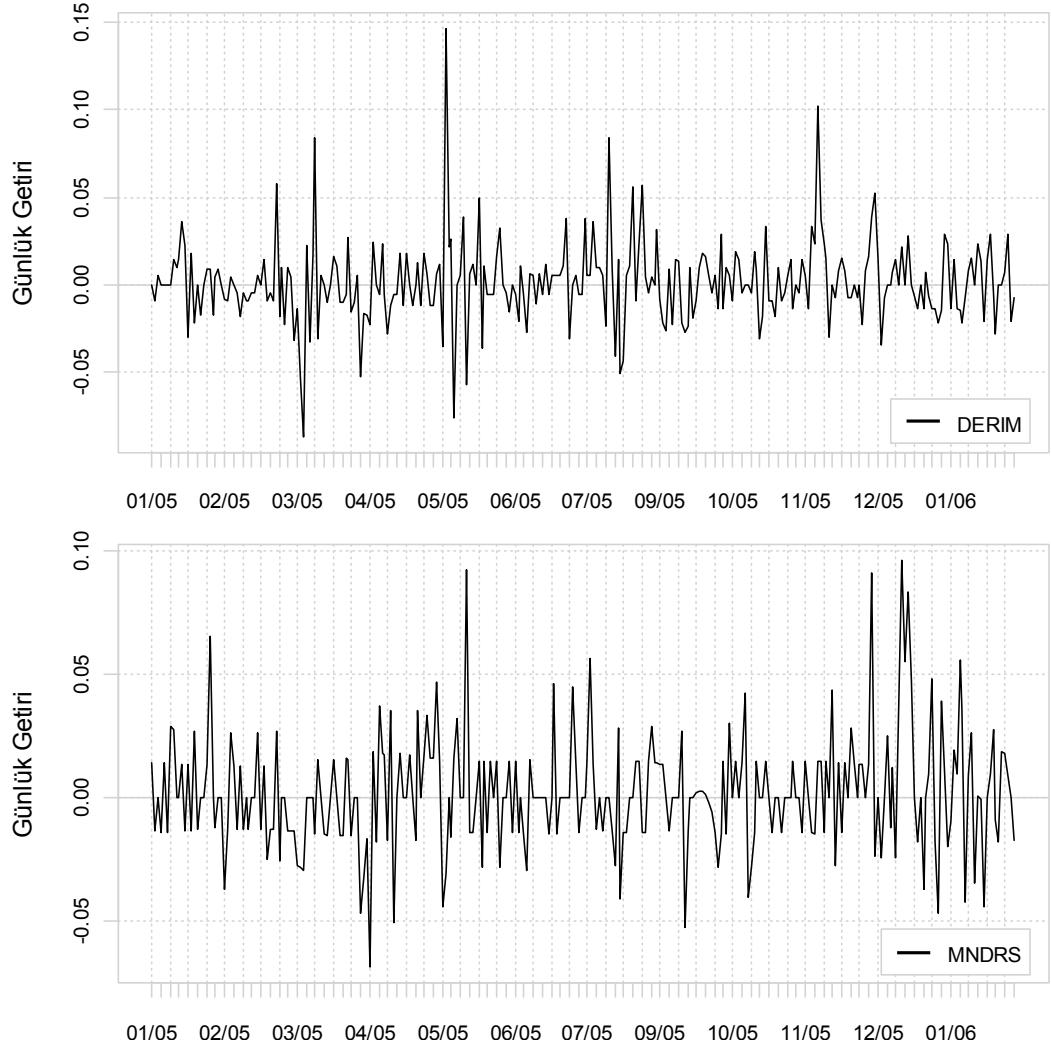
Şekil EK – III.1. DERİM/MNDRS hisse senedi çiftinin kapanış fiyat oran grafiği



