

T.C.
MARMARA ÜNİVERSİTESİ
BANKACILIK VE SİGORTACILIK ENSTİTÜSÜ
SİGORTACILIK BÖLÜMÜ

SİGORTACILIK SEKTÖRÜNDE MÜŞTERİ
İLİŞKİLERİ YÖNETİMİ YAKLAŞIMIYLA
VERİ MADENCİLİĞİ TEKNİKLERİ
VE
BİR UYGULAMA

YÜKSEK LİSANS TEZİ

Evren KASAP

Danışman Prof.Dr. Münevver ÇETİN

İSTANBUL 2007

ÖZET

SİGORTACILIK SEKTÖRÜNDE MÜŞTERİ İLİŞKİLERİ YÖNETİMİ YAKLAŞIMIYLA VERİ MADENCİLİĞİ TEKNİKLERİ

Sigortacılık sektöründe müşteri ilişkileri yönetimi (CRM) ve veri madenciliği çok yeni bir alan olmasına karşın bankacılık sektöründe ve büyük işletmelerde uygulanması daha önceki yıllara dayanmaktadır. Bir ürünün hedef kitlesini belirlemekte kullanılan bu yöntemler iki ana bilim dalının birleştirilmesinden oluşmaktadır.

Müşteri ilişkileri yönetimi (CRM) işletmelerde müşterilerin memnuniyeti esas alınarak yapılan bir grup stratejiden oluşmaktadır. Bu stratejilerin esas amacı maliyeti azaltarak daha fazla satış yapabilmektir. Bu düşünce üzerine kurulan müşteri ilişkileri , teknik analiz desteğini ise istatistik biliminden almışlardır. Verilerin içerisinde istenen amaca uygun verilerin tespiti ve verilerin analizini istatistiksel analizler yardımıyla yapmaktadırlar. Bu teknik kısma veri madenciliği adı verilmiştir.

Özellikle bankacılık sektöründe başarıyla uygulanan müşteri ilişkileri yönetimi (CRM) ve veri madenciliği sigortacılık sektöründe de uygulamaya başlanmıştır. Sektörün ihtiyaçları doğrultusunda birçok sigorta poliçesinin , müşterilerin tercih ve daha önce satın aldıkları poliçeler dikkate alınarak müşteri memnuniyetini artıracak, müşteri bağlılığını güçlendirecek uygun poliçeler önerilerek başarı sağlanmaya çalışılmaktadır. Bu teknik analiz kısımlarının elde edilebilmesi için yüksek teknolojiye ihtiyaç duymaktadır.

Bu çalışma üç bölümden oluşmaktadır. İlk bölümde müşteri ilişkileri yönetimi (CRM) hakkında gerekli bilgiler verilerek yapılan teknik analizlerin amaçları vurgulanacaktır. Ayrıca müşteri memnuniyeti, müşteri sadakati vb. tanımları detaylı bir şekilde açıklanacaktır.

ikinci bölümde ise veri madenciliği hakkında gerekli bilgileri verirken ,burada kullanılan analiz yöntemleri hakkında kapsamlı bir bilgi verilecektir.Temeline istatistiksel analizlerden oluşan veri madenciliği,verilerin gruplanması ya da ayıklanması için kullanılmaktadır.

Üçüncü ve son bölümde ise belirli bir veri setine , veri madenciliği teknikleri uygulanacaktır.Uygulanan veri madenciliği teknikleri sonucunda , veri seti sınıflandırılacak , kümelelere ayrılacak , belirli ilişki modelleri kurulacaktır.Bu bölümde ayrıca bir işletmenin müşterileri hakkında detaylı bilgiyi elde edilerek , satış kampanyaları için hedef kitleler oluşturulacaktır.

İÇİNDEKİLER

ÖZET.....	i
İÇİNDEKİLER.....	iii
ŞEKİL LİSTESİ	v
TABLO LİSTESİ.....	vii
GİRİŞ.....	viii

BÖLÜM 1 MÜŞTERİ İLİŞKİLERİ YÖNETİMİ (CRM).....	1
1.1.Müşteri İlişkileri Yönetiminin Tanımı	1
1.2. Müşteri İlişkileri Yönetiminin Temel Kavramları.....	4
1.2.1.Müşteri İlişkileri Yönetimi Süreci.....	4
1.2.2.Müşteri İlişkileri Yönetiminin Bileşenleri.....	7
1.2.3.Müşteri İlişkileri Yönetiminin Yapıtaşları.....	7
1.2.4.Müşteri İlişkileri Yönetiminin Mimarisi	9
1.2.5.Müşteri İlişkileri Yönetiminin Teknolojileri	10
1.3.Müşteri Yöntemi İle İlgili Diğer Kavramlar	12
1.3.1.Müşteri Tanımı	12
1.3.2.Müşteri Sadakati Kavramı.....	12
1.3.3.Müşteri Tatmini Kavramı.....	14
1.4. Müşteri İlişkileri Yönetiminin Amaçları ve Sağladığı Avantajlar.....	15
1.5. Müşteri İlişkileri Yönetimi Projelerinin Başarılı ve Başarısız Olma Nedenleri.....	17
1.6. Müşteri İlişkileri Yönetiminde Kampanya Yönetimi.....	18
1.6.1.Kampanyanın Planlanması.....	20
1.6.2.Kampanyanın Hayata Geçirilmesi	22
1.6.3.Kampanya Analizi ve Kontrolü	23
BÖLÜM 2 VERİ MADENCİLİĞİ (DATA MINING).....	24

2.1.Verİ MadencilİĐİ Tanımı	24
2.2.Verİ MadencilĐİ Sürecinde Ortaya Çıkabilecek Sorunlar	27
2.3.Verİ Ambarı Kavramı	29
2.4.Verİ MadencilİĐİ Teknikleri	35
2.4.1.Tanımlama ve ayırılma	36
2.4.2.Birlikteli Kuralları	39

2.4.3.Sınıflama ve Öngörü	41
2.4.3.1.Karar Ağaçları	41
2.4.3.2.Yapay Sinir Ağları	43
2.4.3.3. K-Ortalamlar Yöntemi	46
2.4.3.4.Genetik Algoritmalar.....	46
2.4.3.5.Regresyon Analizi	47
2.4.3.6.Zaman Serileri Analizi	53
2.4.4.Kümeleme Analizi	66
2.4.4.1.Kümeleme Methodları.....	70
2.4.5.Sıradışılık Analizi.....	72
2.5.Veri Madenciliğinde Kullanılan Diğer Analiz Teknikleri	73
2.5.1.Faktör Analizi.....	73
2.5.2.Kanonik Kolerasyon Analizi	82
2.5.3. Lojistik Regresyon Analizi.....	85
2.5.4.Çok Boyutlu Ölçkleme.....	98
 BÖLÜM 3 SİĞORTACILIK SEKTÖRÜNDE MÜŞTERİ İLİŞKİLERİ YÖNETİMİ YAKLAŞIMIYLA BİR VERİ MADENCİLİĞİ UYGULAMASI	101
3.1.Araştırmanın Amacı.....	101
3.2.Araştırmanın Kapsamı	101
3.3.Araştırma İle İlgili Uygulama	105
3.3.1. Birliktelik Kuralları Analizi	105
3.3.2. Sınıflandırma Analizleri	109
3.3.3.Kümeleme Analizleri	114
SONUÇ.....	122
KAYNAKÇA	125

ŞEKİL LİSTESİ

Şekil 1.1. Müşteri ilişkileri yönetimi süreci.....	5
Şekil 1.2. Müşteri ilişkileri yönetimi mimarisi.....	9
Şekil 1.3. CRM kampanya yönetim süreci.....	19
Şekil 2.1. Veri tabanlarında bilgi keşfi aşamaları.....	25
Şekil 2.2 .Veri madenciliğinin tarihsel gelişimi.....	29
Şekil 2.3. Veri ambarı mimarisi.....	32
Şekil 2.4. veri ambarı bileşenleri.....	33
Şekil 2.5. Metadata ‘nın veri madenciliğindeki yeri	34
Şekil 2.6. Karar ağacı şekli.....	42
Şekil 2.7. Yapay ağ süreci.....	44
Şekil 2.8. Yapay sinir ağlarının katmanları.....	45
Şekil 2.9. Doğrusal regresyon grafiği.....	48
Şekil 2.10. Eğrisel ilişkiyi gösteren serpilme diyagramı	51
Şekil 2.11. Doğrusal ilişkiyi gösteren serpilme diyagramı.....	51
Şekil 2.12. Zaman serisi trendleri.....	54
Şekil 2.13. Konjonktürel dalgalanma grafiği.....	55
Şekil 2.14. Mevsimsel dalgalanma.....	56
Şekil 2.15. Hareketli ortalama yöntemiyle oluşturulan trend.....	58
Şekil 2.16. Sınıflandırma ağacı.....	66
Şekil 2.17. Benzerlik matrisi.....	67
Şekil 2.18. Veri ölçeleri sınıflandırması.....	68
Şekil 2.19. Kümeleme metodları hiyerarşisi	71
Şekil 2.20. Konanik kolerasyon grafiği.....	83
Şekil 2.21. S ve ters S fonksiyonu.....	87
Şekil 3.1. C 4.5 analizi uygulama süreci	112
Şekil 3.2. C 4.5 analizi sonucu elde edilen karar ağacı diyagramı.....	112

Şekil 3.3. K-ortalamalar analizi uygulama süreci	114
Şekil 3.4. K-ortalamalar analizi sonucu elde edilen kümelerin grafiksel görüntüsü.....	115
Şekil 3.5. K-ortalamalar analizi sonucu elde edilen kümelerin grafiksel görüntüsü.....	116
Şekil 3.6. Elde edilen vektörlerle arasındaki ilişkiyi gösteren grafik.....	119
Şekil 3.7. z_1 vektörü ile satın alma miktarı arasındaki ilişkiyi gösteren grafik.....	120
Şekil 3.8. z_2 vektörü ile satın alma miktarı arasındaki ilişkiyi gösteren grafik.....	121

TABLO LİSTESİ

Tablo 2.1.Anova testi hesap tablosu.....	48
Tablo 2.2. Hareketli ortalamaların hesaplanma yöntemi	58
Tablo 2.3.Skorlama katsayısı matrisi hesaplanması.....	78
Tablo 2.4. Değişkenlerin faktöre katkı oranları hesaplanması.....	78
Tablo 2.5.Değişken etki puanlarının hesaplanması.....	79
Tablo 3.1. Veri seti tablosu.....	101
Tablo 3.2.Cinsiyet değişkeninin özet bilgileri.....	102
Tablo 3.3.Sigorta bedeli değişkeninin özet bilgileri.....	102
Tablo 3.4.Primler değişkeninin özet bilgileri.....	103
Tablo 3.5.Sigorta cinsi değişkeninin özet bilgileri.....	103
Tablo 3.6.Semt değişkeninin özet bilgileri.....	104
Tablo 3.7.Yangın sigortası içindeki kadın-erkek oranı tablosu.....	106
Tablo 3.8.Kaza sigortası içindeki kadın-erkek oranı tablosu.....	107
Tablo 3.9.Konut sigortası içindeki kadın-erkek oranı tablosu.....	108
Tablo 3.10. C&RT analizi sonuç matrisi ve hata oranları.....	109
Tablo 3.11. C&RT analizinin yaprak analiz sonuçları.....	110
Tablo 3.12. C&RT analizi sonucunda elde edilen karar ağacı diyagramı.....	110
Tablo 3.13. K-ortalamalar analizi sonucu elde edilen kümeler.....	114
Tablo3.14. K-ortalamalar analizi sonucu elde edilen kümelerin veri listesi.....	117
Tablo 3.15. Kanonikel diskriminant analizi sonuçları	117
Tablo 3.16. Kanonikel diskriminant analizi sonucu elde edilen vektörler.....	118
Tablo 3.17. Elde edilen vektörlerle değişkenler arasındaki ilişki.....	118

GİRİŞ

Veri madenciliği ve müşteri ilişkileri yönetimi temel olarak veri setine bağlıdır. Veri olmaması durumunda bu tekniklerin geçerliliği ve uygulanabilirliği söz konusu değildir. Ancak günümüz şartlarında çok yoğun bir veri akışı söz konusudur. Veri madenciliği kavramıda tam olarak buradan çıkmaktadır. Çünkü veri madenciliği , çok sayıda veriden amacına uygun verilerinin elde edilme sürecidir. Günümüz şartlarında şirketlerin , doğru hedefe yönelmesinin başarıyı getireceği bilinmektedir.

Bu çalışmanın amacı son yıllarda yaygın olarak kullanılan müşteri ilişkileri yönetimi ve veri madenciliği teknikleri hakkında bilgi vermeyi amaçlamaktadır. Bu amaçla ilk iki bölümde teorik bilgi verildikten sonra üçüncü bölümde sigortacılık sektörü verileri kullanılarak bir örnek uygulama yapılmıştır.

Teknolojik gelişmelerin özellikle iş hayatına yansımaları sonucunda satış ve pazarlama bölümlerindeki gelişmelerin bir sonucu olarak ortaya çıkan müşteri ilişkileri ve veri madenciliği yöntemleri , çok sayıda müşteri verilerini analiz ederek satışları artmasına olanak sağlamıştır.

Yoğun rekabet şartları altında daha başarılı olmanın müşterilerini daha iyi tanımaktan geçtiği bir dönemde müşteri ilişkileri yönetiminde veri madenciliği buna olanak sağlamaktadır.

Ayrıca veri madenciliği teknikleri sağlık , bankacılık , astroloji gibi bir çok alanda da kullanılmaktadır.

Ülkemizde veri madenciliği ve müşteri ilişkileri yönetimi hakkında yeterli kaynak bulunmamaktadır. Bu nedenle yapılan çalışma özellikle sigortacılık sektörüne yapacağı katkı açısından önem arz etmektedir.

BÖLÜM 1 MÜŞTERİ İLİŞKİLERİ YÖNETİMİ (CRM)

1.1.MÜŞTERİ İLİŞKİLERİ YÖNETİMİ (CRM)'İN TANIMI

Müşteri ilişkileri yönetimi (CRM)'in tek bir tanımı olmaktan ziyade birden fazla tanımı vardır.Bunlar şöyle ifade edilebilir .

- Müşteri ilişkileri yönetimi (CRM)'in satış, pazarlama ve hizmet süreçlerinin müşteri odaklı bir felsefe etrafında yeniden tanımlanmasını gerektiren bir süreçtir.¹
- Müşteri ilişkileri yönetimi (CRM)'in , müşteri memnuniyetini kar maksimizasyonuna dönüştürmek amacı taşıyan, istediği müşteriye istediği deneyimi yaşatabilecek kabiliyette bir kurum felsefesi ve bu hedefe ulaşmak için ; gerekli insan, süreç , teknoloji yapılanmasıdır.²
- Müşteri ilişkileri yönetimi (CRM), tüketiciyi ve onların ihtiyaçlarını doğru analiz ederek, eğilimleri ve ilgileri konusunda doğru sonuçları çıkarabilmek ve tüm bunları ortaya koyduktan sonra onların isteklerine uygun ürünlere ulaştırabilmektir.
- Müşteri ilişkileri yönetimi (CRM), müşterileri belli gruplara ayırarak , onların davranış modellerini oluşturan ve müşterilerin ihtiyaçlarını tesbit ederek bu ihtiyaçların karşılamasını amaçlayan bir süreçtir.³
- Müşteri ilişkileri yönetimi (CRM), hem ön ofis (pazarlama, satış ve müşteri servisi) hem arka ofis (muhasabe, üretim ve lojistik) uygulaması olmakla kalmayıp aynı zamanda hem de diğer tüm bölümler, müşteriler ve iş ortakları ile koordinasyonu ve işbirliğini sağlayan müşteri merkezli bir ilişki yönetimi felsefesidir.⁴

¹ Y.ODABAŞ , Satışta ve Pazarlamada Müşteri İlişki Yönetimi , Sistem Yayınları 2005 s3.

² <http://www.biymed.com/pages/makaleler/makale49.htm> 2004.

³ http://www.apluspost.com/bilgi_teknoloji.php 2005.

- Müşteri ilişkileri yönetimi (CRM) , yeni müşteri edinmek, var olan müşteriyi elde tutmak, müşteri sadakatini kazanmak ve karlılığını artırmak için anlamlı iletişimler yoluyla işletme çapında müşteri davranışlarını anlama ve etkileme yaklaşımıdır.⁵
- Müşteri ilişkileri yönetimi (CRM) , genel anlamda insanların satınalma alışkanlıkları hakkında geçmişten elde edilen verilere dayanak ilerideki davranış ve ihtiyaçlarını yüksek teknolojide kullanarak tahmin etme ve bu ihtiyaçları karşılama sürecidir.
- Müşteri ilişkileri yönetimi (CRM) , şirketlerin rekabet avantajını artıran, müşteri odaklı bir iş geliştirme stratejisi olarak algılanmaktadır.⁶
- Müşteri ilişkileri yönetimi (CRM) , en değerli “iş ilişkilerini” seçmeye ve yönetmeye yönelik iş stratejisidir.
- Müşteri ilişkileri yönetimi (CRM), karşılıklı, uzun vadeli bir değer ilişkisi yaratmak için müşteriler ile gerçekleştirilen ilişkilerin etkin biçimde yönetilmesidir. Müşteri ilişkileri yönetimi (CRM) ; satış, pazarlama ve hizmet süreçlerinin müşteri odaklı bir felsefe etrafında yeniden tanımlanmasını gerektirir.⁷

Müşteri ilişkileri yönetimi (CRM)’nin tanımının ihtiyaç duyulan durumlara ve uygulamalara göre değiştiği yukarıdaki tanımlamalardan da anlaşılmaktadır.

Müşteri ilişkileri yönetimi (CRM)’in tarihi gelişimine bakıldığında , şirketlerin müşterilerine ürettikleri hizmet ya da ürünlerin nasıl bir bakışla üretildiği net bir şekilde ortaya koymaktadır.1970’lerde üretim satarız anlayışının yaygın olduğu ve müşterilerden ziyade karlı mal üretiminin ön planda olduğu bir dönem yaşanmıştır.Karlı mal üretimi , 1970’lerde esas olarak seri ve hızlı üretimle bağdaştırılmaktaydı.

⁴ http://www.kobifinans.com.tr/bilgi_merkezi/020305/14318 2007.

⁵ Ronald S. Swift ,Accelerating Customer Relationships ,Prentice Hall , 2001 s12.

⁶ www.tepum.com.tr/Etkinlikler/crm_nedir.pps 2003.

1980'lere gelindiğinde , ürettiğimizi satarız anlayışından ürünün nasıl üretileceği , müşteriye uygun olup olmadığı tartışmalarının ardından yaşanan uzun bir süreçten sonra gerek rekabet , gerekse çeşit sayısının çok olmasından dolayı müşteri odaklı bir bakış açısı şirketlere yerleşmeye başlamıştır.Müşteri ilişkileri yönetimi'nin ortaya çıkışıyla beraber , bazı yeni kavramlar ortaya çıkmıştır.Bu kavramlar temel olarak sadık müşteri, müşteri tatmini , cüzdan payı ,müşteri karlılığı gibi kavramlardır. Bu kavramlara bakıldığında pazarlama satış stratejilerinin üretip satarız , anlayışından müşteri odaklı bir hale geldiği görülebilmektedir.

Müşteri ilişkileri yönetimi (CRM) , kabul edilen temel alt kavramların bir disiplin çerçevesinde ele alınması dünyada son 10 yılda , ülkemizde ise son 5 yılda gerçekleşmiştir.⁸

80'li yıllarda insanların yaşam alışkanlıklarında başlayan , değişim rüzgarı bilgi ve teknoloji sistemlerinin de hızla gelişmesiyle, 90'lı yıllarda firmaları ürün ve hizmet konusunda , seçici, marka sadakati düşük, talepkar müşteriler ile karşı karşıya bırakmıştır.

Şirketler, yoğun rekabet koşullarında, hem müşterisini elinde tutmak, onu memnun etmek, yeni müşteriler kazanmak için çabalarken, hem de kar ve zarar hesabını doğru yapabilmek için yeniden yapılanmaktaydılar.Gelişen bilgi teknolojileri sayesinde, insan gücüne destek olacak yeni uygulamalar yaygınlaştı. Önceleri sadece veri bankası olarak kullanılan sistemlerin fonksiyonlarını, işlem güçlerini arttırmasıyla şirketler tüm süreçlerde daha kontrollü, daha verimli çalışabilme yeteneğine kavuştu.⁹

Müşteri ilişkileri yönetimi (CRM) felsefesinin dünyada tanınmasının, benimsenmesinin ve hayata geçirilmesinin 1989 yılından bu zamana kadar uzun bir

⁷O.C.Gel , CRM Yolculuğu ,Sistem Yayıncılık 2004 s28.

⁸http://blog.inspark.com/blog/2006/10/mteri_lilkileri.html 2006.

⁹ A. PAYNE , Handbook Of CRM: Achieving Excellence in Customer Management , Butterworth - Heinemann Publishers 2005 s 10.

süreç alması, bu kavramın sadece bir veri tabanı sistemi olmadığıının en güzel göstergesidir.

Tüm süreçleri ve tüm fonksiyonları kapsayan bu felsefenin yönetilmesi için; tüm müşteri ilişkilerini her yönüyle görebilecek, ihtiyaçları ve davranışları ölçebilecek, çift yönlü iletişim sağlayabilecek akıllı bütünleşik sistemlere ihtiyaç vardır. Bilgi teknolojilerinin ihtiyaçlara paralel gelişmesiyle birlikte, analiz yapabilen, ihtiyaca uygun özelleştirilebilen, esnek, türlü iletişim araçlarıyla entegre olabilen uygulamalar da hayata geçirilmeye başlamıştır. Böylece bir felsefe ve bu felsefeyi desteklemek üzere hizmet eden teknoloji bütünü olarak Müşteri ilişkileri yönetimi (CRM) ‘ni oluşturmuştur.¹⁰

Müşteri ilişkileri yönetimi (CRM), henüz gelişimini tam olarak tamamlamamış olmakla birlikte, uygulama alanlarının yaygınlaşması ve teknoloji ile bütünleşmiş bir felsefe olduğunun bilincine varılmasıyla Müşteri ilişkileri yönetimi (CRM), yeni yüzyılda firmaların hayatlarını sürdürebilmeleri için benimsenmesi ve uygulanması gereken bir anlayış olarak ortaya çıkmıştır.¹¹

1.2. MÜŞTERİ İLİŞKİLERİ YÖNETİMİ (CRM)’İN TEMEL KAVRAMLARI

Müşteri ilişkileri yönetiminin temel kavramları şöyle sıralanabilir ;

- Müşteri İlişkileri Yönetimi (CRM) ‘in Süreçleri ,
- Müşteri İlişkileri Yönetimi (CRM) ‘in Bileşenleri ,
- Müşteri İlişkileri Yönetimi (CRM) ‘in Yapı taşları ,
- Müşteri İlişkileri Yönetimi (CRM) ‘in Mimarisi ,
- Müşteri İlişkileri Yönetimi (CRM) ‘in Teknolojileridir.

1.2.1. Müşteri İlişkileri Yönetimi (CRM) ‘nin Süreci

Müşteri ilişkileri yönetiminde süreç , ilk olarak müşteriler hakkında olabildiğince bilgi toplamaya dayanır. Bu anlamda daha önce belirtildiği gibi bu sistem (CRM),

¹⁰ O.C.Gel , CRM Yolculuğu ,Sistem Yayıncılık 2004 s9.

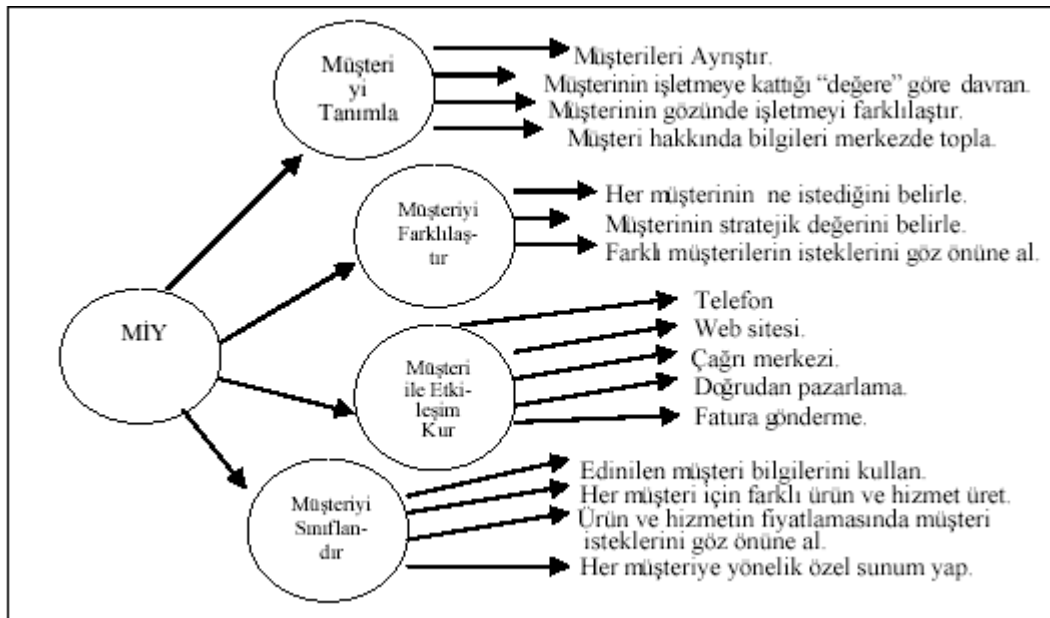
¹¹ http://blog.inspark.com/blog/2006/10/mteri_liklileri.html 2006.

yaratılan felsefenin arkasında yüksek bir teknolojiye ihtiyaç duymaktadır. Elde edilen verilerin ışığı altında müşterilerin gruplandırılması , karakteristik özelliklerinin çıkarılması ve son olarak doğru ürünün doğru müşteriye yönlendirilmesine olanak sağlar.¹²Bu süreç aşağıdaki şekilde açıklanmıştır.