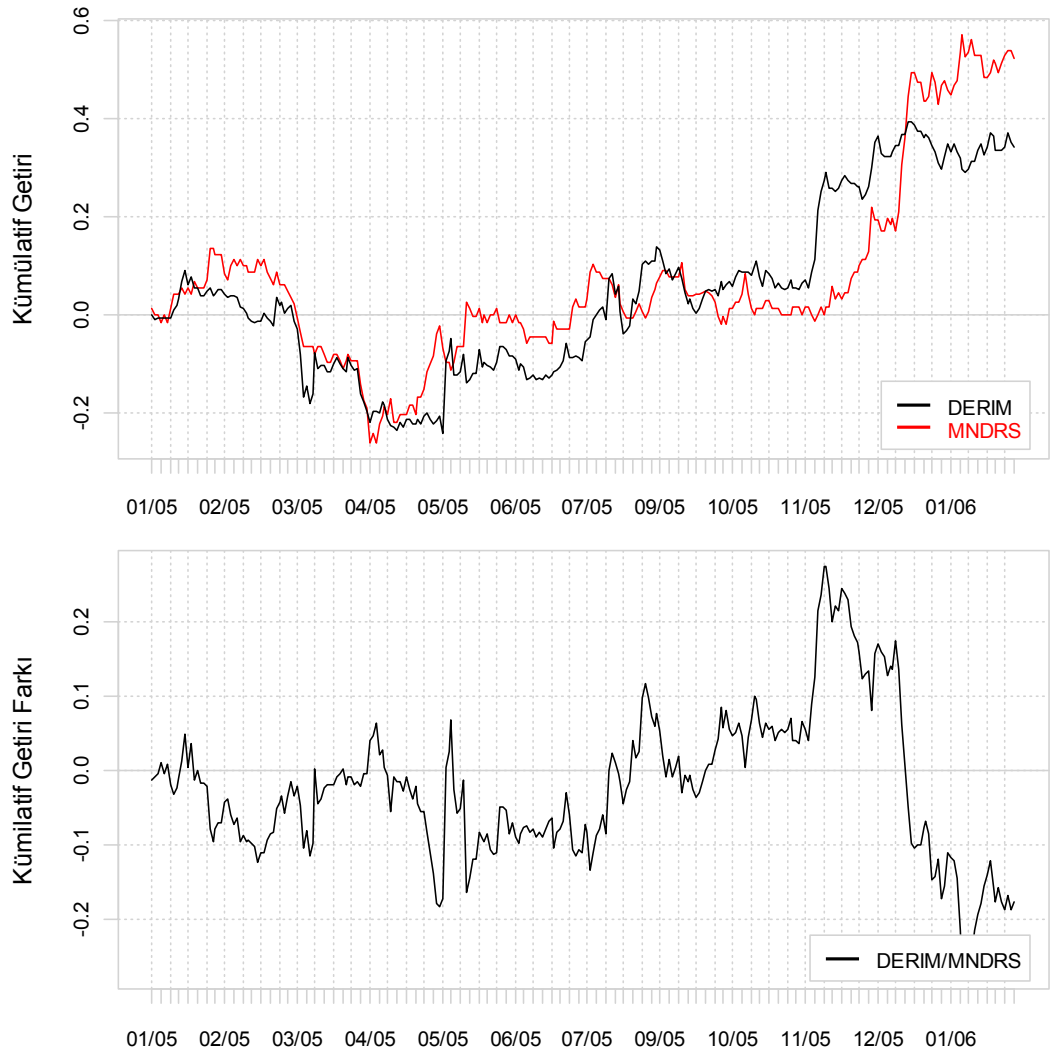
Şekil EK – III.2. DERİM/MNDRS hisse senedi çiftinin kapanış fiyatları grafiği

Çizelge EK – III.1. DERIM/MNDRS çiftinin getiri çizelgesi

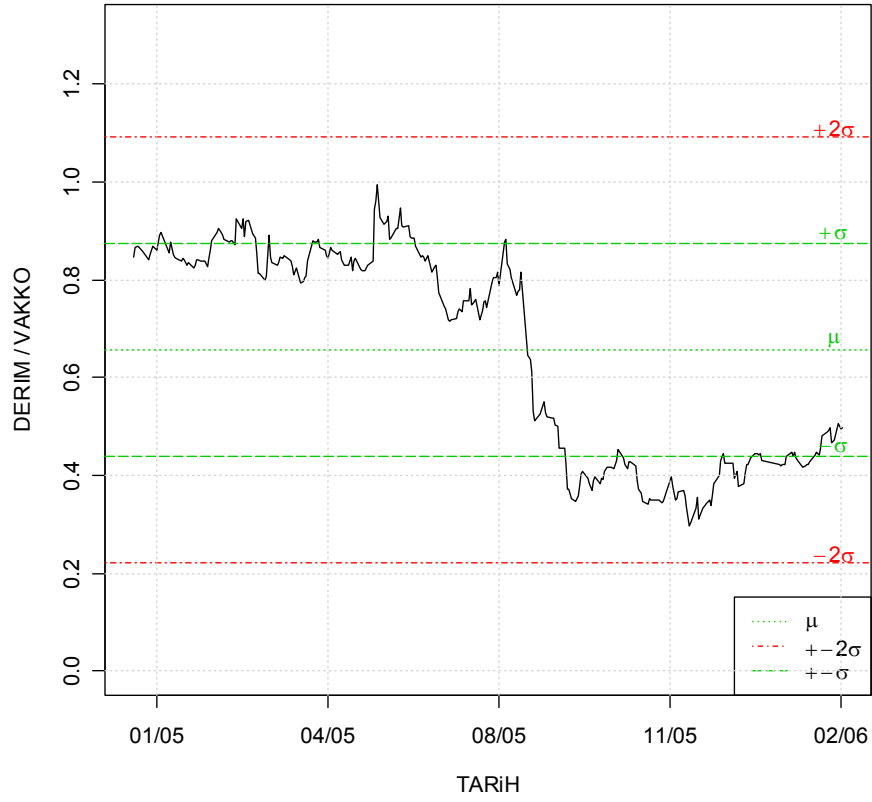
İşlem Tarihi		Uzun/Kısa Pozisyon	Tutar (YTL)		Adet		Getiri (YTL)	
Pozisyon Açma	16.11.2005	DERIM/MNDRS	1.000	1.000	-405	+1.471		
Pozisyon Kapatma	28.12.2005		1.190	1.529			-190	+529



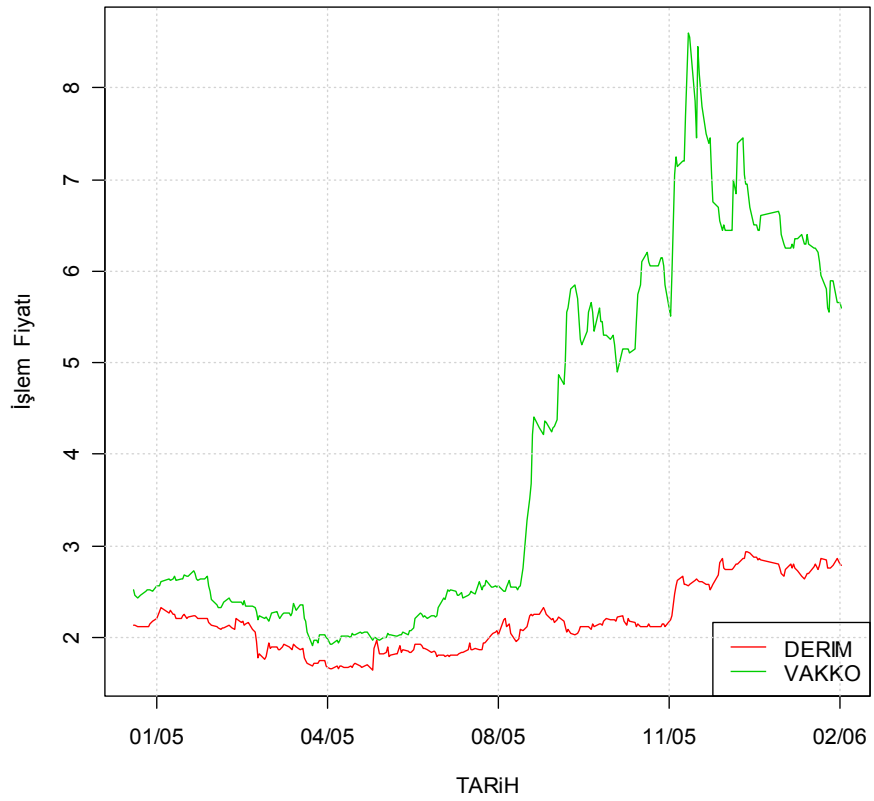
Şekil EK – III.3. DERIM/MNDRS hisse senedi çiftinin günlük getiri grafikleri



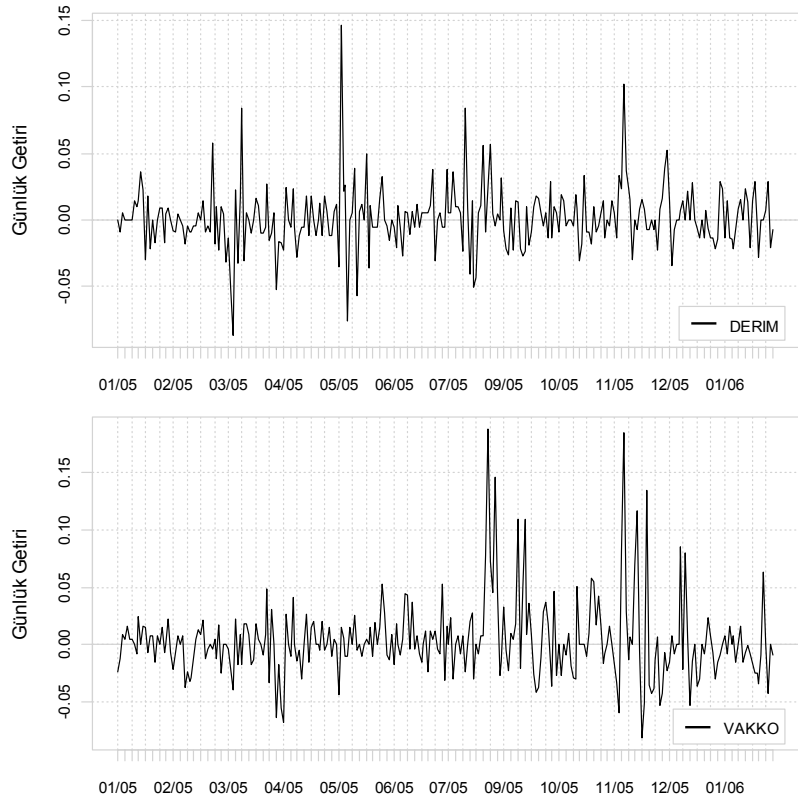
Şekil EK – III.4. DERIM/MNDRS hisse senedi çiftinin kümülatif getiri grafikleri