Müşteri Seçimi

CRM sürecinde bu aşamasında özellikle hedef kitlelerin belirlenmesi , müşterilerin sınıflandırılması , kampanya planlarının yapıldığı aşamadır. Özellikle CRM kampanyasının , temel planının oluşturulduğu ilk aşama olması açısından çok önemlidir. Müşterileri sınıflandırma süreçlerinde amaç en karlı müşteri grubunun tespiti ve bu gruba yönelik kampanyaların planlandığı aşamadır. Bu aşamada yapılan işlemler aşağıda özetlenmiştir.

- Sınıflandırma ,
- Kampanya modelleme ,
- Marka yönetimidir.



Şekil 1.1. Müşteri ilişkileri yönetimi süreci

Kaynak : <http://www.manas.kg/pdf/sbdpdf9/Hamsioglu.pdf> 2004.

¹² A.KIRIM ,Strateji Ve Birebir Pazarlama CRM , Sistem Yayıncılık ,s49 2007.

Müşteriyi elde tutma

Müşterileri elde tutma aşamasında özellikle müşterileri gruplara ayırdıktan sonra net bir şekilde bu grupların ihtiyaçlarının belirtilmesi gerekmektedir.Doğru bir ihtiyaç analizinin yapılması hem müşteri sadakatini hem de firmanın karına doğrudan bir etkide bulunacağı görülmektedir.

Müşteri elde tutmanın bir diğer boyutu ise müşteri memnuniyetini en üst noktada tutabilmektir.Özellikle ürün ve servis memnuniyeti müşterilerde güven yaratmakta ve firmada daha uzun bir süre müşteri olarak kalmaktadır.

Müşteri kazanma

Yeni müşteri kazanma aşamasında özellikle talep analizi ve şikayet yönetimi alanlarında başarıyla uygulanması gerekmektedir.Her memnun müşteri şirketin sözcüsü gibi davranarak , yeni müşteri kazanımında firmaya yardımcı olmaktadır.¹³

Müşterinin ürünle ilgili yaşadığı bir sorunu , en kısa bir şekilde sistemli olarak çözmek müşterilerin kazanılmasında en önemli etkenlerden birisidir.

Müşteriyi büyütme ve derinleştirme

Bu süreçte var olan müşterilere , ürün satışını arttırmak için çapraz satış kampanyaları kullanılır.Bu kampanyalar , müşterilerin veri tabanındaki bilgileri kullanılarak müşteri ihtiyaçlarını tespit edip , ona uygun ürünü müşteriye sunma sürecidir.

Bu çapraz satış kampanyalarına en iyi örnek “amazon.com” dur.Bu firma , kitap ya da ürünü alan müşterisine ürünün en yakın alternatiflerini sunarak müşterilerine , çapraz satış yapmaktadır.

¹³ Y.ODABAŞ , Satışta ve Pazarlamada Müşteri İlişki Yönetimi , Sistem Yayınları 2005 s113.

1.2.2.Müşteri İlişkileri Yönetimi (CRM) 'in Bileşenleri

Müşteri ilişkileri yönetimi (CRM) , üç temel bileşenden oluşmaktadır.Bunlar , insan, proses ve teknoloji yapısıdır. İnsan (çalışanlar), müşterinin beklentilerini anlayıp, standart prosedürlerin ötesinde çözümler üretmektedirler.Proses, çalışanların ilettiği müşteri taleplerini hızlı süreç revizyonları ile müşteri odaklı yapılanmaya dönüştürmektedirler. Teknoloji ise müşteri bilgilerinin tüm temas noktalarından takip edilmesine ve farklı müşteri isteklerine hızlı çözüm üretilmesine imkan tanıyacak şirket içi otomasyonu sağlamaktadırlar.Böylece bu koordinasyon Müşteri ilişkileri yönetimi (CRM) 'i oluşturmaktadır. ¹⁴

1.2.3.Müşteri İlişkileri Yönetimi (CRM) 'in Yapıtaşları

Müşteri ilişkileri yönetimi (CRM) sekiz temel yapıtaşından oluşmaktadır.Bu sekiz yapıtaşı şunlardır .

CRM Vizyonu

CRM'in kurumsal anlamda benimsenmesi için yapılanlar, kurumsal CRM tanımı, pazar pozisyonu, müşterilere önerilmesi planlanan nihai değerler , iş hedeflerinden oluşmaktadır.

Müşteri Stratejisi

Müşteri hedefleri, pazar değerinin yönetimi, CRM'in kurumsal anlamda benimsenmesi için yapılanlar, segmentasyon anlayışları, segment hedeflerinin yönetimi, kurumsal değişim planları, müşteri tabanlı pazarlama planı, CRM esas planlarından oluşmaktadır.

¹⁴ <http://www.biymed.com/pages/makaleler/makale49.htm> 2004.

Müşteri Bakışı

Müşteri ile iletişim, şikayet yönetim sistemi, eskalasyon sistemi, müşteri temas noktalarından oluşmaktadır.

Müşteri Merkezliliği

Değişim dinamikleri, müşteri merkezli süreç yapılanması, bireysel yetkinlik ve uzmanlıkların geliştirilmesi, kariyer planlarına müşteri başarılarının yansımından oluşur. Bu kavram, çoğu zaman müşteri odaklılıkla karıştırılmaktadır. Müşteri odaklılık toplam kalite yönetiminde kullanılan bir kavramdır. Müşteri odaklılık, üretilen malı kitlelere satışı ön görürken, müşteri merkezli bir yaklaşım her müşteriye ihtiyaçları doğrultusunda bir ürün ya da bir hizmet verilmektedir.¹⁵

CRM Otomasyonu

Müşteri yönetimine dönük uygulamalar, yazılım, network ve telefon entegrasyonları, sistem altyapısı ve sistem yönetimlerinden oluşmaktadır.

Müşteri Bilgi Sistemi

Tekil müşteri tanımları, veri entegrasyonu, veri toplama yöntemleri, veri sözlüğü, müşteri verilerinin yönetiminden oluşmaktadır.

CRM Taktik Modelleri

Müşteri performans yönetimi, müşteri mülkiyeti çalışmaları, sadakat programları, ilişki modelleri, mikro pazarlama çalışmaları, hedef odaklı kampanyalarından oluşmaktadır.

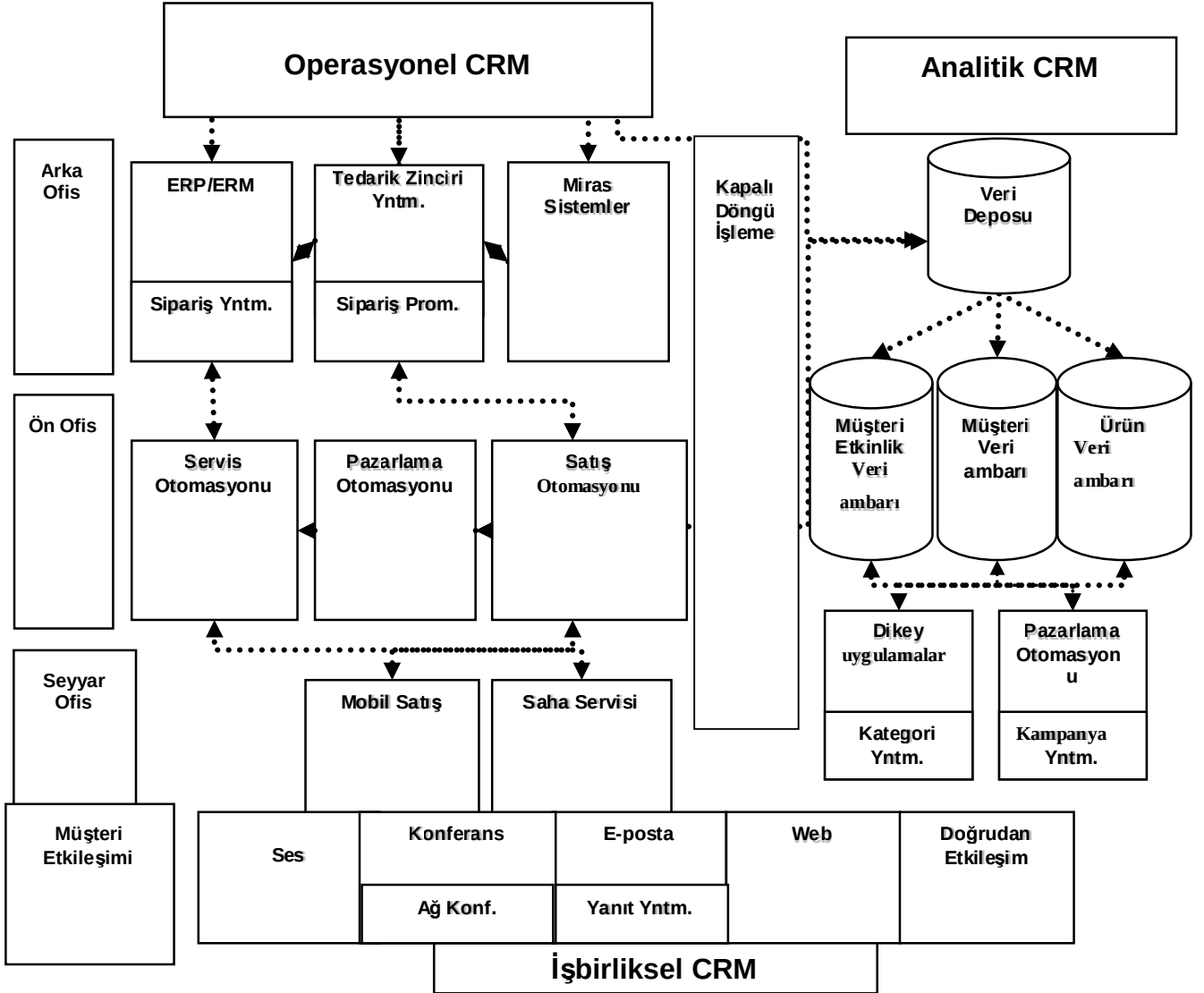
¹⁵ A.KIRIM ,Strateji Ve Birebir Pazarlama CRM , Sistem Yayıncılık , s51 , 2007.

CRM Ölçümleri

Yönetim performans göstergeleri, müşteri stratejisi göstergeleri, operasyonel göstergeler, verimlilik göstergelerinden oluşmaktadır.

1.2.4.Müşteri İlişkileri Yönetimi (CRM) 'nin Mimarisi

CRM çözümleri olarak nitelendirilen uygulamalar ile geçmişteki uygulamalar arasında birçok yapısal benzerlik ve farklılıklar vardır.Farklılıklar, Müşteri ilişkileri yönetimi (CRM) 'in geçirdiği evreleri de ortaya koyması açısından önemlidir.Bu anlamda Müşteri ilişkileri yönetimi (CRM) üç temel aşamadan oluşmaktadır.Bunlar Operasyonel CRM , Analitik CRM , İşbirliğine yönelik CRM dir.CRM 'in mimarisi aşağıda belirtilmiştir.



Şekil 1.2. Müşteri ilişkileri yönetimi mimarisi

Kaynak : http://www.erpcrm.com/crm_anasf/crm_mimarisi.htm 2003.

Operasyonel CRM

Müşteri ilişkileri yönetimi (CRM)'in bu biçimi aslında tipik iş fonksiyonlarının kapsandığı CRM çözümlerinden oluşur. Bu fonksiyonlara örnek olarak müşteri hizmetleri, sipariş yönetimi, faturalama, satış ve pazarlama otomasyonu gibi süreçleri verebilir. Bu çözümler daha çok kurumsal sistem içerisindeki finans, insan kaynakları gibi farklı iş fonksiyonlarının entegre bir yapıya kavuşturulması için kullanılmaktadır.

Analitik CRM

Analitik CRM , kullanıcılara ait verilerin elde edilmesi, depolanması, işlenmesi, analiz ve tahminlere dönüştürülerek raporlanması işlemlerini gerçekleştirmektedir. Böylelikle CRM'in operasyonel ve entegrasyon özellikleri üzerine analiz ve raporlama özellikleri eklenmektedir.

İşbirliğine yönelik CRM

İşbirliğine yönelik CRM , aslında diğerlerinin en uygun birleşiminden oluşmaktadır. Müşteriler ile şirketler arasında tam anlamıyla bir etkileşim ve koordinasyon ağının oluşmasına imkân veren , farklı iletişim kanallarından (web, telefon, e-posta vb) gelen bilgilerin , değere dönüştürülmesini sağlayan bir süreçtir. İşbirliğine yönelik CRM çözümleri müşteri ile etkileşime imkân veren tüm fonksiyonları içermektedir.¹⁶

1.2.5.Müşteri İlişkileri Yönetimi (CRM) 'nin Teknolojileri

Müşteri ilişkileri yönetiminde teknoloji olmazsa olmazdır.Müşteri verilerini analiz edecek gerek ve yeter bir teknoloji , tüm firmalar için şarttır. Müşteri ilişkileri yönetimi (CRM) için kullanılan teknoloji , Müşteri ilişkileri yönetimi (CRM) 'in aşamalarına göre sınıflandırılmıştır.Bunlar operasyonel (sales management..), analitik (data warehousing..) ve işbirlikçi (call center...) gibi ürünler kullanılmaktadır. Kullanılan ürünler detaylı olarak aşağıda belirtilmiştir.¹⁷

¹⁶ http://www.microsoft.com/turkiye/dynamics/crm/crm_nedir.aspx 2007.

¹⁷ A. PAYNE , Handbook Of CRM: Achieving Excellence in Customer Management , Butterworth-Heinemann Publishers 2005 s226- 236.

- İş zekası (Business Intelligence) ,
- Bilgi yönetimi (Knowledge Management) ,
- İletişim yönetimi (Contact Management) ,
- Müşteri etkileşim merkezi (Customer Interaction Center) ,
- Çağrı merkezi (Call Center) ,
- Bilgi sistemleri (Information Systems) ,
- Veritabanı raporlama sistemi (Database Management Report Writing) ,
- Dijital belge yönetimi (Digital Document Management) ,
- Elertronik ticaret (Electronic Commerce) ,
- E-Pazarlama (Emarketing) ,
- E-servis (eService) ,
- İnternet tabanlı çözümler (Web-Based Solutions) ,
- İnternet analizi (Web Analysis) ,
- Pazarlama otomasyonu (Marketing Automation) ,
- Tahmin etme (Forecasting) ,
- Satış otomasyonu (Sales Automation) ,
- Kablosuz veya hareketli veri işleme (Mobile Computing/ Wireless) ,
- Bağlantı yönetimi servisleri (Contact Management Services) ,
- Sipariş yönetimi (Order Management/ Distribution) ,
- Saha gücü otomasyonu (Field Force Automation) ,
- Saha servisi (Field Service) ,
- Ortaklık ilişkisi yönetimi (Partner Relationship Management) ,
- Kişiselleştirme (Personalisation) ,
- Telefonda pazarlama ve satış (Telemarketing/ Telesales) .

Bu teknolojilere rağmen bir kurum , müşteri ilişkileri yönetimi (CRM) ‘i uygulayabilmek için şu temel teknolojilere sahip olmak zorundadır.¹⁸

- Veri tabanı teknolojisi ,
- Call center teknolojisi ,

¹⁸ A.KIRIM ,Strateji Ve Birebir Pazarlama CRM , Sistem Yayıncılık ,s95- 96 2007.

- Ismarlama seri üretim teknolojisidir.

1.3.MÜŞTERİ İLİŞKİLERİ YÖNETİMİ İLE İLGİLİ DİĞER KAVRAMLAR

1.3.1.Müşteri Tanımı

Müşteri , belirli bir mağaza ya da kuruluştan düzenli alışveriş yapan kişi ya da kurumdur.Müşteri ilişkileri yönetimine bakıldığında , müşteri üç gruba ayrılmaktadır.Bunlar sırasıyla , Mevcut müşteri, Yeni müşteri ve de Şirketi terk etmiş müşteridir.Bu müşteri çeşitleri Müşteri ilişkileri yönetiminin bir anlamda amaçlarındadır ortaya koymaktadır. Müşteri ilişkileri yönetimi (CRM) , yeni müşteri kazanmaktansa , var olan müşterilere yapılan satışı arttırarak , satış karlılığını arttırmayı amaçlamaktadır.¹⁹

1.3.2.Müşteri Sadakati Kavramı

Müşteri sadakati kavramını , müşterilerin bir şirkete olan bağlılı olarak tanımlayabiliriz.Müşterilerin neden şirket değiştirdiğiyle ilgili yapılan araştırmalarda , müşterilerin % 68 ‘inin kayıtsız bir tavır hissettikleri için şirket değiştirdiğini ortaya koymaktadır.Bu araştırmanın sonuçlarında sadece müşterilerin % 14 ‘nün üründen memnun kalmadığı için şirket değiştirdiğini , % 9 ‘unun rakip şirketleri tercih ettiği , % 9‘unun ise diğer sebeplerden şirket değiştirdiğini ortaya koymuştur.²⁰

Bu kavram doğrudan sadık müşteri ile ilişkilidir.Sadık müşteri , şöyle tanımlanabilir. Düzenli olarak bir işletmeden alışveriş yapan , aynı şirketten birden fazla ürün ya da hizmet alan , alışveriş yaptığı yeri başkalarına tavsiye eden ve rakip firmalara kaçmayan müşteri olarak tanımlanabilir.²¹

¹⁹ Y.ZENGİN, Değer Yaratan Müşteri İlişkileri Yönetimi ,Yüksek Lisans Tezi, Marmara Üniv. , s 24 2006

²⁰ 65.110.73.19/UploadsNew/Gallery/Seminars/ICT3/III_CRM_Awareness-Handouts.pdf 2006.

²¹ Y.ODABAŞ , Satışta ve Pazarlamada Müşteri İlişki Yönetimi , Sistem Yayınları 20005 s11-12.

Günümüzde yoğun rekabet ortamından dolayı , işletmeler müşteri kavramlarını yeniden tanımlayarak , bir üründen alınan kardan ziyade , müşterileri sadık hala getirerek müşterilerden bir ömür boyu kar elde etmeyi amaçlamaktadır.Bu anlamda müşteri tanımları şöyle sıralanabilmektedir.²²

- İlk kez alan müşteri
- Tekrar alan müşteri
- Sürekli müşteri
- Sözcü

İlk kez alan müşteri

Müşteri sadakatinde yeni müşteriye verilen ilk izlenim müşterinin aldığı ürün ya da hizmetten ne kadar memnun olup olmadığıyla alakalıdır.Genellikle alınan üründen memnuniyetin yüksek olması , müşteri sadakatini arttırmaktadır.

Tekrar alan müşteri

Müşteri sadakatinin asıl amaçlarından biri olan müşterinin tekrar satın alması , işletmelere yüksek kar sağlamaktadırlar.Müşterilerin ürün anlamında , memnuniyetinin yüksek olması ,tekrar satın alan müşteriye yaratmaktadır.

Sürekli müşteri

Bu müşteri türü , genellikle şirketin bir başka şirkete sürekli olarak bir hizmeti vermesinin sonucudur.

Sözcü

Müşterinin sözcü olarak tanımlanması tamamiyle şirketle ürün ya da hizmet aldığı anda , meydana gelen yüksek memnuniyetin ürünüdür. Müşteriler , sözcü

²² Y.ZENGİN, Değer Yaratıcı Müşteri İlişkileri Yönetimi ,Yüksek Lisans Tezi, Marmara Üniv.s32 , 2006

olduklarında çevrelerine , arkadaşlarına ve ailesine şirket hakkında olumlu düşüncelerini yayarak şirketin reklamını yapmakta ve şirketin güvenilirliğini arttırmaktadır.

Bu müşteri zinciri ve yaratılmak istenen müşteri sadakati uzun ve yönetilmesi gereken zorlu bir iştir.Müşterilerin aldığı bir ürün ya da hizmetin kusurlu çıkması veya aldıkları üründen memnun kalmamaları , müşteri sadakatini doğrudan etkilemektedir.Bu gibi durumlarda , şirketin çabuk ve kalıcı çözümleri müşterilerine ulaştırması , müşteride olumlu bir ifade bırakacağı gibi aynı zaman da müşteri sadakatinede olumlu bir katkısı olmaktadır.²³

Müşteri sadakatini etkileyen faktörlere bakıldığında güven , vazgeçilmezlik ve de önemsenmek gibi kavramlar ön plana çıkmaktadır.Bunun gibi kavramların müşteri sadakatine olumlu ya da olumsuz bir etkide bulunacağı kesindir.²⁴

1.3.3.Müşteri Tatmini Kavramı

İşletmeler ayakta kalabilmek için müşterilerin beklenti ve isteklerini dikkate almak zorundadırlar.Müşteri tatmini , işletmelerin performans ve müşteri değerleri yaratmak açısından en önemli kriterlerden birisi olarak kabul edilmektedir.Müşteri tatmini , genel olarak , müşterilerin gerek üründen gerekse kurumdan beklentilerin karşılanmasıdır.

Müşteri tatminin yüksek tutulması , müşteri sadakatini doğrudan etkileyen en önemli faktörlerden birisidir.Müşteri tatminin sonucu olarak , müşteri sadakatinin artması , müşteriden ömür boyu faydalanılmasına olanak sağlamaktadır.Müşteri , satın aldıktan sonra aşağıdaki beş memnuniyet düzeyinden biri yaşanacaktır.²⁵

- Çok hoşnut

²³ A. PAYNE , Handbook Of CRM: Achieving Excellence in Customer Management , Butterworth-Heinemann Publishers 2005 s102-110.

²⁴ O.C.Gel , CRM Yolculuğu ,Sistem Yayıncılık 2004 s50.

²⁵ Y.ZENGİN, Değer Yaratın Müşteri İlişkileri Yönetimi ,Yüksek Lisans Tezi, Marmara Üniv. s27 28,2006

- Hoşnut
- Kayıtsız
- Hoşnutsuz
- Çok hoşnutsuz

Bu düzeylerden Hoşnut - Çok hoşnut olarak memnun ayrılan müşterilerin doğal olarak müşteri tatmini ve sadakati yüksek olacaktır.Bu kavramların karlılık oranına etkisinin olumlu olacağı görülebilmektedir.

Müşterilerin tatminsizliğinin sonucunda oluşabilen davranışlar şunlardır .

- Ürünü boykot etme ve ikame malara yönelme ,
- Markayı boykot etme ve yeni markaya yönelme ,
- Satıcıyı boykot etme , başka satıcılara yönelme ,
- Aile ve yakın çevresine şikayetini iletme ,
- Ürünü iade etme ,
- Basın yayın yoluyla ürün hakkındaki fikirlerini beyan etme, olarak sıralanabilmektedir.

1.4.MÜŞTERİ İLİŞKİLERİ YÖNETİMİ (CRM)'İN AMAÇLARI VE SAĞLADIĞI AVANTAJLAR

- Müşterilerin tam istediği ürün ve hizmetleri sağlamak ,
- Müşteriye daha iyi hizmet sunmak ,
- Daha efektif çapraz satış ,
- Satış ekibinin daha hızlı satış kapatması ,
- Eski ve değerli müşterileri tutmak ve yenilerini kazanmaktır. ²⁶

Müşteri ilişkileri yönetimi (CRM)' in şirketlere olan faydaları şöyle sıralanabilir . ²⁷

²⁶ 65.110.73.19/UploadsNew/Gallery/Presentations/CRM/CRM-distribution.pps 2004.

²⁷ <http://www.sauemk.com/makale.html> 1999.

- Müşterileri sınıflandırmamızı sağlar ,
- En uygun zamanda en uygun pazarlama programı ile en uygun müşteriye yaklaşma olasılığı hesaplar ,
- Müşterinin firmaya daha çabuk ulaşmasını sağlar ,
- Müşterinin daha çabuk karar vermesine olanak tanır ,
- Müşteri sadakatini artırır ,
- Başka firmalarla işbirliği yaparak yeni gelir olanakları yaratır ,
- Müşteri tatmin değerinin yükselmesini sağlar ,
- Birim müşteri gelirinin artmasını sağlar ,
- Müşteri sayısını artırır ,
- Satış giderlerinin azalmasını sağlar ,
- Süreç verimliliklerini artırır ,
- Stok yatırımlarının optimize edilmesini sağlar ,
- Rekabetten önce değişimleri yakalayarak pazar payının artırılmasını sağlar .

Müşteri ilişkileri yönetimi (CRM)' in şirketlere sağladığı faydaları rakamsal olarak ifade etmek gerekirse, CSO Insights şirketinin dünya çapında bin iki yüz elli'nin üzerinde şirketi kapsayan araştırmasında , yüz yirmi'nin üzerinde parametre kullanılarak satış yapılarının ne derece verimli çalıştığı değerlendirilmiştir.

Şirketlerin CRM uygulamalarıyla elde ettikleri rakamsal sonuçlar şunlardır.²⁸

- Müşterilerle %61.9 oranında daha iyi iletişim sağlandığı ,
- Müşterilerin ihtiyaçlarını %50.3 oranında daha iyi tahmin edildiği,
- Yönetim yükünün %41.7 oranında azaldığı ,
- Gelirlerinin ise %30.1 arttığı görülmüştür.

Bu veriler değerlendirildiğinde iş yükünün azalması yani zaman kavramının etkin kullanımı , müşteri memnuniyeti ve sadakati üzerinde etkinliğinin arttığını görebilmekteyiz.

²⁸ Ronald S. Swift ,Accelerating Customer Relationships ,Prentice Hall , 2001 s28.

Müşteri ilişkileri yönetimi uygulamalarının rekabetçi bir ortamda şirketlere sağladığı avantajlar düşünüldüğünde müşteri ilişkileri yönetimi (CRM)'in neden bu kadar önemli ve vazgerçilemez olduğu daha iyi anlaşılmaktadır.

1.5.MÜŞTERİ İLİŞKİLERİ YÖNETİMİ (CRM) PROJELERİNİN BAŞARILI VE BAŞARISIZ OLMA NEDENLERİ

Müşteri ilişkileri yönetimi (CRM) projelerinin başarısız olmasının bir çok nedeni bulunmaktadır.Bunların en önemlilerini şöyle sıralayabiliriz .²⁹

- Projenin tek başına IT personeline bırakılması,
- Vizyon oluşturmada, hedefler belirlenmeden projeye başlanması,
- Pahalı teknolojilerin tek başına yeterli olduğunun düşünülmesi ,
- CRM'in sadece bir programdan ibaret olduğunun düşünülmesi,
- İnsandan çok teknolojiye yatırım yapılması, insana gereken önemin verilmemesi,
- CRM projesini çok kısa zamanda tamamlanmaya çalışılması,
- Tecrübesi olmayan kişilerin, projeyi profesyonel destek almadan yürütmeye çalışması,
- Üst yönetimin destek olmaması,
- Ayrıştırma tekniklerinin uygulanmaması,
- Projenin başına getirilen kişinin bilgisiz ve tecrübesiz olması,
- Yeterince araştırma yapmadan projeye başlanması,
- Müşteri odaklı düşünememe,
- Eksik müşteri bilgileri ile CRM çalışmalarını yürütmeye çalışma,
- Veri ambarının güncel olmaması ve eksik verilere dayanarak CRM çalışmalarını yürütmeye çalışma,
- Projeyi ucuza mal etmeye çalışmadır.³⁰

²⁹ P. Bligh, D. Turk , CRM Unplugged Releasing CRM's Strategic Value , Wiley Publishers 2004 s 33.

³⁰ <http://www.herkesmusteri.com/scrm/crmbasarisiz.aspx> 2007.

Müşteri ilişkileri yönetimi (CRM) projelerinin başarılı olma nedenleri ise şunlardır.³¹

- Satış sürecinin iyi tanımlanması ,
- Üst yönetimin, satış yönetiminin ve satış temsilcilerinin CRM'e bağlılığı ve kararlılığı olması ,
- Etkinliklerin otomasyonu ile daha fazla satış yapılması ve engellerin kaldırılması ,
- Doğru tedarikçilerin veya diğer hizmet sağlayıcıların doğru seçilmesi ,
- Yönetimin değil, satış elemanlarının ve müşterilerin önemi vurgulanmalı,
- Tüm zaman dilimlerinin, kullanıcıların ve iş tarzlarının ihtiyaçlarının karşılanması için artırılmış destek sağlanmalı ,
- Saha satışları için uzaktan iletişim kurulmalı ,
- Satış senaryoları üzerine kurulu bir eğitim programı planlanmalı ,
- Sürdürülebilir ve geliştirilebilir teknolojiye yatırım yapılmalıdır.³²

1.6. MÜŞTERİ İLİŞKİLERİ YÖNETİMİNDE (CRM) KAMPANYA YÖNETİMİ

Bir CRM kampanyasında esas amaç yeni müşteriye elde etmek ve mevcut müşteriye korumaktır. Bu kampanya sürecinde pazarlama kanallarının doğru bir şekilde kullanılmasının yanı sıra müşteriler hakkında detaylı bilgi edinmeyi amaçlamaktadır.

Uygulanan bir kampanyanın başarılı olması için şu dört özelliği taşınmalıdır.³³

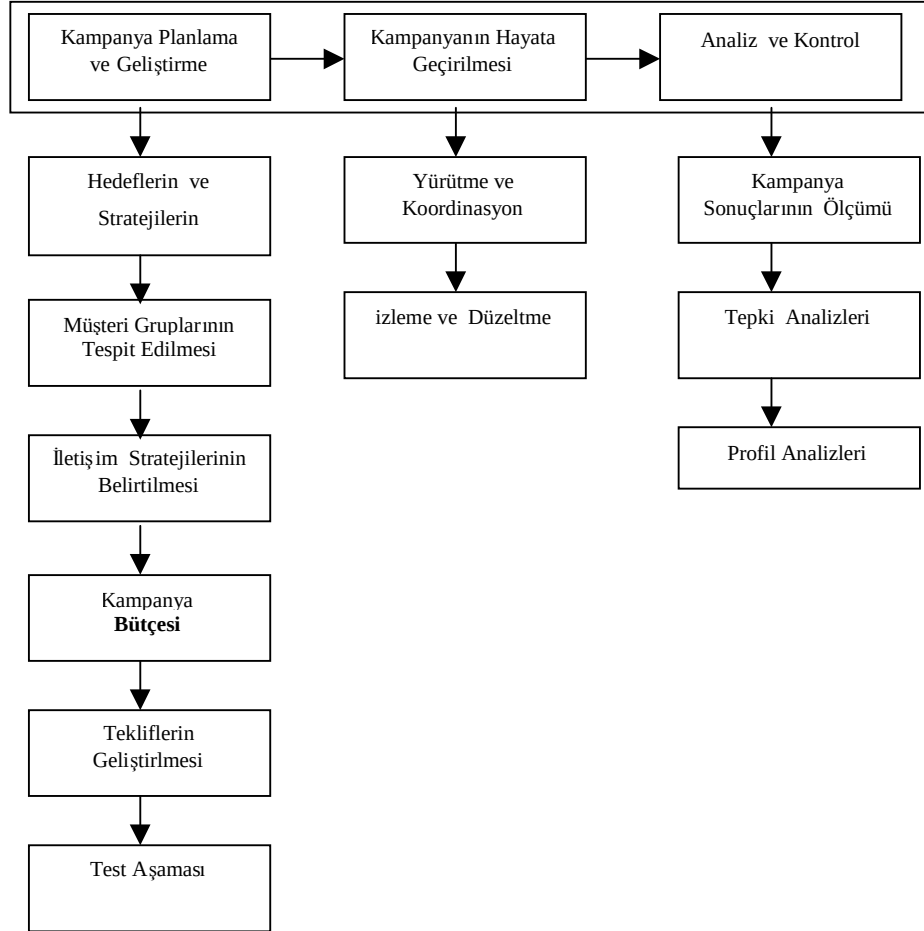
- Elde edilen veriler saklanmalı ,
- Verilere uygun analizler uygulanarak parametreler arasında bir bağ kurulmalı,
- Değişkenler tahmin edilebilir olması ,
- Kampanya sonucunda elde edilen bilgilerin yapılacak olan bir diğer kampanyada kullanılabilir olması gerekmektedir.

³¹ P. Bligh, D. Turk , CRM Unplugged Releasing CRM's Strategic Value , Wiley Publishers 2004 s165

³² www.srdc.metu.edu.tr/webpage/documents/Kosgeb/CRM_KOSGEB.ppt 2005.

Ayrıca yapılan analiz ve değerlendirmelere dikkat edilerek , hazırlanan bir kampanyanın başarı ihtimali daha fazladır.Bu kampanyanın müşterilere ulaşma yöntemleri olarak e-posta , telefon , doğrudan satış , broşür gibi seçenekler uygulanmaktadır.Kampanya yönetim süreci aşağıdaki şekilde belirtilmektedir.

MÜŞTERİ İLİŞKİLERİ YÖNETİMİNDE (CRM) KAMPANYA YÖNETİMİ



Şekil 1.3. CRM kampanya yönetim süreci

Kaynak : <http://www.teamworkcrm.com/Web/Istanbul/TeamWork.nsf/KeyMetin/CRM!OpenDocument&Click> 2004.

Yukardaki şekildedeki görüldüğü gibi başarılı bir kampanya yönetim sürecinin üç temel aşaması vardır.Bunlar planlama , uygulama ve analizdir.

³³ http://www.bendevar.com/v3/makale_326.html 2004.

1.6.1.Kampanyanın Planlanması

Kampanya planlamasına bakıldığında , kampanyanın hedeflerine uygun yönetilmesi için stratjik kararlar alırlar.Hedef tanımı ise , genellikle ulaşılmak istenen müşteri sayısı olarak tanımlanmaktadır.

Yapılan analizler sonucunda hangi müşteriye hangi ürünün uygun olduğunun tespit edilerek , belirlenen ürünler için kampanya yürütülür.Bu kampanyalar , bir grup müşteri hedeflerinden çok , birkaç grup müşteriye kampanyaya dahil edebilmektedir.³⁴

Kampanyaların çok sayıda grubu içine alacak şekilde organize olunmasında teknolojinin sağladığı fayda yadsınamaz .Çok sayıda müşterilerin aynı anda analizi ya da veri kayıtlarının yapılmasına olanak sağlayan teknoloji , kampanyanın daha başarılı olmasına yardımcı olmaktadır.

Kampanyalara hedef belirlerken şu dört grup için hedefler belirlenmelidir.³⁵

- Pazara girmek ,
- Pazarda büyümek ,
- Ürün geliştirmek ,
- Çeşitlendirmek içindir.

Kampanyanın hedefleri belirlendikten sonraki aşama , müşteri gruplarının tespitidir.Müşteri grupları oluşturulması için , çok sayıda grupta analizi verilere uygulanmak zorundadır.Müşteri gruplarını tespit etmek için , iki veri kullanılmaktadır.Bunlar , satınalma davranışları ve müşteri bilgisinin tanımlanmasıdır.

³⁴ N. WOODCOCK , The Customer Management Scorecard: Managing CRM for Profit, Kogan Page Publisherss 2003.

³⁵ http://www.kobifinans.com.tr/yazici.php?Article=8652&Where=bilgi_merkezi&Category=020305 2005.

Satınalma davranışları , müşteri ilişkileri yönetiminden elde edilen müşteri ihtiyaçlarını ön plana alarak müşteri gruplarını oluşturmaktadır.Müşteri bilgisinin tanımlanmasını gruplama olarak kullanmak içinse , sadece müşterilerin geçmiş kampanyadaki verileri gözönüne alarak grupları oluşturmaktadır.

Kampanyanın başarıya ulaşabilmesi için , en kritik aşama müşteri gruplarının oluşturulma sürecidir.Bu süreçte oluşan bir hatanın , kampanyanın başarısında direk etkisi olacağı görülmektedir.Verilerine göre , farklı müşteriler kampanyaya dahil edilebilir.Müşteri çeşitleri , mevcut müşteri , potansiyel müşteri , kaybedilen müşterilerdir.

Kampanyaya yönetiminde gerek gruplama , gerekse diğer gerekli analizlerin kullanılarak grupların oluşturma sürecinin iyi sonuçlar vermesi , tamamiyle eldeki veri setinin kalitesine bağlıdır.

İletişim stratejilerinin geliştirilemesine bakıldığında ise esas konu , kampanya için hedef kitle olarak belirlenen müşteri gruplarına nasıl ulaşılacağıdır.Müşterilere ulaşmak için kullanılan araçlar , dergi , gazete, televizyon , internet ve radyolardır. İletişim planının uygulamadaki amaç hedef kitleye çabuk ve etkili bir şekilde ulaşmayı gerçekleştirmekdir.

Kampanya teklifinin oluşturulması ise , müşterilerin ilgisini daha fazla çekmek için yapılan bir çalışmadır.Bu çalışmanın amacı var olan müşterinin devamlılığını sağlamak , yeni müşterileri ürünü almalarına ikna etmek esasına dayanır.Tekliflere örnek olarak kullanılan promosyonlar , kuponlar vb. verilebilir.Teklif planını oluştururken şunlara dikkat edilmelidir.

- Ürün konumlandırma ,
- Fiyat ,
- Taahhüdün uzunluğu ,
- Ödeme koşulları ,
- Risk azaltma mekanizmalarıdır.

Kampanya bütçesi , yapılan harcamaların ve aktivitelerin maliyetlerine bakılarak belirlenir.Kampanyaya ayrılan bütçe kampanya başarısını doğrudan etkilenmektedir.Bütçe kesin bir şekilde belirlenmekten ziyade , yürütülen kampanya sürecinde belirlenmelidir.Bütçe değerlendirmesinden sonra , dikkate alınması gereken aşama test aşamasıdır.

1.6.2.Kampanyanın Hayata Geçirilmesi

Kampanyayı hayata geçirme süreci tamamiyle operasyonel bir süreçtir.Bu süreç değerlendirme ve izleme olarak ikiye ayrılır.

Yürütme sürecinde kullanılmak üzere , üç tane program hazırlanır.Bunlar , kampanya programı , kampanya cetveli ve aktivite cetvelidir.

Kampanya programı , kampanyanın hayata geçirilmesi sürecinde yapılacak olan herşeyi özet bir şekilde içerir.Bu program , kampanyanın yönetilmesinde bir faaliyeti olan her kişiye verilir.Ayrıca program özellikle yapılacak işlerin bitiş tarihlerininide içermektedir.

Kampanya cetveli , kampanya bileşenlerinin ve de planlanan olayların listelendiği bir cetveldir.Kaynak , zaman , müşteri hedefleri gibi konularda koordinasyonu sağlamak amacıyla kullanılmaktadır.

Aktivite cetveli ise , tamamiyle kendi operasyonel aktivitelerini planlamak ve de koordinasyonunu sağlamak için kullanılan bir listedir.Amaç herhangi bir faaliyetin zamanında ve etkin bir şekilde yapılmasını sağlamaktır.

Kampanyanın izleme aşaması, kampanyanın uygulanma sırasında ortaya çıkan sorunlarını çözmek ve ilk ön bilgiyi elde etmek amacıyla kullanılmaktadır.Bu aşamada ön bilgi almak için özet raporlar hazırlanarak durum değerlendirilmesi yapılmaktadır.³⁶

³⁶ O.C.Gel , CRM Yolculuğu ,Sistem Yayıncılık 2004 s184.

1.6.3.Kampanya Analizi ve Kontrolü

Kampanya yönetiminin son aşaması olan kampanya analizi , elde edilen müşteriler hakkında daha derin bilgi edinmenin yanısıra ilerideki yapılacak kampanyaları geliştirmek için kullanılır.Yapılmış iyi bir kampanya diğer bir kampanya için örnek teşkil etmektedir.

Kampanya analizlerinde kullanılan teknikler şu şekilde sıralanabilir.³⁷

- Performans ölçümleri
- Tepki analizleri
- Profil analizleridir.

Performans ölçümleri , uygulanan kampanyanın sonuçlarını değerlendirmek üzere kullanılan parametrelerdir.Kampanya sonuçları ölçülebilir olmasından dolayı yorumlanması ve de kampanyada değişikliklere gidilmesine olanak sağlamaktadır.

Tepki analizleri , kampanya sürecinde ortaya çıkan sonuçları değerlendirmek için kullanılan yöntemlerdir.Tepki ,şikayet vb gibi kampanya sonrası elde edilen değerler kullanılarak kampanya ile ilgili tepkilerin tahmini değerleri hesaplanmaya çalışılır.Kullanılan istatistik tekniği olarak özellikle Lojistik regresyon kullanılmaktadır.

Profil analizi , kampanyaya katılan müşterilerle işletmenin mevcut müşterileri arasındaki karakteristik farklarını ortaya koymak için kullanılan bir analizdir. Bu analiz kampanyaya katılan müşterilerin profilini belirleyerek kampanyanın hedeflenen gruba ne kadar başarıyla ulaştığının tespitinde de kullanılmaktadır.

BÖLÜM 2 VERİ MADENCİLİĞİ (DATA MINING)

2.1.VERİ MADENCİLİĞİ TANIMI

Zaman içerisinde teknolojinin hızla gelişmesine bağlı olarak çok büyük miktarlarda verilerin elde edilmesine ve de depolanmasına olanak sağlamıştır. Bu veri yoğunluğu içerisinde istenen ya da gizli kalmış yararlı verilerin elde edilme sürecine veri madenciliği olarak tanımlanmaktadır.

Veri madenciliği, veri ambarlarında tutulan verilerden otomatikleşmiş modeller sayesinde anlamlı bilgileri, ilişkileri ve davranışları ortaya çıkarma süreci olarak da tanımlanmaktadır. Bu süreçte, veri içinde önceden pek fazla bilinmeyen veya görülemeyen desenler (pattern) öncelikle ortaya çıkarılmaktadır. Bu desenler genellikle bilgiler arasındaki ilişkilerin, sıralamanın, sınıflandırmanın, veri birlikteliğinin ve tahminlemenin sonucunda elde edilmektedir.³⁸

Veri madenciliği uygulama alanları düşünüldüğünde , en çok kullanılan sektörler şunlardır.³⁹

- Pazarlama ,
- Bankacılık ve sigortacılık ,
- Biyoloji, tıp ve genetik ,
- Kimya ,
- Yüzey analiz ve coğrafi bilgiler ,
- Görüntü tanıma ve robot görüş sistemleri ,
- Uzay bilimleri ve teknolojileri ,
- Meteoroloji ve atmosfer bilgileri ,
- Sosyal bilimler ve davranış bilimleri ,

³⁷ http://www.init.com.tr/news_articles_tr.asp?haber_id=12 2006.

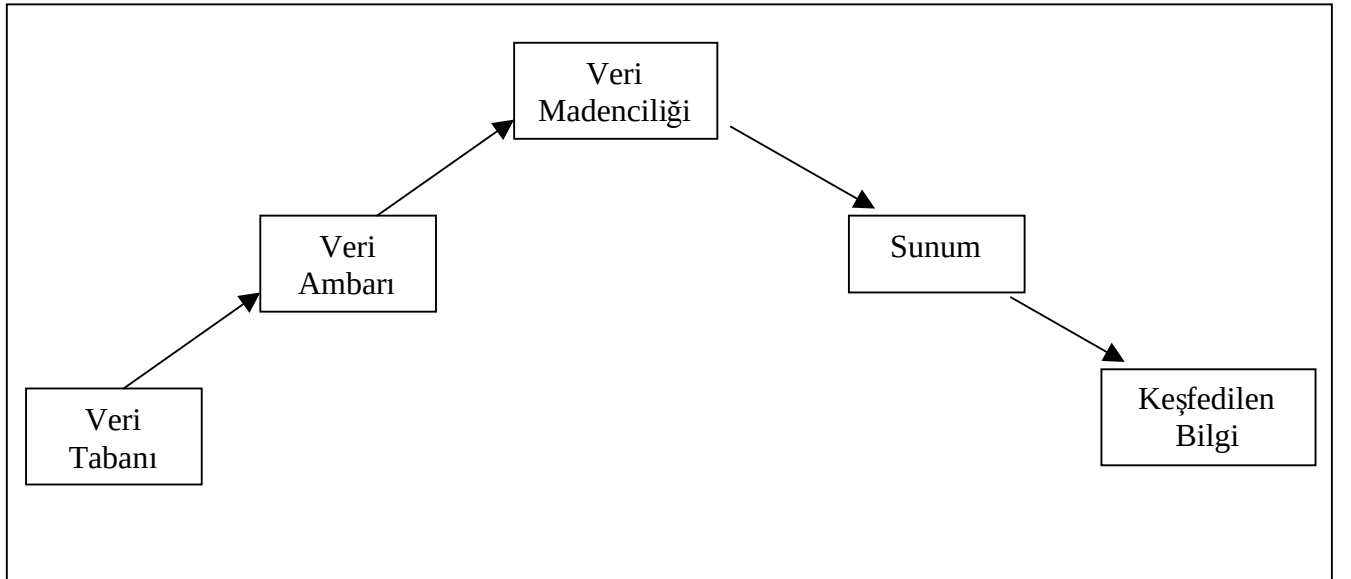
³⁸ S.MITRA, T.ACHARYA, Data Mining : Multimedia, Soft Computing , and Bioinformatics , John Wiley & Sons Publisher 2003

³⁹ T.T.BİLGİN , Veri Madenciliğinde Kavramı Ve Analiz Yöntemi Uygulamaları , Yüksek Lisans Tezi, Marmara Üniv. ,2003

- Metin madenciliği ,
- Web madenciliğidir.

Veri madenciliğinin kullanım alanlarından en önemlisi Müşeri ilişkileri yönetiminin teknik alt yapısında uygulanmasıdır. Bu süreç gerek pazarlama gerekse bankacılık ve sigortacılık sektörlerini kapsamaktadır. Müşteri ilişkileri yönetimi için müşterileri sınıflandırma , grup oluşturma , en değerli müşteri grubunu oluşturma süreçlerinde Veri madenciliğinden yararlanılmaktadır.⁴⁰

Veri madenciliğinde bilgi elde etme süreçleri aşağıdaki şekilde belirtilmiştir. Aşağıdaki şekilde belirtilen süreçlerin başlangıç noktası veri tabanındaki elde bulunan veriden , veri ambarına , oradan veri madenciliğine , son aşama olaraksa gerekli bilginin elde edilmesi aşamasına gelmektedir.



Şekil 2.1 Veri tabanlarında bilgi keşfi aşamaları

Kaynak : http://www.isletme.istanbul.edu.tr/surekli_yayinlar/dergiler/nisan2000/1.htm

Veri madenciliğinin gelişimine bakıldığında veri madenciliğinin gelişmesindeki en önemli nedenin , veri tabanlarındaki büyük teknolojik gelişme olduğu görülebilmektedir. Veri tabanlarının çok büyük miktarlarda veriyi saklayabilme özelliğinin artmasıyla , veri kümelerinden özet ve yararlı verileri ortaya çıkarma

gereksinimine bir cevap olarak Veri madenciliği ortaya çıkmıştır.⁴¹Veri madenciliği sürecinde en önemli konu , her aşamanın dikkatli ve eksiksiz bir şekilde uygulanması gereğidir.Çünkü bir aşama , diğer bir aşamanın girdisi olduğundan dolayı , bir aşamanın eksik yapılması tüm veri madenciliği sürecini etkiler .Veri madenciliği süreci şu aşamalardan oluşmaktadır.⁴²

- Problemin tanımlanması ,
- Verinin hazırlanması ,
- Model kurulumu ,
- Modelin kullanılmasıdır.

Bu aşamalar veri madenciliği sürecini oluşturmaktadır.

Problemin tanımlanması

Veri madenciliğinin en önemli aşamalarından biridir.Problemin ne olduğunu , gerek işletme gerekse müşteri tarafından problemin tanımlandığı ; elde edilmek istenen verinin ne için kullanılacağını tanımlayan bir aşamadır.

Veri madenciliği süreçlerinde amaçlar özellikle işletmeler için , müşteri değerini arttırma , müşteri sadakati yaratma , karı arttırmak , pazar payını artırmak vb. olabilmektedir.

Verinin hazırlanması

Verinin hazırlanması aşaması dört temel aşamadan oluşmaktadır.Bu aşamalar ;

- Verinin toplanması ,
- Verinin birleştirilmesi ,
- Verinin temizlenmesi ,

⁴⁰ T.T.BİLGİN , Veri Madenciliğinde Kavramı Ve Analiz Yöntemi Uygulamaları , Yüksek Lisans Tezi, Marmara Ünv. ,2003

⁴¹ M. DEMİRALAY ,Hiyerarşik Kümeleme Metotları İle Veri Madenciliği , Yüksek Lisans Tezi, Marmara Ünv. s 4,2005

⁴² M .BERRY , Data Mining Techniques , Wiley Publishers 2004 s7.

- Verinin dönüştürülmesidir.

Veri hazırlanması sürecinde , veri toplanması en önemli adımlardan birisidir.Bu aşamada verilerin belirlenen amaçlara uygun elde edilmesi , gerek veri hazırlama aşamasını gerekse tüm veri madenciliği sürecini doğrudan etkilemektedir.

Veri birleştirme sürecinde , farklı kaynaklardan toplanan verilerin aynı formatta çevrilmesi gerekmektedir.Böylece farklı kaynaklardan alınan verilerin hepsine aynı analizler uygulanabilmektedir.

Veri temizleme sürecinde , verilere uygulanan yöntem esas olarak analizi yanlış yönlere sürükleyebilecek olan eksik ya da aykırı verilerin veri topluluğundan çıkarılmasıyla veri madenciliği sürecine etkisi ortadan kaldırılmaktadır.

Veri dönüştürme sürecinde ise , verilerin farklı formlarını analize uygun olacak şekilde dönüştürülmesi sürecidir.

Model kurulumu

Veri madenciliği sürecinde modele dahil edilecek verilerin amaçlara uygun ve de doğru bir şekilde seçilmesi , modelin doğru kurulma olasılığını arttırmaktadır. Doğru kurulan bir model , bilmek istenilen özet bilgiyi de doğru yansıtacağı kesindir.

Modelin kullanılması

Veri madenciliğinin son aşaması olan bu süreç , ortaya çıkarılan modelin benzer amaçlara işletmeler için bir gösterge olacağı kesindir.

2.2.VERİ MADENCİLİĞİ SÜRECİNDE ORTAYA ÇIKABİLECEK SORUNLAR

Veri madenciliğinde ortaya çıkan sorunların temelinde iki unsur yatmaktadır.Bunlar ilk olarak işletmenin hangi amaçla veri madenciliği yaptığı ,

diğeri ise elde bulunan verilerden kaynaklanmaktadır.Karşılaşılan bazı problemler şunlardır.⁴³

- Veri tabanının boyutları ,
- Dinamik veri yapısı ,
- Eksik veri ,
- Gürültü ,
- Eksik değerlerdir.

Veri tabanının boyutları

Veri tabanının boyutlarının veriler için yeterli olmaması durumunda , yaptığımız analizlerin uygulanabilirliği yoktur.Bu problemin çözümünde ise , yapılan örnekleme tekniklerinin ve örnek miktarının azaltılmasıyla bu problem aşılabilmektedir.