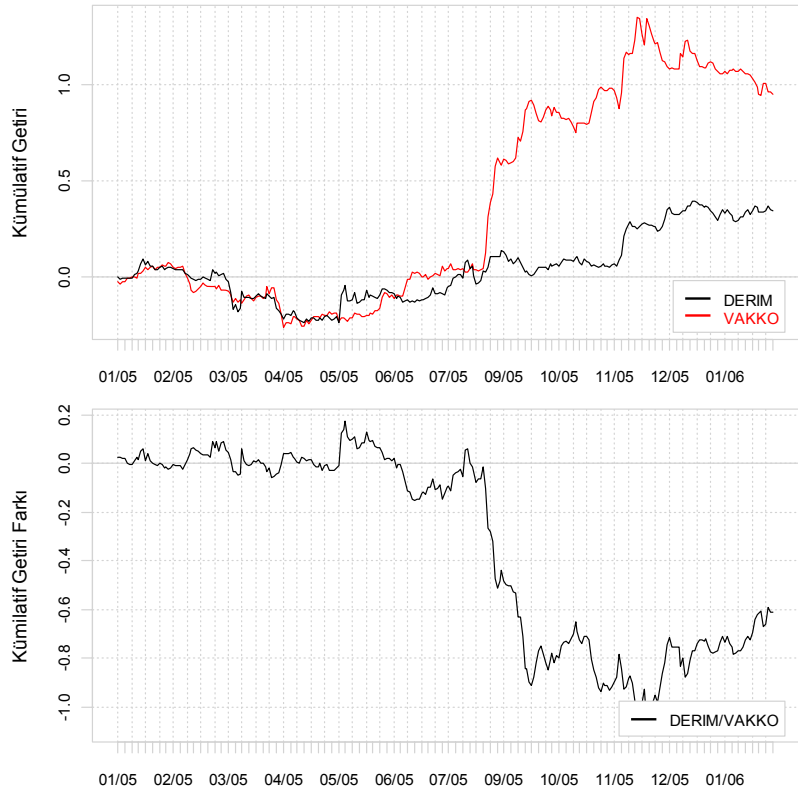
Şekil EK – III.5. DERİM/VAKKO hisse senedi çiftinin fiyat oranı grafiği



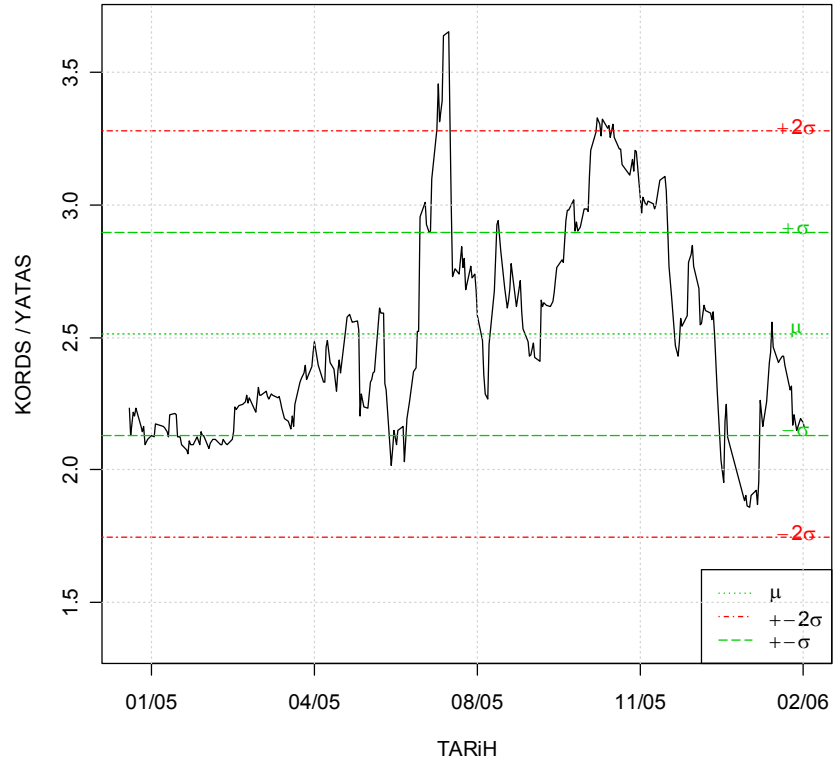
Şekil EK – III.6. DERİM/VAKKO hisse senedi çiftinin kapanış fiyatları grafiği