Dinamik veri yapısı

Veri tabanlarının verileri güncellemesi , ya da yeni müşteri verilerini veri tabanına eklenmesi , veri madenciliği sürecinde olumsuz etkilere yol açabilmektedir.

Eksik veri

Veri madenciliğinde en çok rastlanan problemlerden biri olan eksik veri , analizi ya da yaratılamak istenen modeli yanlış yönlere çekebilmektedir.

Gürültü

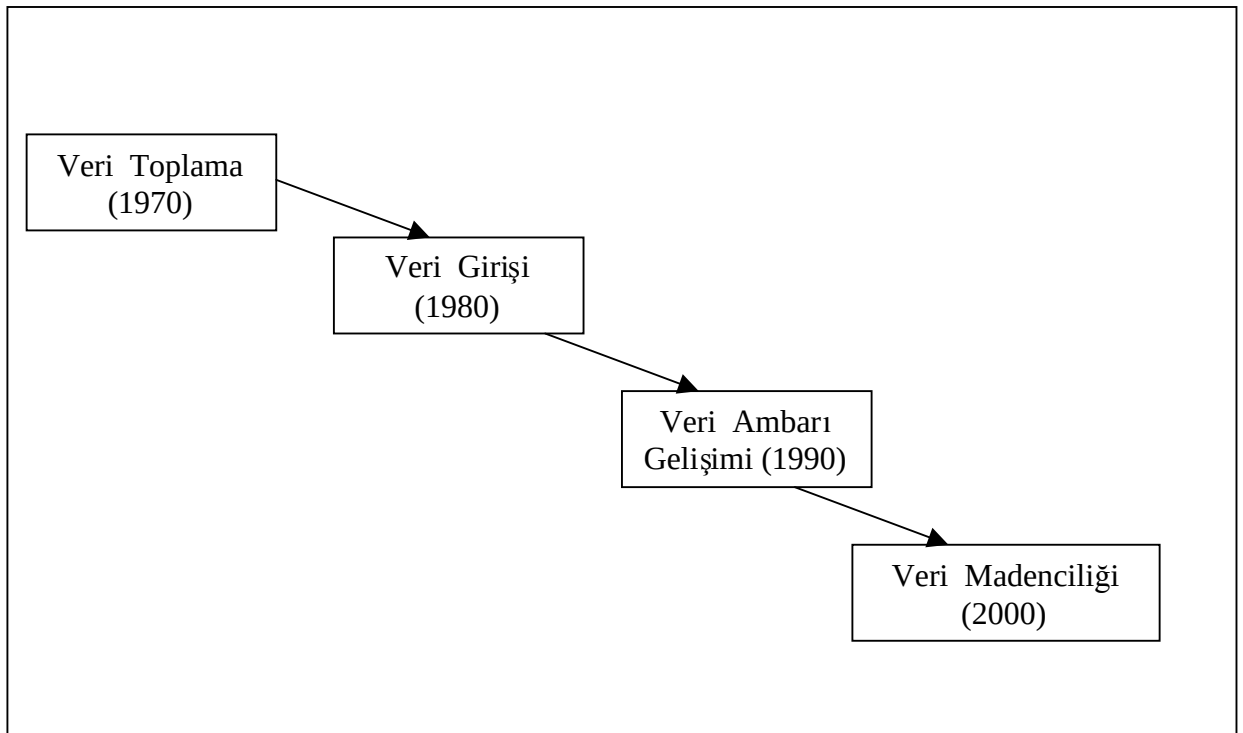
Nitelik değerlerindeki ya da sınıf bilgilerindeki hatalar , gürültü olarak tanımlanır.Ayrıca yanlış girilen veriler de bu tanıma girmektedirler.

⁴³ http://www.isletme.istanbul.edu.tr/surekli_yayinlar/dergiler/nisan2000/1.htm 2000.

Eksik değerler

Yapılan her analiz için büyük bir sorun olan veri değerlerinin hatalı olması , tüm analizi işlevsiz kılacak büyük bir sorundur.Genellikle veri toplama ya da girilme aşamasında oluşan bu hata analizi doğrudan etkilemektedir.

Geçmişten günümüze gelen süreçte , veri madenciliğinin gelişimi şu şekilde olmuştur.



Şekil 2.2 Veri madenciliğinin tarihsel gelişimi

Kaynak : S.SMITH ,Building Data Mining Applications for CRM ,McGraw Hill Publisher 1999,s16 19

2.3.VERİ AMBARI KAVRAMI

Veri ambarı , operasyonel, kalıcı, entegre ve tarihsel derinliği olan verilerin, karar destek sisteminin işlevlerini desteklemek, verilerden anlamlı ilişkiler kurarak sonuçlar çıkarmak üzere modellenmiş süreçlerin toplamıdır. Amaç , verileri organizasyondaki karar vericilerin faydalanmaları için saklanarak, veriye hızlı ve tek kaynaktan ulaşmaları imkanını sağlamaktadır.⁴⁴

⁴⁴ S.MITRA,T.ACHARYA,Data Mining : Multimedia, Soft Computing , and Bioinformatics ,John Wiley & Sons Publisher 2003 s24.

Veri ambarı ayrıca diğer analiz programlarını kullanarak , var olan veriyi sorgulama , analiz etme ve yorumlama imkanı da tanımaktadır.

Veri ambarının en önemli amacına bakıldığında , bunlar bilginin toplanması , dönüştürülmesi ve saklanması olarak ön plana çıkmaktadır.Verilerin toplanmasından sonra , veriler hatalarının giderilmesine çalışılır.Böylece daha doğru verilerle daha doğru analizler yapma imkanı sağlanmış olunur.

Firmaların yoğun rekabet ortamı içinde müşterilerini ve de pazar paylarını ellerinde tutmak ya da genişletmek için veri ambarının sağladığı analiz imkanı ve de raporlama işlemlerinden yararlanmaları gerekmektedir.Firmalar için oluşturulacak veri ambarlarının miktarı , kapasitesi firma ihtiyaçlarına göre farklılık göstermektedirler.

Veri ambarının firmaların operasyon kısmına sağladığı faydalar şöyle sıralanabilir.⁴⁵

- Kolay bir şekilde veriye ulaşma imkanı sağlar,
- Geçmişten ders alınmasını sağlar,
- Operasyon kısmında verimliliği artırır,
- Zaman ve de paradan tasarruf sağlar.

Veri toplamanın öneminin arttığı günümüzde , kurumsal şirketler büyük bir veri ambarı oluşturmak için bütçelerinden büyük meblağlar ayırmaktadır.Müşterilerini daha iyi tanımalarına olanak sağlayan veri ambarı , gerek ürün yapılmasında ya da hangi müşteriye hangi ürünün sunulması gerektiği konusundaki uygulamaların yapılmasına olanak sağlamaktadır.

Veri ambarının taşıması gereken özellikler şöyle sıralanabilir.⁴⁶

⁴⁵ <http://www.breuer.com/benefits.asp> 2007.

⁴⁶ <http://www.breuer.com/features.asp> 2007.

- Zamana bağılı olması ,
- Kalıcı olması ,
- Konuya yönelik olması ,
- Birleştirilmiş olması gerekmektedir.

Zamana bağılı olması

Veri ambarındaki veriler , tarihlere bağılı olarak kaydedilmektedir.Verilerin zamana bağılı olarak kaydedilmesi bazı istatistiksel analizlerin uygulanmasını kolaylaştırmaktadır.Verilerin veri ambarı içerisinde zamana bağılı olmayan verilerin kullanılmaya çalışılması o verilerin güvenilirliğini ve geçerliliğini kaybetmesine sebep olacaktır.

Kalıcı olması

Veri ambarındaki verilerin güncellenmesi aşaması , sadece operasyonel veri tabanları aracılığıyla mümkündür.Esasen veri ambarındaki veri , sabit ve değiştirilemez.Güncelleme , düzeltme ya da veri silme işlemleri ancak operasyonel veri tabanlarında mümkündür.Verit ambarına verileri aktarmak ya da herhangi bir işlemi yapabilmek için , operasyonel veri tabanlarına ihtiyacımız vardır.

Konuya yönelik olması

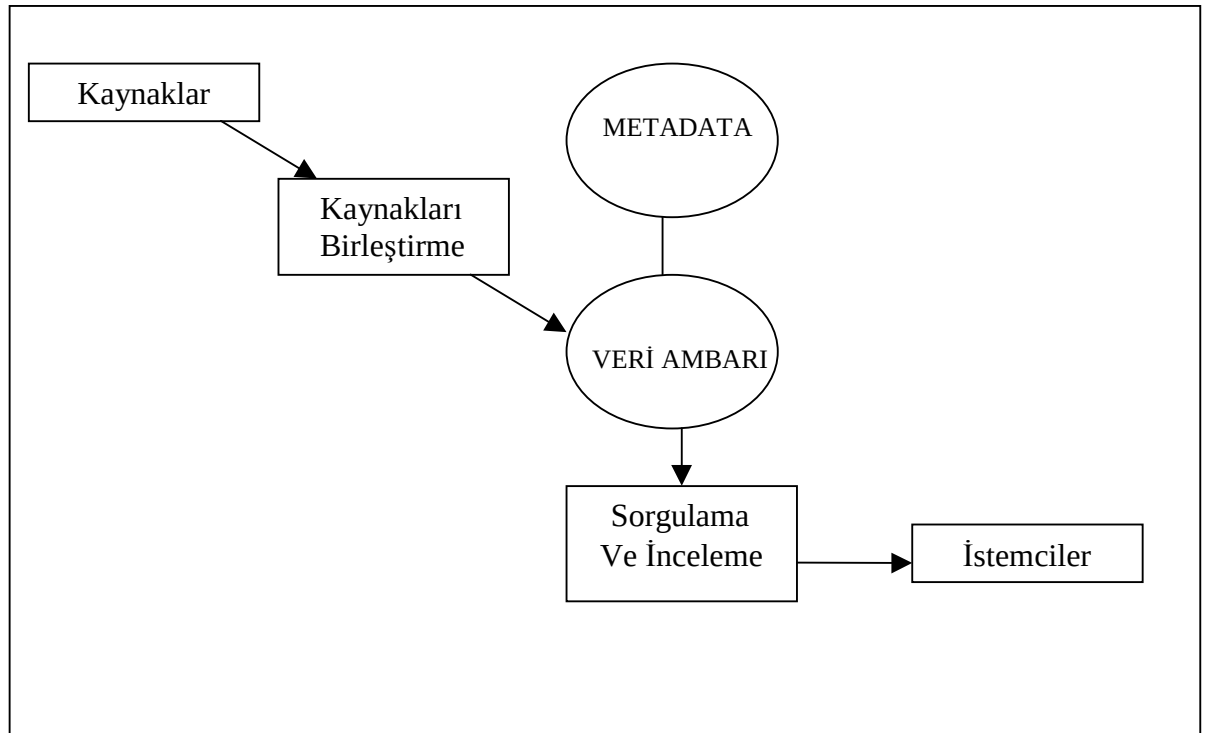
Veri ambarlarının kullanım amaçları uzun ya da orta vadeli kararlar alabilmek için , yararlanılan bir araçtır.Bu amaca uygun bir şekilde gerek analizlerin gerekse kayıtların yapılması gerekmektedir.

Birleştirilmiş olması

Veri ambarlarına kaydedilen veriler , belirli bir formata dönüştürülerek kaydedilirler.Bu dönüşüm verilerin birleştirilmesi ve farklı verilerin aynı analizin uygulanmasına olanak sağlar.

Veri ambarı , bir karar destek sistemi olup veri tabanından farklı olarak sadece verilerin günlük olarak değerlendirilmesine olanak sağlayan bir teknolojidir. Veri ambarının karar destek sistemi olarak OLAP (Online analytical Processing) , Müşteri ilişkileri yönetimi (CRM) , istatistiksel analizler ve raporlama süreçlerinde kullanılmaktadır.⁴⁷

Veri ambarı mimarisi aşağıdaki şekilde görülmektedir.



Şekil 2.3. Veri ambarı mimarisi

Kaynak : <http://web.cs.hacettepe.edu.tr/~fatih/RESEARCH/doc/tez.pdf> 2003.

Veri ambarı , mimarisine bakıldığında alt – orta – üst katman olarak üç esas katmandan oluşmaktadır. Alt katman , operasyonel verinin işlendiği ve verinin temizlenmesine olanak sağlayan ilişkisel bir veri tabanıdır. Alt katmanın en öne çıkan özellik olarak , veri tabanına bağlanıp verilere ulaşmayı sağlaması olduğu söylenebilir. Alt katman , kaynakları birleştirme aşamasından oluşmaktadır.

⁴⁷ T.T.BİLGİN , Veri Madenciliğinde Kavramı Ve Analiz Yöntemi Uygulamaları , Yüksek Lisans Tezi, Marmara Ün. ,2003

Orta katmana bakıldığında ise , veri ambarı bu katmanı oluşturmaktadır.Bu katmanın esas görevi , veri madenciliği için verileri anlaşılır bir hale getirmektir.Özellikle raporlama , analiz gibi işlemler bu katmanda yapılmaktadır.

Son katman olarak üst katman ise , karar destek sisteminden oluşmaktadır.Analizlerin sonuç kısımlarının ve özet verinin değerlendirildiği bir katmandır.Bu katman sorgulama ve inceleme aşamasından oluşmaktadır.⁴⁸

Veri ambarının bileşenleri aşağıdaki şekilde gösterilmektedir.



Şekil 2.4. veri ambarı bileşenleri

Kaynak : <http://web.cs.hacettepe.edu.tr/~fatih/RESEARCH/doc/tez.pdf> 2003.

Veri ambarının oluşturulma amacı iki tanedir.Bunlar ;⁴⁹

- Veri depolamak ve analiz yapmak ,
- Karar destek sistemini oluşturmak ve hızlı bir şekilde bilgiye ulaşmaktır.

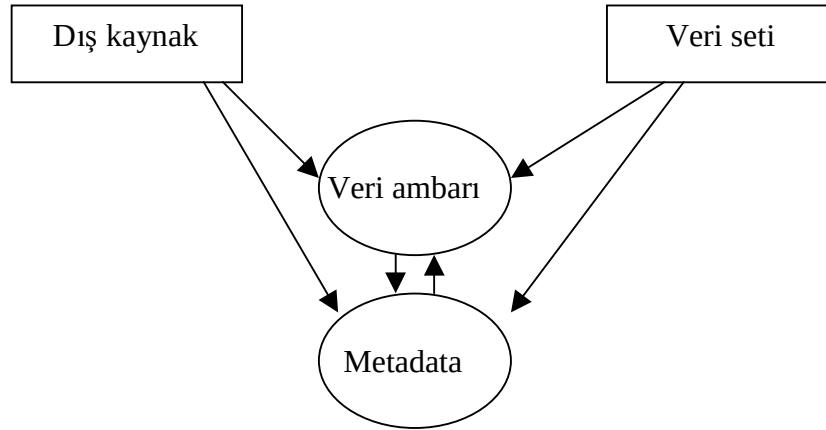
Operasyonel veri tabanları , veri tabanına günlük verileri işlenen verilerdir.Operasyonel veriler , çoğu kez kısa vadeli olarak saklanan verilerdir.

Metadata ise , hem operasyonel veri hem de veri ambarındaki verilerin yapılarını gösteren bir veridir.Gösterdiği veriden ziyade , veri setinin karakteristik bilgilerini taşıyan bir veridir.Metadata , veri biçimi , veri kullanımı , veri kaynaklarını ,

⁴⁸ C .TODMAN , Designing a Data Warehouse: Supporting Customer Relationship Management , Prentice Hall PTR Publisherss 2000 s220.

⁴⁹ http://en.wikipedia.org/wiki/Data_warehouse 2007.

verilerin nerede bulunduđu gibi bilgileri saklayan bir yapıya sahiptir. Metadata'nın veri ambarıyla olan ilişkisi řu řekilde açıklanabilir.



Şekil 2.5. Metadata 'nın veri madenciliğindeki yeri

Kaynak : S.SMITH ,Building Data Mining Applications for CRM ,McGraw Hill Publisher 1999,s30

Veri ambarı , stratejik kararların alınması anlamında kullanılmaktadır.Bu nedenden dolayı , çok miktarda verileri saklamaktadır.Verdi ambarı , alışılmış veri tabanlarının amaç ve çalışma biçimi bakımından farklılık göstermektedir.Verdi ambarları , veri tabanlarından daha fazla veri saklama kapasitesine sahiptir.OLTP (online transaction process) uygulamalarına olanak sağlamıştır.

OLTP uygulamaları veri tabanından veri silme , değıştirme ya da veri ekleme gibi işlemleri içermektedir.Bu işlemler SQL (structured query language) sorgulama dili ile gerçekleştirilmektedir.

OLTP bakıldığında , operasyonel verinin veri tabanına aktarılmasına yardımcı olan bir işlem sürecidir.Bu süreç verilerin kayıt edilmesinden güncellenmesine kadar uzanan bir süreçtir.Verdi ambarından ziyade veri tabanı üzerinde gerçekleşen işlemler , genel olarak veri yüklenmesi aşamasında kullanılmaktadır.OLTP 'nin amacı sadece veri setinin en az hata ile veri tabanına aktarılmasını amaçlamaktadır.

Veri ambarları daha çok karar destek sistemlerine yönelik veri saklama araçlarıdır. Veri madenciliği, OLAP (online analytical process) gibi uygulamalara olanak sağlamaktadırlar.

OLAP ise verilere analiz yapılmasına , raporlanmasına olanak sağlayan bir işlemler sürecidir. Bu süreç tamamiyle veri tabanından bağımsız olarak gerçekleşir. Bütün analizler veri ambarı üzerinde yapılarak on-line işlem sürecini oluşturur. Bu süreç sadece karar vericiye daha sağlıklı karar vermesine olanak sağlayan tüm analizleri veri setine uygulama imkanı vermektedir.⁵⁰

2.4. VERİ MADENCİLİĞİ TEKNİKLERİ

Veri madenciliği teknikleri , verilerin kullanılma amaçları ve de veri yapısına göre farklılaşmaktadır. Temel olarak iki ana grupta bu teknikleri toplayabiliriz. Bunlar ;

- Öngörüşel
- Tanımlayıcı

Tanımlayıcı veri madenciliği teknikleri , daha çok verinin karakteristik özelliklerini ön plana çıkarmak için kullanılırken , öngörüşel teknikler daha çok ileriki yıllarda bir parametreyi tahmin etmek için kullanılmaktadır. Kategorik olarak gruplanan veri madenciliği teknikleri ile , farklı gruplamalar da yapmak mümkündür. Bu gruplamaların en bilineni J.Han kategorileri olarak bilinir.⁵¹

- Tanımlama ve ayırlama ,
- Birliktelem Kuralları ,
- Sınıflandırma ve öngörü ,
- Kümeleme analizi ,
- Sıradışılık analizidir.

⁵⁰ <http://www.olapreport.com/fasmi.htm> 2005.

⁵¹ T.T.BİLGİN , Veri Madenciliğinde Kavramı Ve Analiz Yöntemi Uygulamaları , Yüksek Lisans Tezi, Marmara Üniv. ,2003 .

2.4.1.Tanımlama Ve Ayrılma

Veri setini tanımlama sürecinde amaç , veri hakkında özet bir bilgi elde etmektedir.Ayrılma ise , veri setindeki farklılıkları ortaya koymak için yapılan bir işlem den ibarettir.

Ayrılma işleminde kullanılan en önemli yöntemlerden birisi Diskriminant analizidir.Bu analiz , veri kümesini belirli gruplara atıyarak verileri ayırır.Bu yöntemde kullanılan model , esas olarak veri ayırma işleminde kullanılmaktadır.Ver i Ayırma modeli şu şekildedir.

$$Z_i = b_0 + b_1x_{1i} + b_2x_{2i} + \dots + b_nx_{ni}$$

Z_i : i'ninci bireyin ayırma değeri

b_i : i'ninci değişkenin katsayısı

X_{1i} : i 'ninci bireyin i'ninci değişken üzerindeki etkisidir.

Diskriminant analizi ; örneğin bir işletmede marka alan ile marka almayan müşteri arasında anlamlı bir fark olup olmadığını belirlemek için kullanılabilmektedir.⁵²

Diskriminant analizi, hatalı sınıflandırma olasılığını en aza indirgeyerek birimleri ait oldukları gruplara ayırmak amacına yönelik olan, istatistiksel bir karar verme yöntemidir . Diskriminant analizi, X veri setindeki değişkenlerin iki veya daha fazla gerçek gruplara ayrılmasını belirlemek amacıyla yararlanılan bir yöntemdir.

Diskriminant analizi, genel anlamda ayırma olup, bireylere ait p tane özellikten yararlanarak ait oldukları grupları (yığın) belirlemede veya mevcut grupları birbirinden ayıracak en iyi fonksiyonu bulmada kullanılan , çok değişkenli istatistik tekniklerinden birisidir. Bu analiz , gruplar arasında çeşitli değişkenlere bağlı olarak farklılıklarını ortaya koymasına olanak sağlamaktadır.

⁵² H.TATLIDİL , Uygulamalı Çok Değişkenli İstatistiksel Analiz , Hacettepe ün v. s 258-264 1996.

Diskriminant analizinde , birimler en az hata ile ait oldukları birimlere ayrılmaktadır. Bu analizin temelinde incelenen bireyin kitlesinin belirlenmesini sağlayacak bir fonksiyon bulunmaktadır. Diskriminant analizi iki veya daha fazla gruptaki birimlerin etkileşim seviyelerinin hangi düzeyde olduğu, diğer değişkenler arasında ne gibi farklılıklar bulunduğunu ortaya koymaktadır .

Diskriminant analizi, farklılığın en fazla hangi değişkenlerde yoğunlaştığının belirlenmesi ve böylece grupların farklılaşmasına etkin olan faktörlerin saptanmasını da sağlar. Analiz sonucunda yapılan sınıflama ile orijinal grup üyeliklerinin karşılaştırılması , belirlenen fonksiyonun yeterli olup olmadığını test etmeye olanak sağlar .

Diskriminant analizi, birbirleriyle yakından ilişkili birkaç istatistiksel yaklaşımı kapsayan geniş bir kavramdır . Bu yaklaşımlar iki ana kategoride ele alınabilir. Birinci kategoriye oluşturan yaklaşımlardan , gruplar arası farklılıkları yorumlamada faydalanılırken, ikinci kategori yaklaşımlar birimleri gruplara ayırmak amacıyla kullanılmaktadır. Diskriminant analizi eğer bir ayırma fonksiyonu belirlemeye yönelik olarak uygulanmış ise , tanımlayıcı diskriminant analizi, eğer sınıflama amacıyla uygulanmış ise , tahmin edici diskriminant analizi olarak adlandırılır. Tahmin edici diskriminant analizi, davranış değerleri içinde bulunan temel bilgilerin gruplar için , verilerin nasıl belirleneceği sorusuna işaret eder. Bir girdi eğer tahmini grubun üyesi değilse , yanlış sınıflandırılmış olarak nitelendirilir. Genellikle yanlış sınıflandırma olasılığını ve bedelini düşürmek oldukça önemlidir

Bazı yazarlar diskriminant analizinde ayırma fonksiyonu katsayılarının hesaplanmasında başvuru yöntemlere göre diskriminant analizini, kanonik diskriminant analizi, en çok olabilirlik diskriminant analizi ve bayes diskriminant analizi şeklinde adlandırır.

Genel olarak birimlerin gruplamasında bazı matematiksel eşitliklerden faydalanılır. Diskriminant fonksiyonu olarak adlandırılan bu eşitlikler birbirine en çok benzeyen grupları belirlemeye olanak sağlayacak şekilde grupların ortak özelliklerini belirlemek amacıyla kullanılmaktadır. Grupları ayırmak amacıyla kullanılan karakteristikler ise diskriminant değişkenleri olarak adlandırılmaktadır. Kısaca diskriminant analizi, iki

veya daha fazla sayıdaki grubun farklılıklarının diskriminant değişkenleri vasıtasıyla ortaya konması işlemidir .

Araştıracının, p tane özelliği bilinen gözlemleri belli özelliklerine göre bazı gruplara ayırmak istemesi, elde edilecek somut ve özetleyici bilgiler açısından istatistiksel değerlendirmede önemli bir konudur.

Araştırmacı, hatalı sınıflandırma olasılığını en aza indirgeyerek gözlemleri ait oldukları gruplara ayırmak veya bu gözlemlerin çekilmiş oldukları yığınları belirlemek isteyecektir.

Diskriminant analizinin amaçlarını dört grupta toplanabilir.⁵³

- Analiz öncesi tanımlanmış iki ya da daha fazla grubun (örneğin, mali açıdan başarılı ve başarısız işletmeler) ortalama özellikleri arasında önemli farklar olup olmadığının, bağımsız değişkenlere (açıklayıcı değişken) bağlı olarak istatistiksel olarak test edilmesi,
- Her bir değişkenin, gruplar arasındaki farka katkısının saptanması,
- Grup içi değişime oranla, gruplar arasındaki ayrımı maksimize eden tahmin değişkenleri kombinasyonunun belirlenmesi ve bu sayede başlangıçtaki açıklayıcı değişken sayısından daha az sayıda değişken ile gruplar arasındaki önemli farklılıkların açıklanması,
- Analiz öncesi tanımlanmış grupların atanması ile ilgili yöntemlerin geliştirilmesi, yeni bireylerin hangi gruba ait olduklarının saptanmasıdır.