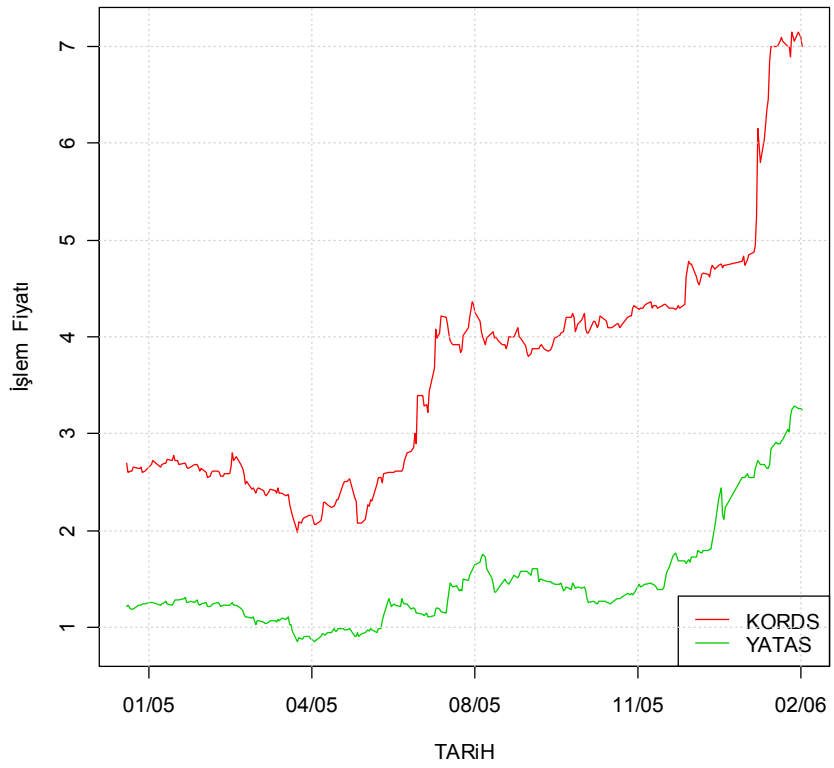
Şekil EK – III.7. DERIM/VAKKO hisse senedi çiftinin günlük getiri grafikleri



Şekil EK – III.8. DERIM/VAKKO hisse senedi çiftinin kümülatif getiri grafikleri



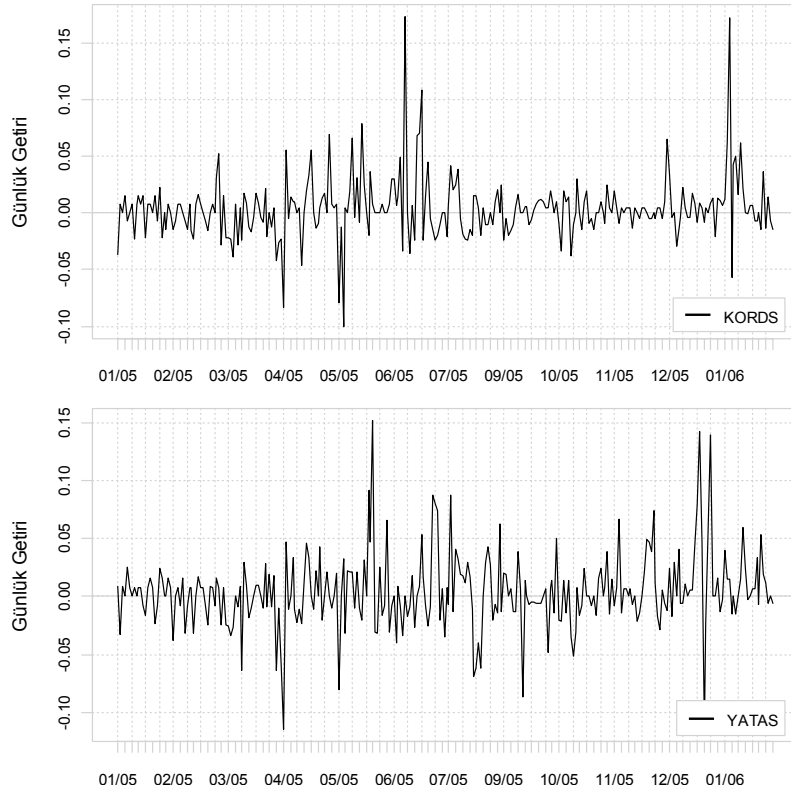
Şekil EK – III.9. KORDS/YATAS hisse senedi çiftinin fiyat oranı grafiği