Tüm istatistiksel ve matematiksel modellerde olduğu gibi, diskriminant analizi de bazı varsayımlara dayanmaktadır. Analizin ayırım gücü, dayandığı varsayımların sağlanmasına ya da bu varsayımlar karşısında sağlam olmasına bağlıdır. Özellikle modelin başarısının, beklenenden düşük çıktığı durumlarda, doğru yorumda

⁵³ <http://people.revoledu.com/kardi/tutorial/LDA/LDA.html> 2006.

bulunabilmek için , bu varsayımların test edilmesi gerekmektedir. Diskriminant analizinin varsayımları şunlardır.⁵⁴

- Anakütle belli özelliklere göre gruplanabilir.Birbirinden farklı iki veya daha fazla grup söz konusu olmalıdır.
- Veriler anakütleden rassal olarak seçilmiştir.
- Bağımsız değişkenler çok boyutlu normal dağılıma sahiptirler.
- Gruplara ait ortalamalar ve kovaryans matrisi önceden bilinir. Grupların kovaryans (sapma) matrisleri eşittir.Bu varsayımın sağlanamadığı durumlarda, diskriminant analizinin karesel formu kullanılabilir.
- Grupların eşit sayıda birimden oluşmadığı durumlarda, üyelerin önsel olasılıklarının bilindiği varsayılır.
- Herhangi bir birimin yanlış sınıflandırmanın maliyeti önceden bellidir.Bu varsayımlardan bir ya da daha fazlasının sağlanamadığı durumda, diskriminant analizi optimum bir sınıflama ortaya koyamayacaktır. Yazında, diskriminant analizinin bu varsayımlar karşısında sağlamlığı tartışmalı bir konudur. Üçüncü varsayımda , bağımsız değişkenler normal dağılıma sahip olduğu belirtilmiştir. Ancak yapılan araştırmalar , mali oranlar kullanılarak yapılan çalışmalarda mali oranların normal dağılıma uygunluk göstermemesi sebebiyle dağılımların normalden ziyade sağa çarpık olduğu göstermektedir. Bu durumda mali oranlar kullanılarak yapılan çalışmalar , değişkenlerin dağılımını normal dağılıma yaklaştırmayı hedeflemektedirler.

2.4.2.Birliktelik Kuralları

Veri kümesinde birliktelik analizi eş zamanlı oluşum , olay vb durumların tesbiti için kullanılan bir analizdir. Veri madenciliği sürecinde en olası modeli

⁵⁴ <http://www.statsoft.com/textbook/stdiscan.html#assumptions> 2003.

ortaya çıkarması açısından önemli analizlerden birisidir. Bu analizler genel olarak çıkardıkları kurallardan dolayı , müşteri davranışlarını tanımlama imkanı sağlamaktadır. Birliktelik kurallarının örnek uygulamaları şöyledir.⁵⁵

- Sepet analizi ,
- Direk satışta başka bir ürün önermek ,
- Kredi kartı sahtekarlıklarını ortaya çıkarmak ,
- Sağlık sigortalarındaki sahtekarlığı ortaya çıkarmak ,
- Standların dükkanda nasıl dizilmesi gerektiğini ortaya koymaktadır.⁵⁶

Bu uygulamalardan en yaygın olarak kullanılan analiz yöntemi , sepet analizidir. Bir ürünü alan müşterinin onun yanında başka ne aldığının tespiti için kullanılan bir analizdir. Sepet analizinde iki önemli kriter kullanılmaktadır. Bunlar;⁵⁷

- Destek kriteri ,
- Güven kriteridir.

Bu kriterler şekilde hesaplanmaktadır. Burada X ve Y mallarının arasındaki ilişki incelenmektedir.

$P(X \cap Y) = \text{X ve Y mallarını almış müşteri sayısı} / \text{toplam müşteri sayısı}$

$P(X \cap Y)$, destek kriteri adı verilmektedir. Destek kriteri X malını alan bir müşterinin Y malını alma olasılığını yani X malını alıp sonra Y malını alma olasılığını gösteren bir değerdir. Bu değer bire yaklaştıkça güçlenmektedir.

$$P(X/Y) = \frac{P(X \cap Y)}{P(Y)}$$

⁵⁵ M.KANTARDZIC, Data Mining : Concept, Models, Methods , and Algorithms, John Wiley & Sons Publisher s 82 ,2003

⁵⁶ S.MITRA, T.ACHARYA, Data Mining : Multimedia, Soft Computing , and Bioinformatics , John Wiley & Sons Publisher s 268 ,2003

$P(X/Y)$, güven kriteri olarak tanımlanmaktadır.Bu kriter Y malını alan bir müşterinin X malını alma olasılığını göstermektedir.Aynı destek kriteri gibi , güven kriteri de bire yaklaştıkça güçlenmektedir.⁵⁸

Birliktelik kurallarına örnek vermek gerekirse , tatil için uçak bileti alan bir kimsenin , belli bir olasılıkla araba kiralaması verilebilir.

2.4.3. Sınıflandırma Ve Öngörü

Veri madenciliğinde sınıflandırma , belirli bir özelliğe göre veri kümesini sınıflara ayırmaya ve yeni elde edilen verilerin hangi sınıfa ait olduğunu gösteren bir analiz tekniğidir.

Öngörü ise , bir parametrenin geçmişteki değerlerine bakılarak gelecekte alabileceği değerleri tespit etme çabasıdır.Gerek sınıflandırma gerekse öngörü işlemleri için kullanılan analizler şunlardır.⁵⁹

- Karar Ağaçları (Decision Tree) ,
- Yapay Sinir Ağları (Neural Networks) ,
- K-Ortalamlar Yöntemi (K-Means) ,
- Genetik Algoritmalar ,
- Regresyon Analizi ,
- Zaman Serileri Analizidir.

2.4.3.1. Karar ağaçları

Karar ağacı , karar vericinin en iyi karara ulaşılabilmesi için yapılan gerek olasılık gerekse maksimum fayda esas alınarak düzenlenen bir tekniktir.Karar ağacı analizi , genellikle seçenekler üzerinde yapılan bir analiz türüdür.Bu analizin veri madenciliğinde kullanılma sebepleri şunlardır.⁶⁰

⁵⁷ M .BERRY , Data Mining Techniques , Wiley Publishers 2004 s289.

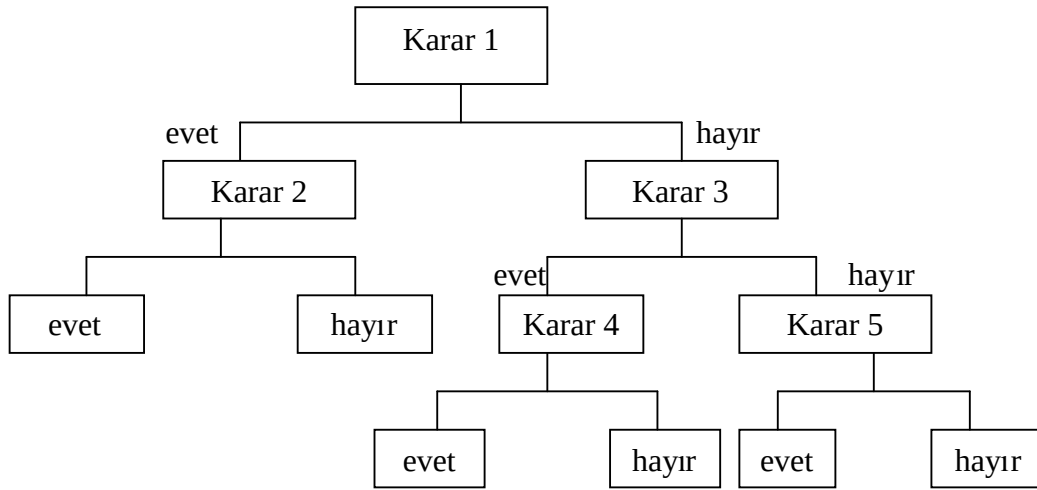
⁵⁸ S.MITRA , Data Mining Multimedia , Soft Computing and Bioinformatics , Wiley Publishers 2003 s71.

⁵⁹ J. BIGUS , Data Mining With Neural Networks ,McGraw Hill Publishers , 1996 s12.

⁶⁰ W.CHU, Foundations and Advances in Data Mining ,Springer Publisherss 2005 s25 , s100.

- Maliyeti azdır.
- Anlaşılması ve yorumlanması kolaydır.
- Veri tabanına kolay entegre edilebilmektedir.
- Güvenirliliği yüksektir.

Bu analizin uygulamasında veri seti iki kısma ayrılır.İlk veri seti karar ağacını oluşturmak , ikinci kısım ise karar ağacını kontrol etmek amaçlı kullanılmaktadır.Karar ağacı şu şekildedir.



Şekil 2.6.Karar ağacı şekli

Kaynak : M. KANTARDZIC ,Data Mining Concepts, Models, Methods, and Algorithms , John Wiley & Sons Publishers 2003 .

Karar ağacı analizinde kullanılan algoritmalar şunlardır.⁶¹

- C&RT ,
- CHAID ,
- C4.5 ,
- Quest .

⁶¹ M. KANTARDZIC ,Data Mining Concepts, Models, Methods, and Algorithms , John Wiley & Sons Publishers 2003 s142.

C&RT , ikili ağaç analizi olarakta bilinmektedir.Bu analiz doğru homojen yapıya ulaşılmıncaya kadar , veriyi ikili alt kümelere ayırmaktadır.1984 yılında bulunan bu yöntem kullanışlılık açısından , büyük verilerde çok uzun zaman aldığından dolayı pek tercih edilmemektedir.

CHAID , bu karar ağacı tekniği grupları oluşturmak için Ki-Kare analizinden yararlanmaktadır.Bu algoritma esasen benzer verileri birleştirerek farklı olanlarla analize devam eder.Böylece ulaşılmak istenen değerler kesin ve kolay bir şekilde elde edilir. Karar ağacının ilk dallarını oluşturmak için en iyi tahmin edici değişkenler seçilir.Bu algoritmada değişkenler sürekli ise F-Testi değerleri , nominal ya da ordinal bir değişkense Ki-Kare Test değerleri kullanılır. CHAID algoritması ikili bir algoritma değildir.Kullanım açısından yaygın ve popüler bir uygulamadır.

C4.5 algoritması en iyi karar ağacı algoritmasıdır. Karar ağacını oluştururken kayıp verileri hesaba katmaz .Özellikle hasas ve anlamlı veriler elde etmek için kullanılmaktadır.⁶²

Quest algoritması , ikili bir algoritmadır.En önemli özelliği gerek değişken gerekse ayırım noktalarının belirtilebilmesidir.1997 yılında bulunan bu yöntem yaygın bir karar ağacı algoritmasıdır.⁶³

2.4.3.2.Yapay sinir ağları

Yapay sinir ağları , öğrenme yolu ile yeni bilgi elde edebilmeyi sağlayan bir tekniktir.Tarihi gelişimine bakıldığında , 1950 'lilerden günümüze uzanan hızlı gelişmeler olduğu görülebilmektedir. Yapay sinir ağlarının yapısına bakıldığında;

W_i : kendi ağırlık değeri ,

I_i : n adet girdi değeri ,

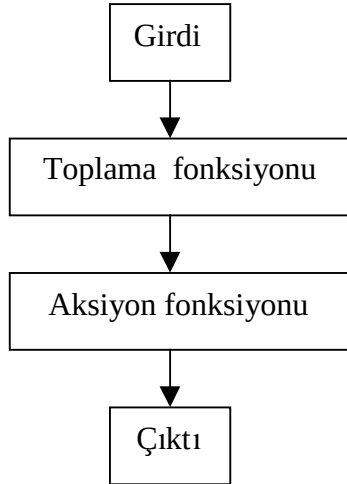
⁶² M. KANTARDZIC ,Data Mining Concepts, Models, Methods, and Algorithms , John Wiley & Sons Publishers 2003 s154.

⁶³ N.YE , The Handbook Of Data Mining , Lawrence Erlbaum Associates Publisherss 2003 s3.

$\sum_{i=1}^n I_i W_i$: Toplama fonksiyonu ,

$F(\sum_{i=1}^n I_i W_i)$: Aksiyon fonksiyonu bu şekilde hesaplanır.⁶⁴

Yapay sinir ağların süreci şu şekildedir.



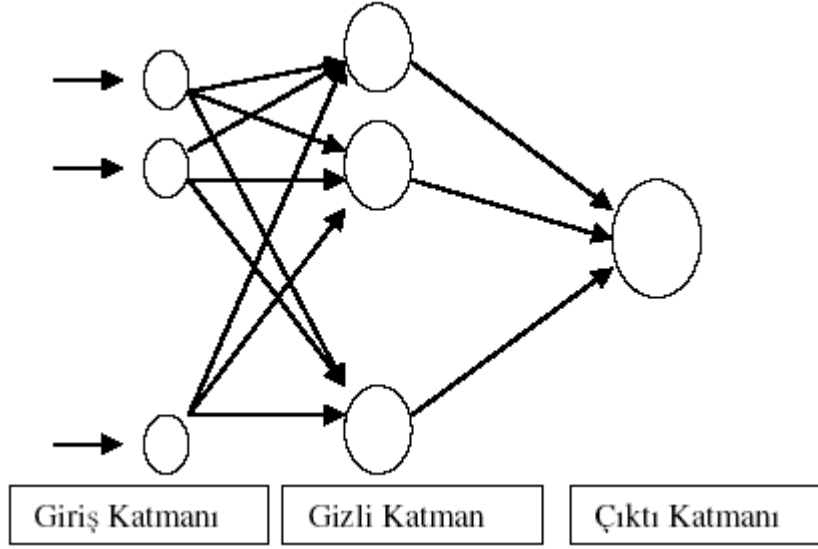
Şekil 2.7. Yapay ağ süreci

Kaynak : http://tr.wikipedia.org/wiki/Yapay_sinir_a%C4%9Flar%C4%B1 2007.

Bir yapay sinirin öğrenme yeteneği , kullanılan ağırlık oranıyla doğrudan alakalıdır.Süreçte kullanılan girdiler , dışarıdan elde edilen bilgilerdir.Toplama fonksiyonu bir hücreye gelen net girdi miktarı olarak tanımlanabilir.Aktivasyon fonksiyonu , bu fonksiyon öğrenilme sonucu oluşan değerlerin ortaya çıkarılması için kullanılan bir fonksiyondur.Son olarak çıktı ise , aktivasyon fonksiyonundan elde edilen değer olarak tanımlanabilir.⁶⁵Yapay sinir ağlarının katman olarak işleyişi aşağıdaki şekilde gösterilmektedir.

⁶⁴ M. KANTARDZIC ,Data Mining Concepts, Models, Methods, and Algorithms , John Wiley & Sons Publishers 2003 s222.

⁶⁵ W.CHU, Foundations and Advances in Data Mining ,Springer Publisherss 2005 s23.



Şekil 2.8.Yapay sinir ağlarının katmanları

Kaynak : http://tr.wikipedia.org/wiki/Yapay_sinir_a%C4%9Flar%C4%B1 2007.

Yapay sinir ağları , ağın yapısına göre sınıflandırıldığında iki gruba ayrılmaktadır.Bunlar ;⁶⁶

- İleri beslemeli ağlar ,
- Geri beslemeli ağlardır.

Yapay sinir ağları öğrenme türüne göre sınıflandırıldığında ise , denetimli ve denetimsiz öğrenme olarak iki gruptan oluşmaktadır.

Yapay sinir ağlarının kullanıldığı yerlere bakıldığında ilk olarak akla genel veri madenciliği olmasına karşın , birçok alanda yapay sinir ağları kullanılmaktadır.Bunlar ;⁶⁷

- Kredi kartı sahtekarlığının tespiti ,
- Kalite kontrol ,
- Üretim planlama ve çizergeleme ,
- Ürünlerin performans tahmini gibi konularda kullanılmaktadır.

⁶⁶ N.YE , The Handbook Of Data Mining , Lawrence Erlbaum Associates Publisherss 2003 s71.

⁶⁷ <http://www.backpropagation.netfirms.com/ysauygulama.htm> 2005.

2.4.3.3. K-ortalamalar yöntemi (K-Means)

K-en yakın komşu yöntemi 1967 yılında Mac QUEEN tarafından bulunmuştur.Kümeleme algoritması olan k-ortalamaları , k sayıda veriden küme oluşturmaktadır.Verilen ağırlıklı ortalamalara bakılarak en yakın değerleri birbirine atayarak kümeler oluşturmaktadır.

Bu yöntem ilk olarak veri setini k tane küme olacak şekilde ayırır.K değeri analizi yapan kişi tarafından belirtilmektedir.Daha sonra her veri , merkez noktalara en yakın olduğu kümeye dahil edilerek kümeleme işlemi yapılır.Oluşturulan kümelerin tekrar ağırlıklı ortalaması hesaplanarak merkez değerleri yeniden oluşturulur.Böylece elde edilen kümeler homojen bir şekilde oluşturulmuş olurlar.⁶⁸

2.4.3.4.Genetik algoritmalar

Genetik alagoritmalar , çok değişkenli fonsiyonları optimize etmeyi amaçlayan sayısal bir araçtır.Bu algoritma parametre yerine onların kodlanmış biçimlerini kullanarak en iyiye ulaşmaya çalışır.Yapay zekanın bir uygulaması olan genetik algoritma , kısa sürede çözümleri ortaya çıkarması bakımından önemli bir tekniktir.⁶⁹

Genetik algoritmalarının uygulama alanlarına bakıldığında ise , kromozon ve gen hesaplamaları , havuz problemi çözümü , uygunluk fonksiyonunun hesaplanması gibi genel problemlerde uygulanabildiğini görmekteyiz.⁷⁰

Bu uygulama alanlarını sınıflandırmamız gerekirse ;

- Deneysel çalışmaların optimizasyonu ,

⁶⁸ M. KANTARDZIC ,Data Mining Concepts, Models, Methods, and Algorithms , John Wiley & Sons Publishers 2003 s134.

⁶⁹ M. KANTARDZIC ,Data Mining Concepts, Models, Methods, and Algorithms , John Wiley & Sons Publishers 2003 s222.

⁷⁰ M .BERRY , Data Mining Techniques , Wiley Publishers 2004 s421.

- Pratik endüstriyel uygulamalar ,
- Sınıflandırma çalışmalarıdır.

Genetik algoritmaların veri madenciliğinde uygulamalarına bakıldığında ise ;

- Kümeleme ,
- Model tahmini ,
- İlişki kuralları oluşturma ,
- Müşteri grupları oluşturma ,
- Sınıflandırma çalışmalarında kullanılmaktadır.

2.4.3.5.Regresyon analizi

Tahmin yöntemlerinden biri olan regresyon analizi bir bağımlı değişkenin birden fazla bağımsız değişkenle olan ilişkisini gösteren basit bir fonksiyondur.Bu analizde amaç geçmişte ilişkisi olduğu varsayılan bağımlı ve

$$Y=\alpha+\beta X+\varepsilon$$

bağımsız değişkenler arasında bir ilişki fonksiyonu oluşturarak gelecekte alabilecekleri değerleri tahmin etme esasına dayanmaktadır.Regresyon fonksiyonu şu şekildedir.⁷¹

Örneklem içinse formül şu şekli almaktadır.

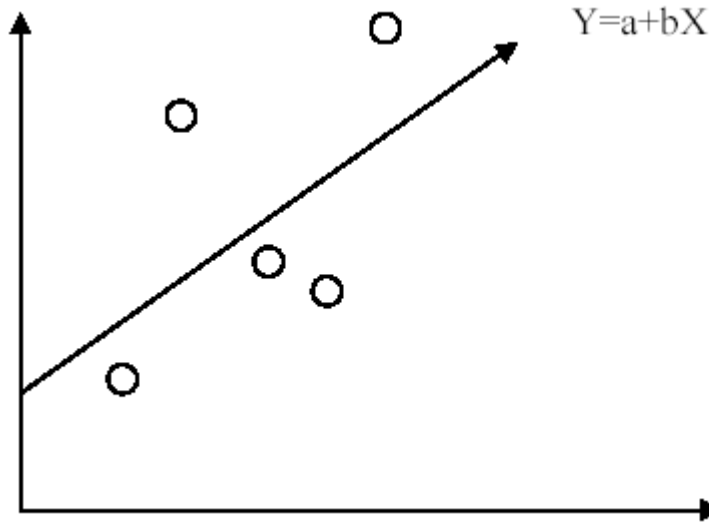
$$Y=a+bX+e$$

$$b = \frac{\sum (x_j - \bar{x})(y_j - \bar{y})}{\sum (x_j - \bar{x})^2}$$

$$a = \bar{y} - b\bar{x}$$

Doğrusal regresyon modeli grafikte şu şekilde gösterilir.

⁷¹ N.ORHUNBİLGE , Uygulamalı Regresyon ve Kolerasyon Analiz , İÜ. Yayınları 1996 s14.



Şekil 2.9.Doğrusal regresyon grafiği

Kaynak : Ö.SERPER,Uygulamalı İstatistik 2 , Ezgi Yayınevi 2000 s220

$Y = a + bx + e$ olan örneklem için doğrusal modelde a sabit katsayı , b bağımsız değişkenin modeldeki ağırlı , e hata terimleri olarak tanımlanmaktadır.Bu modelin katsayıları şu şekilde tahmin edilmektedir.⁷²

Tahmin edilen $Y = a + bx + e$ modelinin güvenirliliği F-testi olan ANOVA ile araştırılır.Araştırılan hipotez aşağıdaki gibidir.⁷³

H_0 : X ve Y değişkeni doğrusal arasında ilişki yoktur.

H_1 : X ve Y değişkeni doğrusal arasında ilişki vardır.

Bu hipotez şu formülle test edilir.

Tablo 2.1.Anova testi hesap tablosu

Kaynak	S.D.	K.T.	K.O.	F
Regresyon	k	$\sum(\hat{Y} - \bar{Y})^2$	K.T./S.D.	$\frac{KO_{Reg}}{KO_{Artık}}$
Artık	n-k-1	$\sum(Y - \hat{Y})^2$	K.T./S.D.	
Genel	n-1	$\sum(Y - \bar{Y})^2$		

Kaynak : K.ÖZDAMAR , Paket Programlar ile İstatiksel Veri Analizi , Kaan Kitabevi 1999 s274

⁷² M.SPIEGEL, İstatistik , Bilim Teknik Yayınevi 1995 s143.

⁷³ K.ÖZDAMAR , SPSS ile Bioistatistik , Kaan Kitabevi 2001 s313.

Hesaplanan F değeri F-test değerinden büyük ise H_0 hipotezi red edilir H_1 hipotezi kabul edilir.Yani değişkenler arasında doğrusal bir ilişkinin var olduğu söylenebilmektedir.

Eğer bağımsız değişkenlerin sayısı birden fazla olduğu durumlarda regresyon modeli şu şekli almaktadır.⁷⁴

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \varepsilon$$

Yukarıdaki modelin değişkenlerinin tanımı şu şekildedir.

- Y Bağımlı değişken ,
- X_1 Birinci bağımsız değişken ,
- X_2 İkinci bağımsız değişken ,
- β_0 Modelin sabit katsayısı ,
- β_1 X_1 'in modeldeki ağırlığını gösteren katsayı ,
- β_2 X_2 'nin modeldeki ağırlığını gösteren katsayı ,
- ε Hata terimlerini göstermektedir.

Çoklu regresyon modelinin uygulanabilmesi için bazı varsayımlar modeldeki değişkenlerin uyması gerekmektedir.Aksi takdirde değişkenler ile regresyon modeli kurulamamaktadır.Bunlar ;⁷⁵

- $\varepsilon \sim (0, \sigma_\varepsilon)$, hata terimleri normal dağılıma uygun olmalı,
- $Cov(X_i, X_j) = 0$, bağımsız değişkenler arasında hiçbir ilişki olmaması gerekmektedir.

Örnekten hesaplanan regresyon denkleminin verilere uyum düzeyini, dolayısıyla denklemin başarısını ölçmede belirleme katsayısı R^2 istatistiği kullanılmaktadır. Belirleme katsayısı, regresyon denkleminin başarısını ölçme yanında, denklemin tahmin gücünü de yansıtan bir parametre olması bakımından modelin , anlamlılık

⁷⁴ K.KURTULUŞ ,Pazarlama Alıştırmaları ,Avcıol Yayını ,1998 s390.

testinden sonra gelen en önemli parametredir. R^2 istatistiği şu şekilde hesaplanmaktadır.

$$R^2 = \frac{r_{x_1y}^2 + r_{x_2y}^2 - 2r_{x_1y}r_{x_2y}r_{x_1x_2}}{1 - r_{x_2y}^2}$$

Çoklu regresyon modelinin testide , aynı doğrusal regresyon modelindeki gibidir.Hipotezlerin kullanımı aynıdır.

H_0 :Bağımsız değişkenlerle bağımlı değişken arasında ilişki yoktur.

H_1 : Bağımsız değişkenlerle bağımlı değişken arasında ilişki vardır.

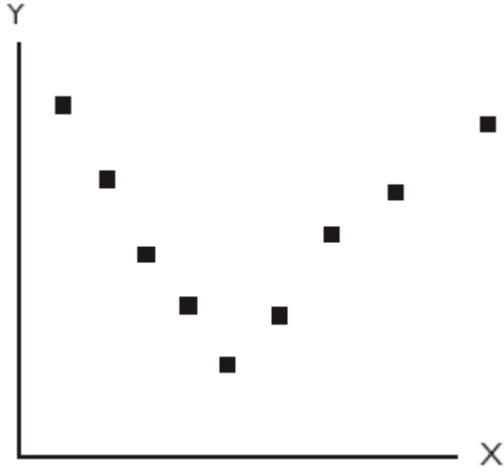
Bu hipotez aşağıda belirtilen F değeri ile test edilmektedir.

$$F = \frac{R^2 / k}{(1 - R^2) / (n - k - 1)}$$

k değişken sayısı , n gözlem sayısı olarak tanımlanmıştır . $F_{k-1,n-k-1}$ F-test istatistiği ile karşılaştırılan F değeri , test istatistiğinden büyük olması durumunda H_0 hiporezi red edilirken H_1 hipotezi kabul edilerek modelin anlamlılığına karar verilmektedir.