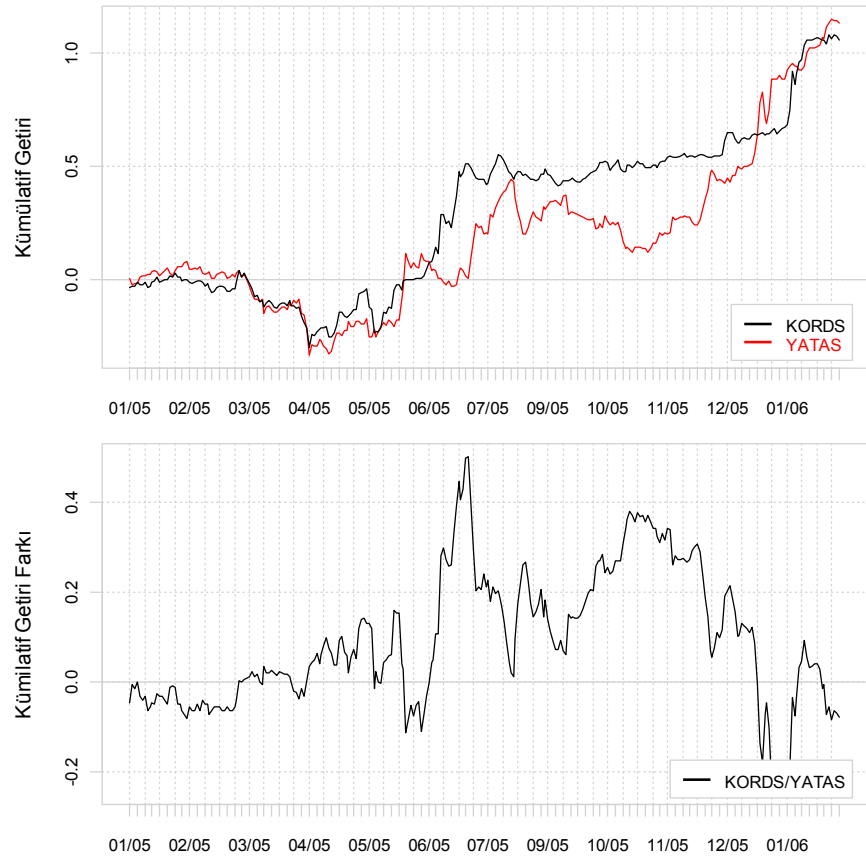
Şekil EK – III.10. KORDS/YATAS hisse senedi çiftinin kapanış fiyatı grafiği