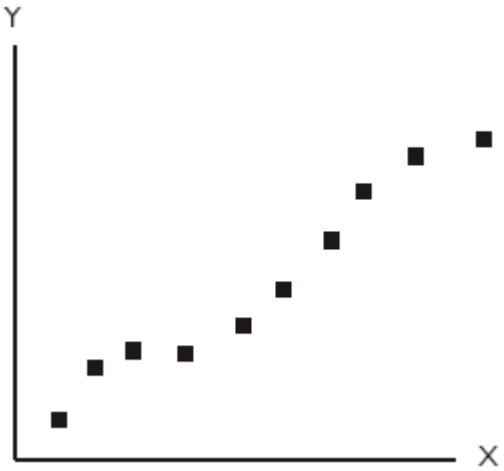
Regresyon analizinin bir diğer önemli konusu ise , regresyon modelinin belirlenmesi sürecidir.Bu süreç verilerin bir x-y eksenli bir grafikte dağılımının ne olduğuna bakılarak hangi regresyon modelinin uygulanacağına karar verilen bir süreçtir.Bu süreçte kullanılan grafiğe , serpilme diyagramı denmektedir.Doğrusal model ve eğrisel model için serpilme diyagramı şu şekildedir.

⁷⁵ D.LAROSE , Data Mining Methods and Models , Wiley Publisherss 2006 s34 .



Şekil 2.10.Eğrisel ilişkiyi gösteren serpilme diyagramı

Kaynak : D.GUJARATI , Temel Ekonometri , Literatür Yayınevi 1999 s17-18



Şekil 2.11.Doğrusal ilişkiyi gösteren serpilme diyagramı

Kaynak : D.GUJARATI , Temel Ekonometri , Literatür Yayınevi 1999 s17-18

Serpilme diyagramında noktaların durumu ve genel seyri, iki değişken arasında ilişki olup olmadığını ; varsa ilişkinin ne tür bir fonksiyon tipine uyduğunun belirlenmesinde yardımcı olur.

Serpilme diyagramı , yalnız ilişkinin olup olmadığını ve fonksiyonel şeklini göstermekle kalmaz, ilişkinin derecesi hakkında da bilgi verir. Bunun için, noktaların

en dışta kalanları birleştirilerek, bir şekil elde edilir. Söz konusu şeklin durumuna göre ilişkinin derecesi hakkında tahminde bulunulur. Eğer şekil, oldukça dar bir elipse benziyorsa, ilişki kuvvetlidir. Elips genişledikçe ilişki zayıflamaktadır.

Regresyon analizinde tahmin edilen parametrelerin güven aralıklarını hesaplamak için ilk olarak aşağıdaki formüllerden parametrelerin standart hatalarının hesaplanması gerekmektedir.⁷⁶

$$S_{b_1} = \frac{\sqrt{\frac{\sum (Y_i - \hat{Y}_i)^2}{n-k}}}{\sqrt{\sum (X_i - \bar{X})^2}} \quad S_{b_1}, \beta_1 \text{ 'in tahmini değerinin standart hatasıdır.}$$

$$H = \sqrt{\frac{\sum (Y_i - \hat{Y}_i)^2}{n-k}}$$

$$S_{b_0} = H * \sqrt{\frac{1}{n} + \frac{\bar{X}^2}{\sum (X_i - \bar{X})^2}}, \quad S_{b_0} \text{ ise } \beta_0 \text{ 'in standart hatasıdır.}$$

t örnekleme dağılımına göre β_0, β_1 'in güven aralıkları şu şekilde hesaplanmaktadır.⁷⁷

$$\beta_0 \pm S_{b_0} * t_{n-2}$$

$$\beta_1 \pm S_{b_1} * t_{n-2}$$

⁷⁶ H.TATLIDİL , Uygulamalı Çok Değişkenli İstatistiksel Analiz , Hacettepe ünv. 1996 s381.

⁷⁷ M.AYTAÇ , Matematiksel İstatistik , Ezgi Kitabevi 1999 s345 .

β_0, β_1 'in güven aralıkları 0.95 güvenle hesaplanmaktadır.

2.4.3.6.Zaman serileri analizi

Zaman serileri analizi , zamana bağlı olarak verileri analiz ederek parametrelerin gelecekteki değerlerini tahmin etmeye yarayan bir analiz tekniğidir.Bu analizin regresyon analizinden temel farkı , zaman esasına göre verilerin değerlendirmesidir.

Bu zaman serileri , yıllara göre milli gelirin, istihdamın veya ihracatın kaydettiği gelişme gibi iktisadi zaman serileri olabileceği gibi, bir mağazanın aylık satışlarını, mevsimlere göre sıcaklık değerlerini veya tıp veya meteoroloji konuları ile ilgili serilerde olabilmektedirler. İşletme ve iktisat alanlarında zaman serilerinin büyük önem taşımalarının sebebi, önceki dönemlere ait gözlemlerin incelenmesi ve ileriye dönük tahmin yapabilmenin mümkün olmasıdır.⁷⁸

Zaman serileri, bir yıldan fazla genelde 5, 10 , 15, ve 20 yıla dayanan uzun dönem planlama ve tahminleme için kullanılmaktadır.

Zaman seriler analizi uygulayabilmek için , seriyi oluşturan bileşenlerin ayrıştırılması gerekmektedir. Bir seriyi bileşenlerine ayırmak için , kapsadığı dört bileşen arasında belli bir ilişki bulunduğu varsayılmalıdır. Bunun için kullanılan yöntem , zaman serisinin birkaç bileşenini toplamı ya da çarpımından meydana geldiği varsayımıdır. Zaman serisi şu dört etkiye maruz kalmaktadır.⁷⁹

- Uzun dönemli genel trend (T) ,
- Konjonktür dalgalanmaları (C) ,
- Mevsimsel dalgalanmalar (S) ,
- Varyasyon ve düzensiz rastgele hareketlerdir. (I)

⁷⁸ R.S.TSAY ,Analysis Of Financial Time Series ,Wiley publishers 2005 s24.

⁷⁹ M.SPIEGEL, İstatistik , Bilim Tekik Yayınevi 1995 s225.

Bu etki bileşenleri kullanılarak elde edilen denklem şu şekildedir.

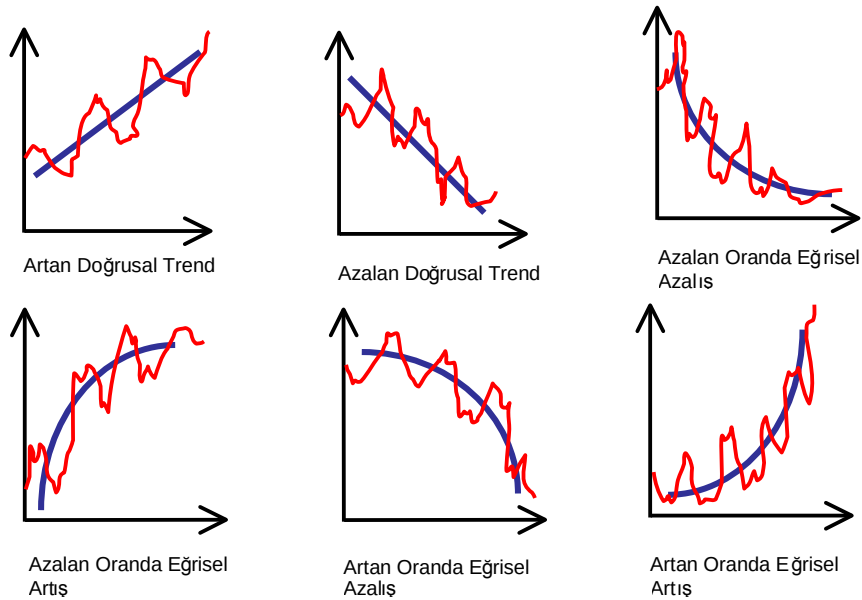
$$Y = T * C * S * I$$

Zaman serilerinin amacı , bu dört elemanın her birinin aldığı değerlerin model üzerinde ne ölçüde etkili olduğunu araştırmaktır.

Uzun dönemli genel trend (T)

Uzun dönemli genel trend , olayın bağlı olduğu temel ve taşıdığı özellikler bakımından uzun dönemde çok fazla değişmediği için , bu trendlere uzun dönem genel trendleri denmektedir.

Trend , belirli bir zaman aralığında serinin uzun dönemde belirli bir yöne doğru gösterdiği gelişme veya ilerlemedir. Trend analizi bir uzun dönem analizi olduğundan verilerin aylık veya mevsimlik olarak verilmiş olması tahlilin sonucunu etkilemeyecektir. Trendin yön ve şiddeti her zaman sabit kalmaz. Trend doğrusal ya da eğrisel olabilir. Bu trendler şu şekildedir.⁸⁰

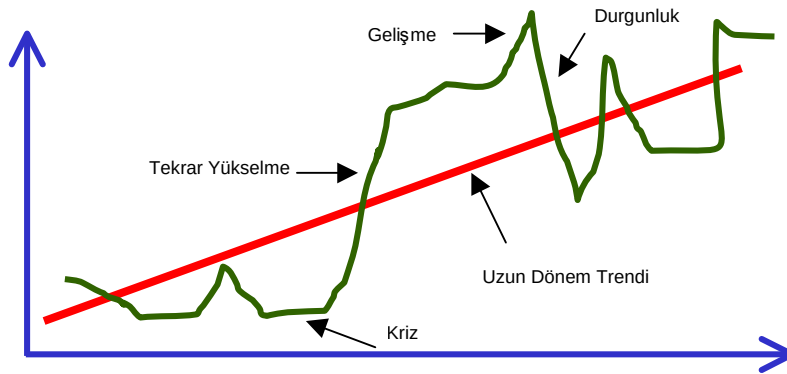


Şekil 2.12. Zaman serisi trendleri

Kaynak : <http://img92.imageshack.us/img92/2669/zamanserileri1fv0.png> 2005.

Konjonktür dalgalanmaları (C)

Konjonktür dalgalanmalar , genel trend doğrusundan verilerin yapmış olduğu sapmalara konjonktürel dalgalanmalar denmektedir.Bu hareketler mevsim dalgalanmalarına benzer şekilde periyodik olarak tekrar etmekte iseler de periyotların uzunluğu ve sürelerin belirsizliği ile dikkati çekmektedir.Konjonktürel dalgalanmalar aşağıdaki grafikte gösterilmektedir.⁸¹



Şekil 2.13.Konjonktürel dalgalanma grafiği

Kaynak : Ö.SERPER,Uygulamalı İstatistik 2 , Ezgi Yayınevi 2000 s329

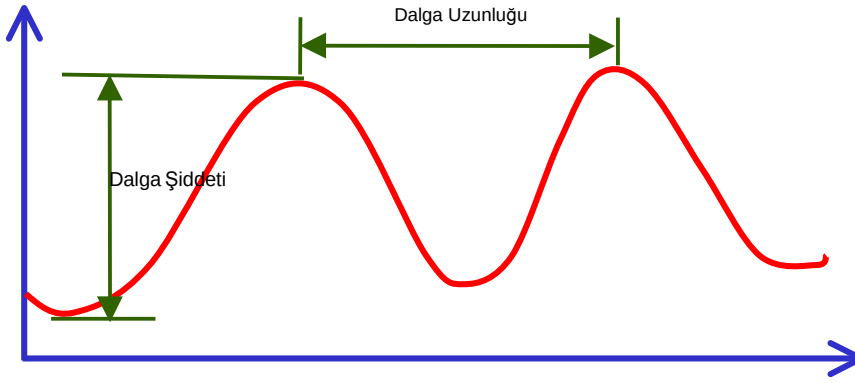
Mevsimsel dalgalanmalar (S)

Mevsimsel dalgalanmalar ise , ekonomik olayların zaman içinde izlendiği doğal ve sosyal nedenlerden dolayı , mevsime göre oluşan değişimler mevsimsel dalgalanmalar olarak adlandırılır. Mevsimsel dalgalanmaların dalga uzunluğu 12 ay olmaktadır.Mevsimsel dalgalanmaya örnek verilmesi gerekirse , ekonominin okulların açıldığı zamanki hareketliliği gösterilebilir.Mevsimsel dalgalanma şu şekildedir.⁸²

⁸⁰ M.SPIEGEL, İstatistik , Bilim Teknik Yayınevi 1995 s474- 483.

⁸¹ Ö.SERPER ,Uygulamalı istatistik 2 , Ezgi Kitapbevi 2000 s331.

⁸² D.GUJARATI , Temel Ekonometri , Literatür Yayınevi 1999 s709 .



Şekil 2.14.Mevsimsel dalgalanma

Kaynak : Ö.SERPER,Uygulamalı İstatistik 2 , Ezgi Yayınevi 2000 s333

Trendin hesaplanmasında kullanılan yöntemler şunlardır.⁸³

- Basit grafik yöntemi ,
- Hareketli ortalamalar yöntemi ,
- En küçük kareler yöntemidir.

Basit grafik yöntemi

Bu metoda göre , inceleme konusu olan zaman serisi gözlem sayısı itibariyle iki eşit kısma bölünür ve her kısımdaki gözlemler için birer aritmetik ortalama hesaplanır. Bu ortalama değerleri grafiğe işaretlendikten sonar araları bir doğru ile birleştirilerek bir trend doğrusu elde edilir. Tek ve çift sayıdaki örneklemeler için , ayırım şu şekilde olmaktadır.

- Seri çift sayılı ise , seri eşit olarak tam ortadan iki parçaya ayrılır.
- Seri tek sayılı ise , tam ortadaki eleman dikkate alınmadan seri iki eşit parçaya ayrılır.

⁸³ M.SPIEGEL, İstatistik , Bilim Tekik Yayınevi 1995 s232- 247.

Bu yöntemin uygulanabildiği veriler , doğrusal bir trende sahip olmaları ve her iki kısmında da konjonktürel etkilerinin aynı olması gerekmektedir.

Hareketli ortalamalar yöntemi

Hareketli ortalamalar yöntemi , konjonktürel ve mevsimsel dalgalanmaları yok etmek amacıyla kullanılmaktadır.

Hareketli ortalamalar bir zaman serisine ait her değerin yerine, o değer ve daha önce ve sonra gelen birkaç değerın ortalamasının bu değer yerine yazılması suretiyle elde edilen bir zaman serisidir. Örneğin yıllık verilerde üçer yıllık hareketli ortalamalar hesaplamak istiyorsak, her yılın değeri bir önceki ve bir sonraki değerlerle toplanarak üçe bölünür ve bulunan değer fiili değerin yerine konulur. Benzer şekilde daha çok yılı veya ayları içine alan hareketli ortalamalar hesaplanabilmektedir.

Hareketli ortalamalar trend hesabının sağlıklı olabilmesi için gerekli koşullar şunlardır.

- Olayın trendi doğrusal olması ,
- Serideki dalgaların uzunluğu aynı olması ,
- Serideki dalgaların şiddeti aynı olmalı ,
- Hareketli ortalamalar yönteminin aşamaları şunlardır.

İlk olarak , hareketli ortalama yönteminin uygulama sürecinde dalga uzunlukları bulunur. Dalga uzunlukları minimumdan minimuma ya da maksimumdan maksimuma olan uzaklık olarak hesaplanır. Daha sonra kaçarlı hareketli ortalama hesaplanacağı bulunur. Bunun için kullanılan formül aşağıda verilmiştir.

$$KHO = (\text{Dalga Uzunlukları Toplamı}) / (\text{Dalga Sayısı})$$

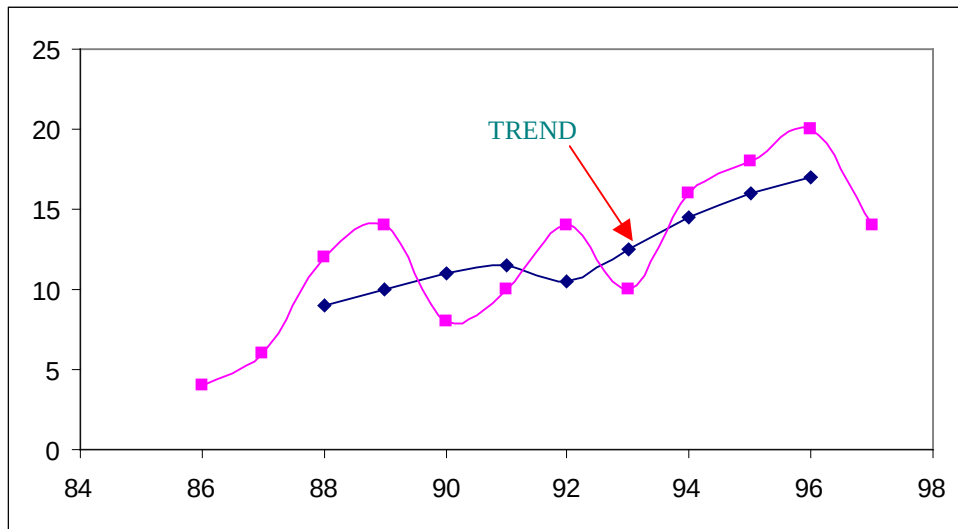
Hareketli ortalama sayısı tek sayı bulunursa yapılan hesaba $(n-1) / 2$ eleman az alınır. Örneğin üçerli hareketli ortalama hesaplandığında $(3 - 1) / 2 = 1$ bulunur . Bunun anlamı, baştan ve sondan birer trend değerinin hesaplanmayacağıdır.

Hareketli ortalama sayısı çift sayı bulunursa yapılan hesaba (n) / 2 eleman az alınır. Dörderli hareketli ortalama bulunduğunu varsayarsak $4/2 = 2$ elde edilir. Bu sayı, baştan ve sondan ikişer trendin hesaplanmayacağını gösterir. Üçerli hareketli ortalamalar ile trend değerleri şu şekilde hesaplanır.

Tablo 2.2. Hareketli ortalamaların hesaplanma yöntemi

Yıllar	Değerler	Üçerli hareketli ortalamalarla trend değerleri
1990	Y_1	$Y_2 = (Y_1 + Y_2 + Y_3) / 3$
1991	Y_2	$Y_3 = (Y_2 + Y_3 + Y_4) / 3$
1992	Y_3	
...	...	
1995	Y_6	$Y_7 = (Y_6 + Y_7 + Y_8) / 3$
1996	Y_7	...
.	.	...

Hareketli ortalamalar yöntemiyle hesaplayarak trend aşağıdaki şekilde gösterilmektedir.



Şekil 2.15. Hareketli ortalama yöntemiyle oluşturulan trend

Kaynak : M.SPIEGEL , İstatistik , Bilim Teknik Yayınevi 1995 , s473

Hareketli ortalamalar için gerekli şartlar sağlanamazsa, doğru sonuçlar elde edilmez. Hareketli ortalamalar tekniği ile elde edilen sonuçlar serideki uzun ve şiddetli dalgaların etkisi altındadır. Serinin başındaki ve sonundaki bazı elemanların hareketli ortalaması bulunamaz.

En küçük kareler yöntemi

Bu yöntem , zaman ile sonuçlar arasındaki fonksiyonel ilişkiyi ortaya çıkarmaktadır.⁸⁴

Trendi en iyi tanımlayacak fonksiyon tipinin seçilmesi için veriler X eksenini zaman olmak üzere bir diyagrama aktarılır. Oluşturulan diyagrama bakılarak trendin fonksiyonu belirlenmeye çalışılır. Trend fonksiyonları şu şekillerde olabilmektedir.⁸⁵

- Doğrusal Denklem ($Y = a + bX$), bu eşitlik genelde artma ve azalmaları sabitlik gösteren seriler için kullanılır.
- Parabol Denklemi ($Y = a + bX + cX^2$), bu eşitlik, yükselmeyi alçalmanın takip ettiği yön değiştirmeyi veren veriler için kullanılır.
- Üstel Denklem ($Y = a.bX$), bu denklem, artma yada azalma oranı sabit olan serileri hesaplamayı sağlar.
- Hiperbol Denklemi ($1/Y = a + bX$)
- Kübik Denklem ($Y = a + bX + cX^2 + dX^3$), serinin grafiğinde alçalma ve yükselmeleri içeren iki bükülme varsa kullanılır.
- Geometrik Denklem ($Y = a . X^b$)

Yukarıdaki denklemlerde kullanılan parametreler şunlardır.

⁸⁴ Ö.SERPER ,Uygulamalı istatistik 2 , Ezgi Kitapbevi 2000 s340.

Y : Değerler ,

X : Yıllar ,

a, b, c : Modelin katsayılarını temsil etmektedir.

Doğrusal denklem yöntemi kullanılarak trend şu şekilde hesaplanır.

$$\begin{aligned}\sum Y &= na + b \sum X \\ \sum XY &= a \sum X + b \sum X^2\end{aligned}$$

Seri tek sayıda ise , zaman bloğunda, tam ortadaki değer orijin olarak alınarak, sıfır değeri konur. Bundan öncekilere -1,-2,-3, vb; sonrakilere ise +1, +2, +3 değeri verilir.

Seri çift sayıda ise , zaman sütununda tam ortadaki iki değerden büyük olana +1, küçük olana – 1 değeri verilerek +3, +5 , ... ve -3, -5, ... konularak işleme devam edilir. Doğrusal denklem yöntemi kullanılarak standart hata şu şekilde hesaplanır.⁸⁶

n < 30 ise

n ≥ 30 ise

Formül;

$$S_{yx} = \sqrt{\frac{\sum (Y - Y')^2}{n - 2}}$$

$$S_{yx} = \sqrt{\frac{\sum (Y - Y')^2}{n}}$$

Formül ;

$$S_{yx} = \sqrt{\frac{\sum Y^2 - a \sum Y - b \sum XY}{n - 2}}$$

$$S_{yx} = \sqrt{\frac{\sum Y^2 - a \sum Y - b \sum XY}{n}}$$

⁸⁵ B, KEDEM , Regression Models For Time Series Analysis , Wiley Publishers 2002 s1-4.

S_{yx} : Tahminin standart hatasını ,

Y : Bağımlı değişkenin gözlenen değerini ,

Y' : Bağımlı değişkenin modelden hesaplanan değerini ,

n : Gözlem sayısını ,

a ve b : model katsayıları olarak ifade etmektedir.

Parabol denklem için ($Y = a + bX + cX^2$) a, b, c katsayıları şu yöntemle hesaplanır.

$$\begin{aligned}\sum Y &= na + b \sum X + c \sum X^2 \\ \sum XY &= a \sum X + b \sum X^2 + c \sum X^3 \\ \sum X^2 Y &= a \sum X^2 + b \sum X^3 + c \sum X^4\end{aligned}$$

n : Trend periyodundaki zaman ögesinin sayısını,

Y : Zamana göre serinin değerini ,

X : Zamanı temsil eder.

Parabol denklem yöntemi kullanılarak standart hatanın hesabı

$n < 30$ ise

$n \geq 30$ ise

Formül ;

$$S_{yx} = \sqrt{\frac{\sum (Y - Y')^2}{n - 3}} \quad S_{yx} = \sqrt{\frac{\sum (Y - Y')^2}{n}}$$

Formül ;

$$S_{yx} = \sqrt{\frac{\sum Y^2 - a \sum Y - b \sum XY}{n - 3}} \quad S_{yx} = \sqrt{\frac{\sum Y^2 - a \sum Y - b \sum XY}{n}}$$

S_{yx} : Tahminin standart hatasını ,

⁸⁶ Ö.SERPER ,Uygulamalı istatistik 2 , Ezgi Kitapbevi 2000 s348-349.

Y : Bağımlı değişkenin gözlenen değerini ,
Y' : Bağımlı değişkenin regresyondan hesaplanan değerini ,
n : Gözlem sayısını ,
a ve b : Regresyon katsayılarını ifade etmektedir.

Üstel denklem için ($Y = a.b^X$) , bağımlı değişkene ilişkin veriler, logaritmik duruma getirilirse doğrusal ilişki şu şekilde olur.⁸⁷

$$\log Y' = \log a + X \log b$$

$\log Y'$: Bağımsız değişken için trend değerinin logaritmasını,
 $\log a$: $X = 0$ durumunda $\log Y'$ nin değerini,
 $\log b$: Trend doğrusu eğrisinin logaritmasını ,
X : Bağımsız değişkeni ifade eder.

a ve b katsayıları şu formüllerden hesaplanır.

$$\begin{aligned} \sum \log Y &= n \log a & \sum \log Y &= n \log a + \sum X \log b \\ \sum X \log Y &= \sum X^2 \cdot \log & \sum X \log Y &= \sum X \log a + \sum X^2 \log b \end{aligned}$$

$$\begin{aligned} \log a &= (\sum \log Y) / n \\ \log b &= (\sum X \log Y) / \sum X^2 \end{aligned}$$

Mevsim dalgalanmalarına bakıldığında , mevsimsellik şu şekilde ölçülmektedir. Bir seride iklim ve sosyal sebeplerden dolayı, her yıl düzenli olarak tekrar eden periyodik değişimler “mevsim dalgalanmaları” olarak adlandırılmaktadır. Mevsimselliği hesaplamak için ilk önce şu düzeltme yapılmalıdır. Aylık verilerin gün sayısına göre ayarlanması gerekmektedir.

$$\text{Düzeltilme Faktörü} = (\text{Ort. Bir Aydaki Gün Say.}) / (\text{O Aydaki Gerçek Gün S.})$$

Ortalama bir aydaki gün sayısı ;

⁸⁷ B, KEDEM , Regression Models For Time Series Analysis , Wiley Publishers 2002 s62.

Normal Yıllarda : $(365)/12=30,41667$

Artık Yıllarda : $(366)/12=30,5$

Normal Bir yıl için Düzeltme Faktörü Hesaplanması

Şubat Ayı için: D.F. = $(30,41667)/28=1,08631$

30 günlük Ay için: D.F. = $(30,41667)/30=1,013889$

31 günlük aylar için: D.F. = $(30,41667)/31=1,981183$

Mevsim dalgalanmalarının ölçülmesinde kullanılan yöntemler aşağıda belirtilmiştir.⁸⁸

- Genel ortalamayı oranlama yöntemi ,
- Trende oranlama yöntemi ,
- Hareketli ortalamalar yöntemidir.