Çizelge EK – III.2. KORDS/YATAS çiftinin getiri çizelgesi

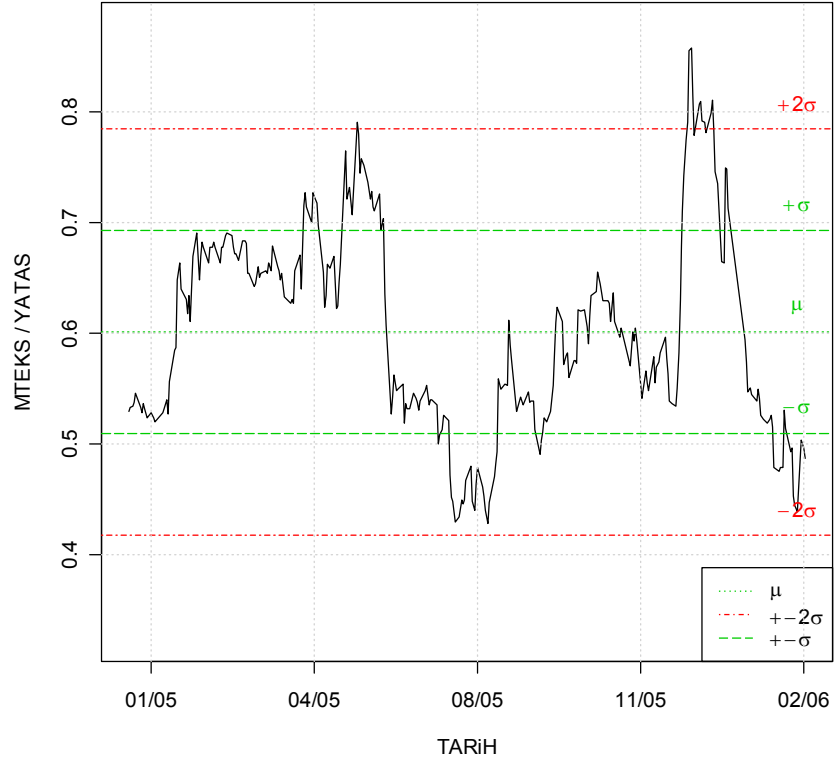
İşlem	Tarih	Hisse Senetleri	Tutar (YTL)		Adet		Getiri (YTL)	
Pozisyon Açma	11.07.2005	KORDS/YATAS	1.000	1.000	-272	+892		
Pozisyon Kapatma	08.08.2005		1.130	1.491			-130	+491
Pozisyon Açma	18.10.2005	KORD/YATAS	1.000	1.000	-240	+800		
Pozisyon Kapatma	05.12.2005		1.034	1.392			-34	+392



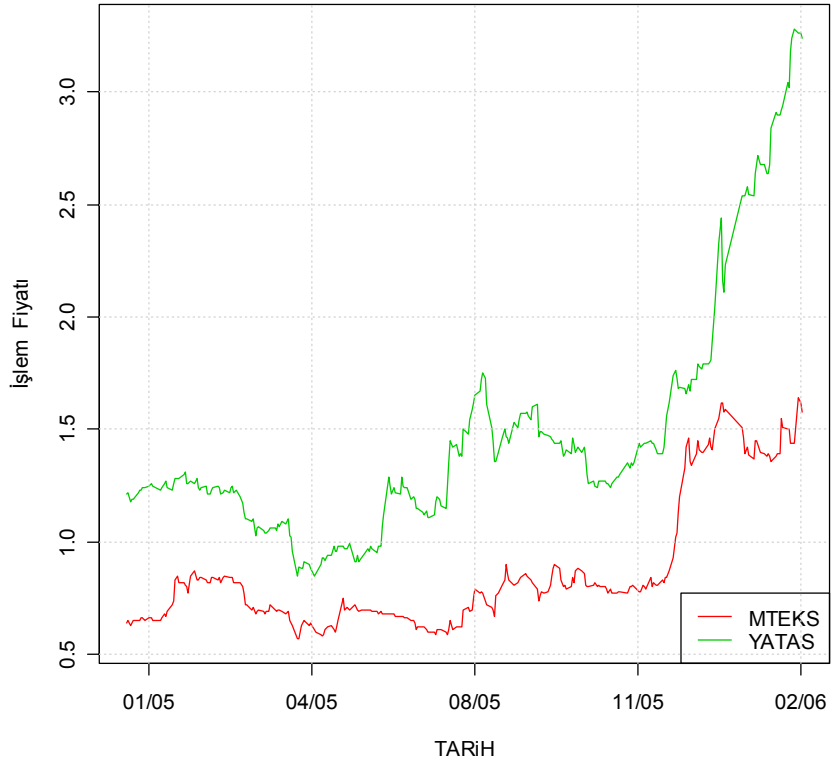
Şekil EK – III.11. KORDS/YATAS hisse senedi çiftinin günlük getiri grafikleri



Şekil EK – III.12. KORDS/YATAS hisse senedi çiftinin kümülatif getiri grafikleri



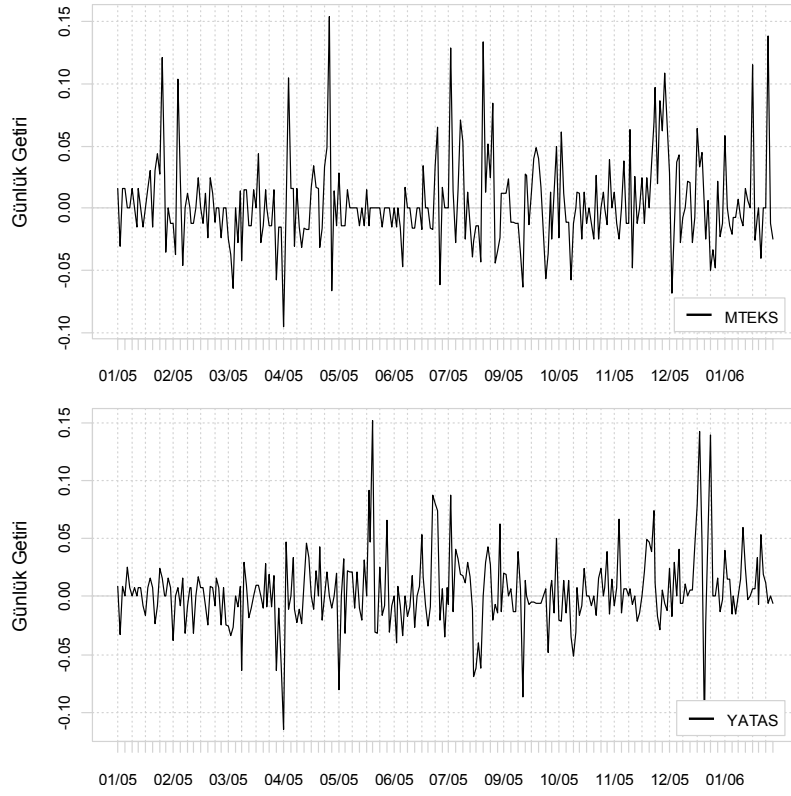
Şekil EK – III.13. MTEKS/YATAS hisse senedi çiftinin fiyat oranı grafiği



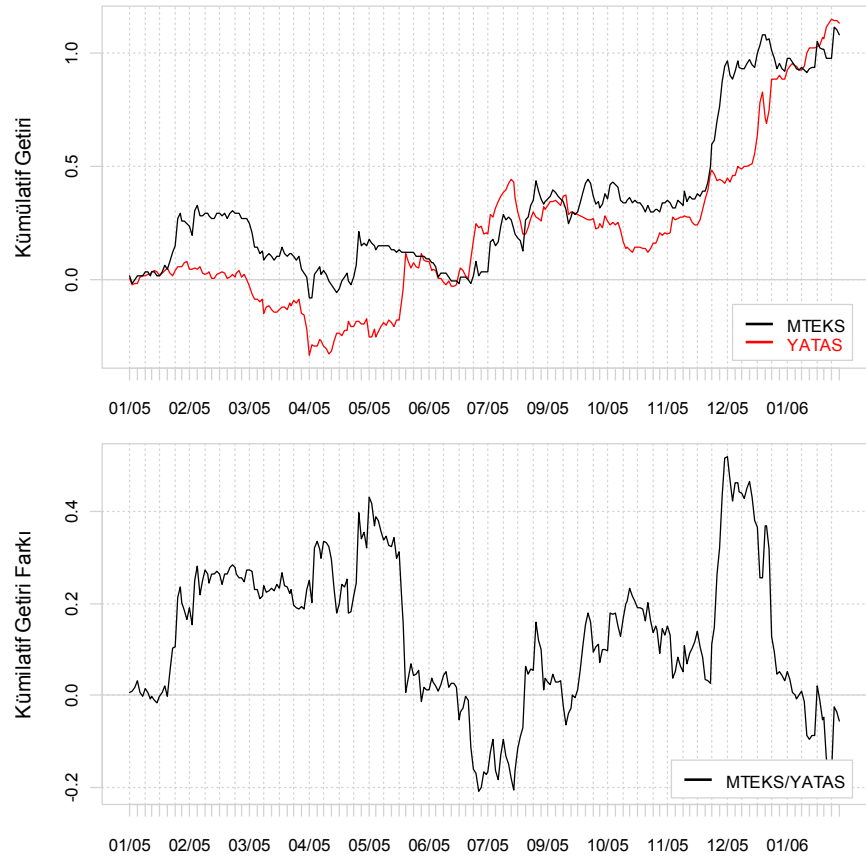
Şekil EK – III.14. MTEKS/YATAS hisse senedi çiftinin kapanış fiyatı grafiği

Çizelge EK – III.3. MTEKS/YATAS çiftinin getiri çizelgesi

İşlem	Tarih	Hisse Senetleri	Tutar (YTL)		Adet		Getiri (YTL)	
Pozisyon Açma	23.05.2005	MTEKS/YATAS	1.000	1.000	-1.389	+1099		
Pozisyon Kapatma	13.06.2005		944	1.418			+54	+418
Pozisyon Açma	12.12.2005	MTEKS/YATAS	1.000	1.000	-752	+599		
Pozisyon Kapatma	16.01.2006		1.135	1.512			-135	+512



Şekil EK – III.15. MTEKS/YATAS hisse senedi çiftinin günlük getiri grafikleri



Şekil EK – III.16. MTEKS/YATAS hisse senedi çiftinin kümülatif getiri grafikleri