Genel ortalamayı oranlama yöntemi

Bu yöntemde her ayın aritmetik ortalaması bulunur. Daha sonra bu aritmetik ortalamalardan genel aritmetik ortalama hesaplanır. Bu yöntem şöyle formüllemektedir.

$$GOOY = (\hat{Y}_{aylık} / \hat{Y}_{genel}).100$$

(GOOY) Genel ortalamayı oranlama yöntemi ,

$$\hat{Y}_{genel} : \Sigma \hat{Y}_{aylık} / 12 \text{ dir.}$$

Bu yöntemde trendin durgun olduğu varsayılmaktadır. Halbuki ekonomik zaman serilerinin çoğunlukla bir hareketli bir trendi vardır. Bu yöntemle elde edilen sonuçlar, hem mevsimin hem de trendin ortak ölçüsünü vermektedir.

⁸⁸ Ö.SERPER ,Uygulamalı istatistik 2 , Ezgi Kitapbevi 2000 s363-372.

Trende oranlama yöntemi

Trende oranlama yöntemiyle mevsimin indeksi bulunarak seride ilk olarak trendin etkisi giderilir. En küçük kareler yöntemiyle elde edilen trend denklemiyle trend değerleri $\hat{Y}_{aylık}$ hesaplanır.

Trendin etkisi şu şekilde ortadan kaldırılır.

$$[(\hat{Y}_{aylık}) / (\hat{Y}'_{aylık})] \cdot 100$$

n yılının aylarına ilişkin oranların aritmetik ortalaması hesaplanır. Mevsim indeksi değerini veren S' lerin toplamın 1200 olması istenir. 1200 olmadığında düzeltme faktörü kullanılır. Trende oranlama yöntemi dezavantajları ise , mevsim ineksinin belirlendiği aylık değerlerde, ekonomiksel zaman serilerine etki eden dört faktöründe etkisi mevcuttur. Trende oranlama yönteminde, gerçek değerleri trende oranlamakla, trendin etkisi giderilir.

$$(Y_{aylık} / Y_{aylık}) - (T)(C)(S)(I) / (T) = (C)(S)(I)$$

Hareketli ortalamalar yöntemi

Mevsim indeksinin hesaplanmasında en iyi yöntem hareketli ortalamalar yöntemidir.

Mevsimsel dalgalanmaların dalga boyu 12 olması nedeniyle 12 şerli hareketli ortalamalarla trend değerleri bulunulur.⁸⁹

$[(\hat{Y}_{aylık}) / (\hat{Y}'_{aylık})] \cdot 100$ formülü ile gerçek değerlerin mevsimin etkisiyle hangi oranda değiştiği hesaplanabilmektedir.

Değişik yıllara ait aynı ay oranlarının aritmetik ortalaması hesaplanır.

$\Sigma S' \neq 1200$ ise düzeltme faktörüyle S' lerin ayarlanması gereklidir.

⁸⁹ D.GUJARATI , Temel Ekonometri , Literatür Yayınevi 1999 s737.

$$D.F = (1200) / \sum_{i=1}^{12} S$$

$$S = (S') (D.F.)$$

Konjonktürel dalgalanmaların ölçülmesinde ise iki hesaplama yöntemi kullanılır.

- Yıllık serilerde konjonktürün ölçülmesi ,
- Aylık serilerde konjonktürün ölçülmesidir.

Yıllık serilerde konjonktürün ölçülmesi

Yıllık seriler üzerinde mevsimin etkisi söz konusu değildir. Bundan dolayı üç faktörün sonucu olan gerçek değerler trend değerlerine bölündüğünde, trendin etkisi giderilecektir. Kalan kısım, konjonktürel ve düzensiz hareketler faktörlerinin etkisini verir.

$$(Y / Y') = [(T)(C)(I) / (T)] = (K)(I)$$

$$(K)(I) = [(Y) / (Y')] \cdot 100$$

Bu eşitlikte konjonktürün önceki periyotlardaki etkisi hesaplanabilir. Ancak konjonktürün gelecekteki etkisini tahminlemek mümkün değildir.

Aylık serilerde konjonktürün ölçülmesi

Aylık serilerde konjonktürün ölçülmesi iki aşamada gerçekleştirilir.

İlk aşamada , aylık serilerde konjonktür dalgalanmalarının etkisini belirleyebilmek için , ilk olarak aylık trend değerleri ilgili ayların mevsim indeksleri ile çarpılarak her ay için konjonktürün etkisini içermeyen normal değerleri hesaplanır.⁹⁰

$$(Y'_{\text{aylık}}) (S)$$

⁹⁰ Ö.SERPER ,Uygulamalı istatistik 2 , Ezgi Kitapbevi 2000 s373.

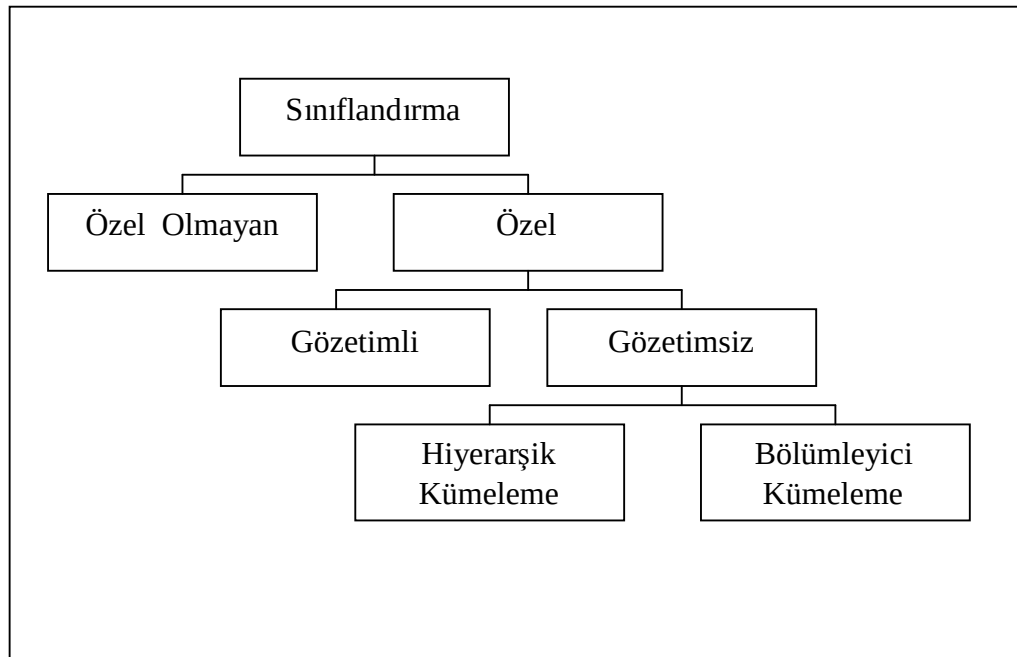
İkinci aşamada ise , gerçek değerler, normal değerlere bölünerek oran şeklinde ifade edilir.

$$K. I = [(Y_{aylık}) / (Y'_{aylık}) (S)] . 100$$

Böylelikle her ay gerçek değerlerin konjonktürel etki nedeniyle normalden ne kadar saptığı belirlenmiş olur.

2.4.4. Kümeleme Analizi

Kümeleme analizinde , sınıflandırma işlemini önceden belirten sınıflar ya da kriterler yoktur. Analiz sonucunda ortaya çıkan kümeler benzerliklerine göre oluşturulmuştur. Veri kümesinin sınıflandırma işlemleri şunlardır.



Şekil 2.16.Sınıflandırma ağacı

Kaynakça : H.TATLIDİL , Uygulamalı Çok Değişkenli İstatistiksel Analiz , Hacettepe ünv. s 330-38, 1996

Sınıflandırma yapısındaki ilk ayırım , özel (Exclusive) - özel olmayan (Nonexclusive) olarak ayrılmaktadır. Özel sınıflandırmada , her veri yalnızca bir gruba aittir. Özel olmayan sınıflandırmada ise , bir veri birden fazla gruba ait olabilmektedir. İkinci sınıflandırmada ise , gözetimli (Supervised) – gözetimsiz

(Unsupervised) olarak yapılmaktadır. Gözetimli sınıflandırmada veriler önceden tanımlı ve etiketlenmiş olarak yapılırken ; gözetimsiz sınıflandırmada herhangi bir etiket olmaksızın verilerin yakınlık matrisine göre sınıflandırılmaktadır. Kümeleme analizi bu sınıflandırma ağacına göre , Özel – Gözetimsiz bir sınıflandırma analizidir.⁹¹

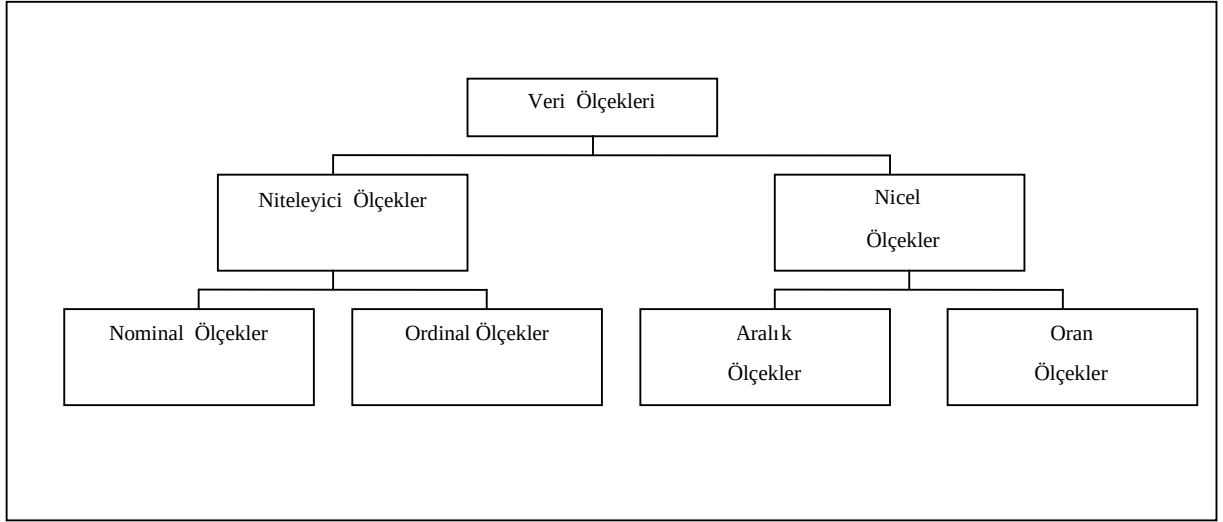
Kümeleme analizinde , kümeleri gruplamamıza yardımcı olan en önemli araç Benzerlik matrisidir (Proximity Matrix). Benzerlik matrisi, matris eksenleri sıfır değerinden oluşan d_{ij} değerleri kümelerin birbiriyle olan ilişkilerini gösteren bir matristir. d_{ij} değerleri , sıfıra yakın pozitif bir değer aldığında kümelerin benzerlikleri de o kadar artmaktadır. Benzerlik matrisi aşağıdaki şekilde gösterilmektedir.⁹²

$$D = \begin{bmatrix} 0 & d_{12} & d_{13} & \dots & d_{1n} \\ & 0 & d_{23} & \dots & d_{2n} \\ \text{Sim} & & 0 & \dots & \vdots \\ & & & 0 & d_{n-1,n} \\ & & & & 0 \end{bmatrix}$$

Şekil 2.17. Benzerlik matrisi

Benzerlik matrisinin elde edilme sürecine bakıldığında , veri ölçeklerine göre farklılık gösterdiği görülmektedir. Veri ölçeklerine bakıldığında ise iki ana başlıkta toplanmaktadır. Veri Ölçekleri Şekil 2.18 ‘de gösterilmektedir.

⁹¹ M. DEMİRALAY ,Hiyerarşik Kümeleme Metotları İle Veri Madenciliği , Yüksek Lisans Tezi, Marmara Üniv. ,2005



Şekil 2.18.Veri ölçekleri sınıflandırması

Kaynak : M.SPIEGEL, İstatistik , Bilim Tekik Yayınevi 1995 s14

Veri ölçekleri şöyle sıralanabilir.⁹³

Nominal Ölçekler

En az kısıtlı ancak en güçsüz ölçeklerdir.Bu ölçeklerde matematiksel işlemler yapılamazlar.Ölçek yalnızca bir kategori göstergesi olarak kullanılmaktadır.Ayrıca nominal ölçeklere uygulanan istatistiksel analizlerde oldukça kısıtlıdır.

Ordinal (Sıralı) Ölçekler

Nominal ölçeklerden daha güçlü ölçeklerdir.Bu ölçeklerin esas niteliği sıralayıcı bir ölçek olmasıdır.Ordinal ölçeklerin en çok kullanıldıkları durumlar insan davranışları , eğilimleri , tercihleri gibi subjektif konulardır.Ayrıca bu ölçekteki verilere kısıtlı bir istatistiksel analiz uygulanabilmektedir.

⁹² K.KURTULUŞ ,Pazarlama Araştırmaları ,Avcıol Yayını ,1998 s495.

⁹³ H.ARICI , İstatistiksel Yöntemler ve Uygulamaları , Hacettepe ünv. Yayınları 1998 s14-19.

Aralık Ölçekler

En önemli ölçekler arasındadır. Bu ölçekler başlangıç noktası keyfi seçilmesine rağmen , ölçü biriminin sabit olması en önemli özelliğidir. Ayrıca ölçeklenen verilerin aralıklarının bir birinin katı olması diğer önemli bir özelliğidir.

Oransal Ölçekler

En güçlü ölçeklerdir. En önemli özellikleri başlangıç noktasının ve ölçü biriminin değişmemesidir. Bu ölçekle ölçeklenen veriler birbirlerinin katı olabilmektedir.⁹⁴

Benzerlik matrisini elde etme yöntemini seçerken , veri kümesinin hangi ölçekle ölçeklendirildiğinin önemli olduğu bilinmektedir. Bu anlamda nicel ölçekler ve nitel ölçekler için benzerlik matrisi şu şekilde hesaplanmaktadır.⁹⁵

Nicel Ölçek için Minkowski Uzaklığı olarak bilinen bir yöntem kullanılmaktadır.

$$d_{\lambda}(x_i, x_j) = \left[\sum_{k=1}^p |x_{ik} - x_{jk}|^{\lambda} \right]^{1/\lambda}; \lambda \geq 1$$

formülünden elde edilmektedir. $\lambda = 1$ için bu formül Manhattan City –Block Uzaklığı olarak bilinen formüle dönüşmektedir. Bu formül şu şekildedir.

$$d_1(x_i, x_j) = \left[\sum_{k=1}^p |x_{ik} - x_{jk}| \right]; \lambda = 1$$

Nitel Ölçekli veriler için d_{ij} 'leri hesaplamakta Bhattacharyya uzaklığı formülü kullanılmaktadır. Bu formül ;

$$d(x_i, x_j) = \frac{1}{p} \left[\sum_{k=1}^p w_k |x_{ik} - x_{jk}| \right]$$

⁹⁴ K.KURTULUŞ ,Pazarlama Alıştırmaları ,Avcıol Yayını , s 338-346 ,1998

⁹⁵ H.TATLIDİL , Uygulamalı Çok Değişkenli İstatistiksel Analiz , Hacettepe üniv. 1996.

$$w_k = \frac{1}{m} \quad m = k\text{'nıncı değişkenin dağılım aralığıdır.}$$

Bu yöntemin kullanılması için bazı koşulların sağlanması gerekmektedir.Bunlar ;

- $d(x_i x_j) = d(x_j x_i)$ simetri özelliği ,
- $d(x_i x_j) \geq 0$ negatif olmama özelliği ,
- $d(x_i x_j) = 0$ ise $i = j$ tanım özelliği,
- $d(x_i x_j) \leq d(x_i x_1) + d(x_1 x_j)$ üçgen eşitsizliği özelliği bulunmalıdır.⁹⁶

Kümeleme analizinin metodolojisine bakıldığında ise , şu aşamalarda gerçekleşmektedir.⁹⁷

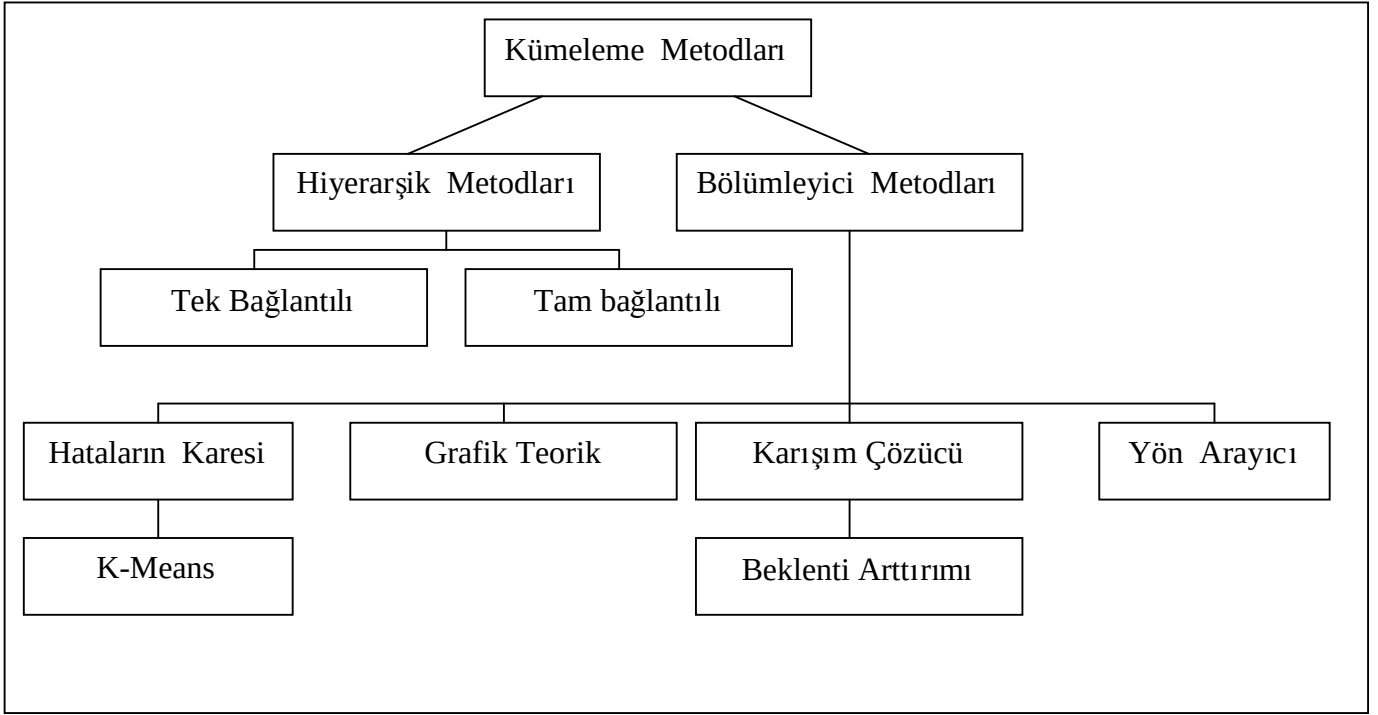
- Örgütsel sunum özellik seçme ,
- Örgütsel yakınlık ölçüsü tanımlama ,
- Kümeleme ,
- Veri soyutlama ,
- Sonuçları değerlendirme süreçleridir.

2.4.4.1.Kümeleme Metodları

Kümeleme analizinde kullanılan kümeleme metotları iki ana başlık altında toplanmaktadır.Bunlar ;

- Hiyerarşik metodlar ,
- Bölümeyici metodlardır.

⁹⁶ H.TATLIDİL , Uygulamalı Çok Değişkenli İstatistiksel Analiz , Hacettepe ünv. s 330-342, 1996.



Şekil 2.19.Kümeleme metodları hiyerarşisi

Kaynak : H.TATLIDİL , Uygulamalı Çok Değişkenli İstatistiksel Analiz , Hacettepe ünv. s 336-339 ,1996

Hiyerarşik Metodlar

Bu metotta , başlangıçta herbir veri bir kümeyken analizin sonunda tüm veriler bir küme oluştururlar.Hiyerarşik metodların süreçleri şu şekilde açıklanabilir.⁹⁸

- n tane birey n tane küme ile işe başlanır.
- En yakın iki küme (d_{ij} değeri en küçük olanlar) birleştirilir.
- Küme sayısı bir indirgenerek yinelenmiş uzaklıklar matrisi bulunur.
- İlk iki adım n-1 kez tekrarlanır.

⁹⁷ W.Hardle , Multivariate Statistics:Exercises And Solutions , Springer Publishers 2007 s210.

⁹⁸ N.H.TIMM ,Applied Multivariate Analysis,Springer Publishers 2002 s523-530.

Bölümleyici Metodlar

Bu metodlar hiyerarşik metodlardan daha güçlü tekniklerdir. Bu teknikler n tane veriden oluşan veri kümesini k tane kümeye ayıran metodlardır. Benzerlik matrisindeki değerlere göre benzerliklerine göre gruplanırlar.

Kümeleme analizinin son aşaması olarak , küme sayısının belirlenmesi gerekmektedir. Bunun için kullanılan formül pek güvenilir olmamakla beraber pratikte çok sık kullanılmaktadır. Küme sayısının hesaplanacağı formül şu şekildedir.

$$k \cong (n / 2)^{1/2}$$

k : küme sayısı ,

n : veri kümesindeki veri sayısıdır.

2.4.5. Sıradışılık Analizi

Sıradışılık analizi , veri kümesinde oluşan aykırı gözlemi tespit etmek için kullanılan bir analiz türüdür. Özellikle sahtekarlık ve dolandırıcılık gibi konularda en başta kullanılan analiz yöntemidir. Bu analiz iki teknikten oluşmaktadır.⁹⁹

- İstatistiksel tabanlı yöntem ,
- Yoğunluk tabanlı yöntemdir.

İstatistiksel yöntem , temel istatistik parametrelerindeki aykırılığı gözönüne alarak aykırı gözlemi tespit etmeye çalışırken , yoğunluk tabanlı yöntemde ise x-y ekseninde dağılan bir grafik şeklinde analiz yapılmaktadır.

⁹⁹ http://www.togaware.com/datamining/survivor/Outlier_Analysis.html 2007.

2.5.VERİ MADENCİLİĞİNDE KULLANILAN DİĞER ANALİZ TEKNİKLERİ

Veri madenciliği sürecinde uygulanan bir çok teknik olmasına karşın veri madenciliğinde kullanılan istatistiksel teknikler şunlardır.¹⁰⁰

- Faktör analizi ,
- Kanonik korelasyon analizi ,
- Lojistik regresyon analizi ,
- Çok boyutlu ölçekleme olarak tanımlayabiliriz.

2.5.1. Faktör Analizi

Çok değişkenli bir istatistik analizi olan faktör analizi , verileri özetleyen ve daha anlamlı bir halde yorumlamamıza olanak tanıyan bir analizdir.Değişkenlerin arasındaki ilişkileri araştırarak özet bilgiye ulaşma imkanı vermektedir.

Değişkenlerin ya da verilerin birbirleriyle olan bağılılıklarını ortadan kaldırmak için kullanılan en yaygın analizlerden biridir.Değişken bağımlılıklarını ortadan kaldırarak daha sağlıklı bir veri seti oluşturulmasına imkan vermektedir.¹⁰¹

Faktör analizi kullanım amaçlarına bakıldığında , iki temel amaç ön plana çıkmaktadır.Bunlar ;

- Değişken sayısını azaltmak ,
- Değişkenler arasındaki ilişkiden yararlanarak bazı özel yapılar ortaya çıkarmaktır.

¹⁰⁰ M. KANTARDZIC ,Data Mining Concepts, Models, Methods, and Algorithms , John Wiley & Sons Publishers 2003 s82.

¹⁰¹ H.TATLIDİL , Uygulamalı Çok Değişkenli İstatistiksel Analiz , Hacettepe ünv. 1996 s167.

Kullanıldığı yerlere bakıldığında ise esas olarak , tüketici eğilimleri, davranışları, tüketici karakteristiklerinin ortaya çıkarılması için kullanıldığı bilinmektedir.

Faktör analizinde faktörlerin belirlenmesi için birçok yöntem kullanılmaktadır. Bunlar kullanım sıklıklarına göre şu şekilde sıralanabilir;¹⁰²

- Temel Bileşenler Yöntemi ,
- En Çok Olabilirlik Yöntemi ,
- Ağırlıksız En Küçük Karalar Yöntemi ,
- Genelleştirilmiş En Küçük Kareler Yöntemi ,
- Ana Eksen Faktörizasyonu Yöntemi ,
- Alfa Faktörizasyon Yöntemi ,
- İmge Faktörizasyonu Yöntemidir.

Bu yöntemler içinde en yaygın kullanılan yöntemler ise , temel bileşenler analizi ve en çok olabilirlik yöntemidir.

Faktör analizi kullanılma koşullarına bakıldığında ise , gerek maliyet gerekse işlem zorlukları nedeniyle profesyonel uygulamalara ihtiyaç duymaktadır.

Yöntemin ana amacı, fazla sayıdaki değişkenlerin gruplanarak faktör değişkenler olarak ifade edilip edilemeyeceğinin belirlenmesi ve bu mümkün ise hangi değişkenlerin hangi faktör içinde yer alacağının bulunmasıdır.

Bu sayede araştırmacı faktörler içine dahil edilen değişkenleri inceleyerek ilgili faktörün ne anlam ifade ettiğini yorumlayabilecektir. Faktör analizinin algoritması kısaca aşağıdaki şekliyle ifade edilebilir.¹⁰³

$$Y = \alpha + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + ... + \beta_n x_n + \varepsilon \quad (1)$$

$$Y = \alpha + A_1 F_1 + A_2 F_2 + A_3 F_3 + ... + A_k F_k + u \quad (2)$$

$$F_k = W_{k1} x_1 + W_{k2} x_2 + W_{k3} x_3 + ... + W_{kn} x_n \quad (3)$$

¹⁰² N.H.TIMM ,Applied Multivariate Analysis, Springer Publishers 2002 s496.

¹⁰³ W.Hardle , Multivariate Statistics: Exercises And Solutions , Springer Publishers 2007 s185-188.

Denklem 1’de verilen çoklu doğrusal regresyon denkleminde $x_1x_2x_3...x_n$ ile ifade edilen değişken sayısının fazla olması, bağıntının karmaşıklığını arttırmakta ve kullanımını çoklu doğrusallık nedeniyle güçleştirmektedir. Faktör analizi sonucunda belirlenen k sayıda faktör, $F_1F_2F_3...F_k$ kullanılarak ifade edilen çoklu doğrusal regresyon ilişkisi denklem 2’de verilmektedir. Burada her faktör denklem 1’de verilen $x_1x_2x_3...x_n$ değişkenlerinin bir fonksiyonu olarak faktör katsayıları ($W_{k1}W_{k2}W_{k3}...W_{kn}$) yardımıyla denklem 3’de verildiği şekliyle ifade edilmektedir.

Faktör analizinin etkin olabilmesi için ‘k’ faktör sayısının mümkün olduğunca ‘n’ parametre sayısından küçük olması gerekmektedir ($k < n$). Aksi takdirde faktör analizi değişken sayısını azaltamayacağı için faktör analizi uygulanamaz. Faktör analizi iki aşamada uygulanmaktadır.¹⁰⁴

- Değişkenlerin faktör gruplarını oluşturmaya uygun olup olmadıklarının tayini,
- Faktörlerin belirlenmesi ve faktör skor katsayılarının ($W_{k1}W_{k2}W_{k3}...W_{kn}$) hesabıdır.

Değişkenler faktör grupları içine dahil edilemiyorsa, faktör analizinin kullanımı da mümkün değildir. Bu durum, ilk aşamada faktör analizi için uygunluk kriterleri ile araştırılmaktadır.

Faktör analizinin ikinci aşamasında değişkenlerin ait olduğu faktör gruplarına karar verilmektedir. Faktör sayısı Asal Bileşen Analizi ile belirlenmektedir. Bu yöntemde bağımsız değişkenlerin varyansları ayrı ayrı belirlendikten sonra, toplam varyansı büyük oranda (>%70) temsil eden değişken sayısı kadar faktör seçilmektedir.

Asal Bileşen Analizi faktör analizinden bağımsız bir teknik olup, ana kullanış amacı, regresyona dahil edilecek ve çoklu doğrusallığa yol açabilecek bağımsız değişkenlerin teşhis edilmesidir. Bu teknik , özellikle işlem verimliliği açısından diğer benzer tekniklere göre üstünlük arz ettiğinden bilgisayar uygulamalarında hesap süresini azaltmaktadır.

¹⁰⁴ L.L.HARLOW , The Essence Of Multivariate Thinking : Basic Themes And Methods, LEA Publishers 2005 s222-223.

Ayrıca k faktör sayısını belirlemek için varyans analizine dayalı şu yöntemler kullanılmaktadır.¹⁰⁵

- Temel Eksen faktörü ,
- En Çok Olabilirlik Tekniği ,
- Ağırlıksız En Küçük Kareler ,
- Genelleştirilmiş En Küçük Kareler ,
- image faktörü,
- alpha faktörüdür.

Her bağımsız değişkenin seçilen faktörler cinsinden aşağıdaki denklemde verilen doğrusal regresyon denklemleri kurulmaktadır.

$$x_i = \overline{\beta}_1 F_1 + \overline{\beta}_2 F_2 + \overline{\beta}_3 F_3 + ... + \overline{\beta}_k F_k \quad (4)$$

$$F_1 = x_1, F_2 = x_2, F_3 = x_3, ..., F_k = x_k$$

Seçilen faktörlere karşılık gelen bağımsız değişkenler ise ,

$$\overline{\beta}_1, \overline{\beta}_2, \overline{\beta}_3, ..., \overline{\beta}_k$$

faktör ağırlıklarını ifade etmektedir.

Denklem 4'te her faktörün $\overline{\beta}$ katsayısı, aynı zamanda faktör ile x_i değişkeni arasındaki korelasyon katsayısını vermektedir. Korelasyon katsayısı ya da faktör ağırlıklarının karelerinin toplamı x_i değişkeninin kullanılan faktörlerle temsil edilebilen toplam varyans yüzdesini ya da katkı değerini ifade etmektedir.

¹⁰⁵ H.TATLIDİL , Uygulamalı Çok Değişkenli İstatistiksel Analiz , Hacettepe üniv. 1996 s171.

Aşağıdaki formül katkı değeri hesaplamak için kullanılır.

$$x_{i_{katz}} = (\overline{\beta_1})^2 + (\overline{\beta_2})^2 + (\overline{\beta_3})^2 + ... + (\overline{\beta_k})^2 \quad (5)$$

Katkı değerleri , 0 ile 1 arasında bir değer almaktadır. Katkı değeri düşük olan değişkenlerin, belirlenen faktörlerle bir ilişkilerinin olmadığı kabul edilerek, faktör analizinde kullanılmamaları ve regresyon denklemine bağımsız değişkenler olarak dahil edilmeleri gerekmektedir. Her bir bağımsız değişkenin $x_1x_2x_3...x_n$ hangi faktör altında yer aldığı korelasyon katsayılarına ya da faktör ağırlıklarına bakılarak karar verilmektedir. İdeal olarak her değişkenin en fazla bir faktör ile yüksek bir korelasyon katsayısı vermesi istenir. Ancak korelasyon matrisine bakıldığında , kimi değişken birden fazla faktöre dahilmiş gibi algılanabilmektedir. Bu gibi durumlarda ortogonal döndürme tekniği kullanılarak döndürülmüş korelasyon matrisi elde edilir. Döndürülmüş korelasyon matrisi sayesinde , her değişkenin kesin olarak hangi faktör altında yer aldığı kolayca algılanabilmektedir. Döndürme teknikleri ortogonal döndürme ve eğimli açı ile döndürme olmak üzere iki farklı teknik ile ele alınabilmektedir. Bu çalışmada ortogonal döndürme tekniği , birbirinden tamamen bağımsız faktör grupları oluşturması nedeniyle tercih edilmiş, ileride lineer regresyon analizine giren bu faktörler arasında çoklu doğrusallık olması engellenmiştir. Bu durumda, eğimli açı ile döndürme tekniğinin parametreler arasında kısmi bir ilişki olduğunun varsayılabilirdiği durumlarda kullanılması daha uygundur. ¹⁰⁶

Faktör analizinde son aşama, denklem 3'te verilen $W_{k1}W_{k2}W_{k3}...W_{kn}$ faktör skor katsayılarının hesaplanması ve böylece faktör değerlerinin belirli hale getirilmesi gerekmektedir. Faktör analizinden elde edilen faktörler ve bu faktörlerle ilişkisi bulunmayan bağımsız değişkenler tahminleme modelinin kurulmasında kullanılmaktadır. Kurulan tahminleme modeli ,

$$Y_{tahmin} = \alpha + \beta_1 F_1 + \beta_2 F_2 + ... + \beta_{k+1} F_{k+1} + ... + \beta_{k+n} F_{k+n} + U \quad \text{dir.}$$

¹⁰⁶ W.Hardle , Multivariate Statistics: Exercises And Solutions ,Springer Publishers 2007 s186-188.

Burada $F_1, F_2, F_3, \dots, F_k$ faktör analizinden hesaplanan faktör değerlerini, $x_{k+1}, x_{k+2}, x_{k+3}, \dots, x_{k+n}$ ise faktör analizinde ilişkisi bulunamamış bağımsız değişkenleri ifade etmektedir. Faktörlerin etki değerleri, $\beta_1, \beta_2, \beta_3, \dots, \beta_{k1}$ katsayıları ile, bağımsız parametrelerin etki değerleri ise $\beta_{k+1}, \beta_{k+2}, \beta_{k+3}, \dots, \beta_{k+n}$ katsayıları ile temsil edilmektedir.

Faktörler içinde yer alan değişkenlerin etkinlik düzeyleri aşağıdaki adımlar dahilinde belirlenebilmektedir.

İlk olarak , faktör skor matrisi kullanılarak her bir değişkenin faktör skor katsayıları toplanmaktadır. Aşağıdaki şekil skorlamanın nasıl yapıldığı göstermektedir.

Tablo 2.3. Skorlama katsayısı matrisi hesaplanması

	F_1	F_2	F_3	...	F_k
x_1	W_{11}	W_{12}	W_{13}	...	W_{1k}
x_2	W_{21}	W_{22}	W_{23}	...	W_{2k}
x_3	W_{31}	W_{32}	W_{33}	...	W_{3k}
...	...	...	...	...	...
W_n	W_{n1}	W_{n2}	W_{n3}	...	W_{nk}
	$\sum W_{n1}$	$\sum W_{n2}$	$\sum W_{n3}$		$\sum W_{nk}$

Kaynak : H.TATLIDİL , Uygulamalı Çok Değişkenli İstatistiksel Analiz , Hacettepe ünv. 1996 s199

İkinci olarak , incelenen değişkenin her faktöre katkısı, W_{ij}/W_{nj} oranından hesaplanmaktadır.

Tablo 2.4. Değişkenlerin faktöre katkı oranları hesaplanması

	F_1	F_2	F_3	...	F_k
x_1	$\frac{W_{11}}{\sum W_{n1}}$	$\frac{W_{12}}{\sum W_{n2}}$	$\frac{W_{13}}{\sum W_{n3}}$	...	$\frac{W_{1k}}{\sum W_{nk}}$
x_2	$\frac{W_{21}}{\sum W_{n1}}$	$\frac{W_{22}}{\sum W_{n2}}$	$\frac{W_{23}}{\sum W_{n3}}$	...	$\frac{W_{2k}}{\sum W_{nk}}$
x_3	$\frac{W_{31}}{\sum W_{n1}}$	$\frac{W_{32}}{\sum W_{n2}}$	$\frac{W_{33}}{\sum W_{n3}}$	...	$\frac{W_{3k}}{\sum W_{nk}}$
...	...	...	...	...	...
x_n	$\frac{W_{n1}}{\sum W_{n1}}$	$\frac{W_{n2}}{\sum W_{n2}}$	$\frac{W_{n3}}{\sum W_{n3}}$	...	$\frac{W_{nk}}{\sum W_{nk}}$

Kaynak : H.TATLIDİL , Uygulamalı Çok Değişkenli İstatistiksel Analiz , Hacettepe ünv. 1996 s199

Son aşama olarak denklem 6'dan elde edilen her faktör için beta katsayıları, ' $\beta_1, \beta_2, \beta_3, \dots, \beta_k$ ', ile W_{ij}/W_{nj} oranlarının çarpımlarının toplamı değişkenin etki puanı olarak hesaplanmaktadır.

Tablo 2.5. Değişken etki puanlarının hesaplanması

	F_1	F_2	F_3	...	$\frac{\sum (\beta_k W_{ik})}{\sum W_{nk}}$
x_1	$\frac{\beta_1 W_{11}}{\sum W_{n1}}$	$\frac{\beta_2 W_{12}}{\sum W_{n2}}$	$\frac{\beta_3 W_{13}}{\sum W_{n3}}$	...	$\chi_{1_{etki}}$
x_2	$\frac{\beta_1 W_{21}}{\sum W_{n1}}$	$\frac{\beta_2 W_{22}}{\sum W_{n2}}$	$\frac{\beta_3 W_{23}}{\sum W_{n3}}$	...	$\chi_{2_{etki}}$
x_3	$\frac{\beta_1 W_{31}}{\sum W_{n1}}$	$\frac{\beta_2 W_{32}}{\sum W_{n2}}$	$\frac{\beta_3 W_{33}}{\sum W_{n3}}$	...	$\chi_{3_{etki}}$
...	...	...	...	...	...
x_n	$\frac{\beta_1 W_{n1}}{\sum W_{n1}}$	$\frac{\beta_2 W_{n2}}{\sum W_{n2}}$	$\frac{\beta_3 W_{n3}}{\sum W_{n3}}$	...	$\chi_{n_{etki}}$

Kaynak : H.TATLIDİL , Uygulamalı Çok Değişkenli İstatistiksel Analiz , Hacettepe ünv. 1996 s199

Değişken etki puanları $\chi_{1_{etki}}, \chi_{2_{etki}}, \chi_{3_{etki}}, \dots, \chi_{n_{etki}}$ 'çok etkili', 'etkili', 'az etkili' ve etkisiz' olmak üzere dört ayrı kategoride değerlendirilmektedir. Bu sınıflamada 'etkili'-'az etkili' sınırı t-dağılım testinden belirlenmektedir. 'çok etkili'-'etkili' ve 'az etkili'-'etkisiz' sınırları ise t testinden elde edilen sınır değerinin altında ve üzerinde kalan değerlerin ortanca değerleri hesaplanarak elde edilmektedir.

Kullanılan parametrelerin tanımları aşağıda belirtilmiştir.

A_k	:Faktör beta katsayısı
F_k	:Faktör değeri
k	:Faktör sayısı
n	:Veri sayısı
u	:Faktör analizi hata terimi
W_{kn}	:Faktör skor katsayısı
x	:Bağımsız değişken
$x_{i\text{Katki}}$	:Parametre katkı değeri
$\lambda_{n\text{etki}}$	:Parametre etki değeri
Y	:Bağımlı değişken
Y_{kira}	:Kira değeri
β	:Regresyon katsayısı
β_k	:Faktör yük katsayısı
ε	:Hata terimi

Faktör analizi, birbirleriyle ilişkili veri yapılarını birbirinden bağımsız daha az sayıda yeni veri yapılarına dönüştürmek, bir oluşumun nedenini açıkladıkları varsayılan değişkenleri gruplamak ve ortak faktörleri ortaya koymak, majör ve minör faktörleri tanımlamak amacıyla başvurulmuş bir çok değişkenli istatistiksel analiz türüdür .

Yani faktör analizi, veriler arasındaki ilişkilere dayanarak verilerin daha anlamlı ve özel bir biçimde sunulmasını sağlayan bir yöntemdir. Faktör analizinin temel amacı boyut indirgeme ve bağımlılık yapısını yok etmektir. Faktör analizinin diğer amaçlarını şu şekilde sıralamak mümkündür. Değişkenler arasındaki karşılıklı bağımlılığın kökenini araştırmaktır. Faktör, gözlenen değişkenlerin doğrusal bir bileşimidir.¹⁰⁷

Faktör analizinde , kovaryans veya korelasyon matrisinden hareket ederek bilgi kaybı olmadan daha az sayıda faktör adını verdiğimiz yeni değişkenlere ulaşılmaya çalışılır.Karşılıklı olarak aralarında ilişki olan değişkenleri bir araya toplamak ve böylece veri grubunu daha az sayıda değişken ile temsil edilebilecek duruma dönüştürebilmektir .

¹⁰⁷ <http://www.statistics.com/resources/glossary/f/factoran.php> 2007.

Faktör analizi, gruplandığı değişkenler arasındaki ortaklaşa ilişkileri inceleyerek birbirleriyle yüksek ilişki kuran değişkenleri bir grupta; daha az ilişki içinde bulunan değişkenleri diğer bir grupta toplayabilir.

Faktör analizinin başlıca varsayımları, veri matrisinin analiz öncesi kriter ve tahmin değişkenleri alt matrislerine bölüştürülmemesi ve değişkenler arasındaki ilginin doğrusal olduğudur .

Bu açıklamaların ışığında faktör analizinin, değişkenler arasındaki tüm ilişkilerin gücünün ve bu arada bu ilişkiyi temsil edecek değişkenlerin saptanmasını amaçlayan, esas olarak değişkenlerle ilgilenen, veri matrisinin kriter ve tahmin değişkenlerinin alt matrislerine bölüştürülmediği ; değişkenler arasındaki ilişkinin doğrusal olduğunun varsayıldığı ve genel kural olarak aralıklı ölçekle ölçülmüş verilere gereksinme gösteren bir çok değişkenli istatistiksel analiz olduğu söylenebilir.

Örneklem büyüklüğü faktör analizi için önemlidir. Gözlem sayısı değişken sayısından fazla olmalıdır. Başarılı bir faktöranalizi uygulamasında, elde edilen faktör sayısı değişken sayısına göre çok daha az olmalıdır. Ayrıca faktörlerin yorumlanabilir olması aranılan diğer bir özelliktir.¹⁰⁸

Faktör analizinde faktörlerin ortaya çıkarılması için yapılan faktörleşme işleminde farklı yöntemler kullanılmaktadır. Bu yöntemlerden bazıları; temel bileşenler, en büyük benzerlik, ağırlıksız en küçük kareler, genelleştirilmiş en küçük kareler, ana eksen faktörizasyonu, alfa faktörizasyon, imge faktörizasyonu, çoklu gruplandırma ve maksimum olabilirlik yöntemleridir.

Temel Bileşenler Yöntemi, bütün değişkenlerdeki maksimum varyansı açıklayacak faktörü hesaplar. Kalan maksimum miktardaki varyansı açıklamak için, ikinci faktör hesaplanır. Ancak, birinci faktörün ikinci faktör ile ilişki göstermemesi için sınırlama vardır. Söz konusu süreç, değişkenlerdeki bütün varyansın açıklanmasına kadar devam eder. Normal olarak bu noktaya faktör sayısı değişken sayısına eşit olunca ulaşılır. Ancak değişken sayısı kadar faktör olması hiçbir şeyi basitleştirmeyeceği

¹⁰⁸ K.KURTULUŞ ,Pazarlama Alıştırmaları ,Avcıol Yayını ,1998 s482.

için; özdeğer istatistiği kullanılarak analize kaç faktörün dahil edileceğine karar verilir.

2.5.2. Kanonik Korelasyon Analizi

Kanonik korelasyon analizi , birden çok bağımlı ve bağımsız değişken arasındaki ilişkiyi ortaya koymak için kullanılan bir analizdir.Çok değişkenli istatistiksel analiz tekniği olan kanonik korelasyon analizi , şu amaçlar için kullanılmaktadır.¹⁰⁹

- İki değişkenin birbirinden bağımsız olup olmadığının tespit edilmesi,
- Kümeler arasında en büyük ilişkilere sahip kümelerin ortaya çıkarılması ,
- Korelasyonu maksimum yapacak modelin belirtilmesi için kullanılmaktadır.

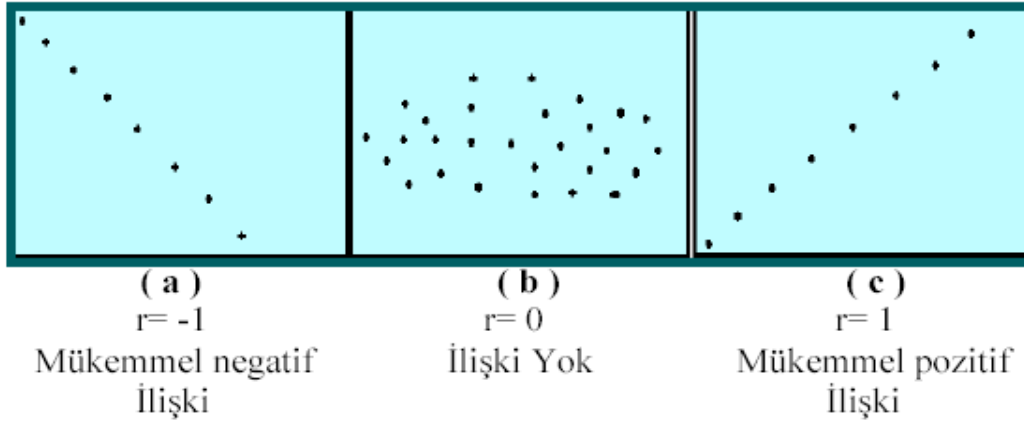
Kanonik korelasyon analizine örnek vermek gerekirse , bir firmanın ürününü alan müşterinin cinsiyeti , medeni hali, yaşı ve de satın aldıkları ürünler arasındaki ilişkiyi ortaya koyan bir analizdir.

İki değişken arasındaki ilişki hakkında genel bir bilgi edinmek için kullanılan grafiklere saçılım grafiği denir.Ancak, ilişkinin miktarı konusunda yorum yapabilmek için korelasyon katsayısının hesaplanması gerekmektedir.¹¹⁰

Korelasyon katsayısı (r), iki değişken arasındaki ilişkinin ölçüsüdür ve -1 ve +1 arasında değişim gösterir.Aşağıdaki şekilde değişkenler arasındaki korelasyon grafiklerle gösterilmektedir.

¹⁰⁹ N.H.TIMM ,Applied Multivariate Analysis, Springer Publishers 2002 s477.

¹¹⁰ H.TATLIDİL , Uygulamalı Çok Değişkenli İstatistiksel Analiz , Hacettepe ün.v. 1996 s216.



Şekil 2.20.Konanik kolerasyon grafiği

Kaynak : H.TATLIDİL , Uygulamalı Çok Değişkenli İstatistiksel Analiz , Hacettepe ünv. 1996 s216

$r = -1$ olan grafikte değişkenlerden birisinin artışına bağlı olarak diğesinde azalması yönünde bir ilişki olduğunu göstermektedir.

$r = 0$ olan grafikte iki değişken arasında ilişki olmadığını göstermektedir.

$r = 1$ değişkenlerden birisindeki artışa bağlı olarak diğesinde de artış olacağını göstermektedir.

Korelasyonun katsayısının gücü ile ilgili olarak aşağıdaki tanımlamalar yapılmıştır.

- 0 - 0.25 Çok zayıf ilişki ,
- 0.26 - 0.49 Zayıf ilişki ,
- 0.50 - 0.69 Orta ilişki ,
- 0.70 - 0.89 Yüksek ilişki ,
- 0.90 - 1.0 Çok yüksek ilişkidir.

Korelasyon katsayısı, örneklem büyüklüğünden etkilenmektedir. Küçük hacimli örneklerde, elde edilen korelasyon katsayısı büyük bile olsa istatistiksel olarak önemli bir değer olmayabilir.Dolayısıyla, elde edilen değerın hipotez testinin yapılması gerekmektedir.¹¹¹

¹¹¹ K.KURTULUŞ ,Pazarlama Alıştırmaları ,Avcıol Yayını ,1998 s453.

Değişkenlerin türlerine göre korelasyon katsayıları sınıflanabilir.İki nitel değişken arasındaki ilişkinin belirlenmesi için kullanılan korelasyon katsayıları ; Phi katsayısı ,Cramer V katsayısı ,Olaganlık katsayısı ,Lambda katsayısıdır.

Kesikli/sürekli nicel değişkenler arasındaki ilişkinin belirlenmesinde kullanılan korelasyon katsayıları ise Pearson korelasyon katsayısı , Spearman korelasyon katsayısıdır.

Kısmi korelasyon katsayıları , iki değişken arasındaki ilişkiyi gösterirken diğer değişkenlerin etkilerini dikkate alınmadan ya da diğer değişkenlerin etkisi arındırıldıktan sonra iki değişkenin biribiri ile olan ilişkilerini ortaya koyan bir korelasyon katsayısıdır.

2.5.3.Lojistik Regresyon Analizi

Lojistik regresyon analizi , geleceğe dönük tahmin yapmak için kullanılan ve son yıllarda popüler olan bir analizdir.Bu analiz değişkenlerin kesikli değerler almasına ve de esnek modeller kurulmasına olanak sağladığı için günümüzde tercih edilmektedir.Kesikli değişkenlere örnek olarak medeni hal , cinsiyet vb. örnek olarak verilerbilir.¹¹²

Lojistik regresyon modellerinin yaygın bir şekilde kullanılır hale gelmesi, katsayı tahmin yöntemlerinin geliştirilmesi ve lojistik regresyon modellerinin daha ayrıntılı incelenmesine sebep olmuştur. Cornfield (1962), lojistik regresyondaki katsayı tahmin işlemlerinde diskriminant fonksiyonu yaklaşımını ilk kez kullanarak popüler hale getirmiştir. Lee (1984) basit dönüşümlü deneme planları için linear lojistik modeller üzerinde durmuştur. Bonney (1987) lojistik regresyon modelinin kullanımı ve geliştirilmesi üzerinde çalışmıştır. Robert ve ark. (1987) lojistik regresyonda standart ki-kare, olabilirlik oran, en çok olabilirlik tahminleri, uyum mükemmelliği ve hipotez testleri üzerine çalışmalar yapmışlardır.

Duffy (1990) lojistik regresyonda hata terimlerinin dağılışı ve parametre değerlerinin gerçek değerlere yaklaşımını incelemiştir. Hsu ve Leonard (1995) lojistik regresyon fonksiyonlarında Bayes tahminlerinin elde edilmesi işlemleri üzerine çalışmışlar ve lojistik regresyonda Monte Carlo dönüşümünün kullanılabileceğini göstermişlerdir. Akkaya ve Pazarlıoğlu (1998) lojistik regresyon modellerinin ekonomi alanında kullanımını örneklerle incelemişlerdir.

Çeşitli varsayım bozulmaları olduğunda Lojistik regresyon analizi, diskriminant analizi ve çapraz tablo uygulamalarına alternatif olarak uygulanmaktadır. Kullanım nedeni olarak en temel yaklaşım doğrusal regresyon analizinde yapılabilir; bağımlı değişken 0 ve 1 gibi ikili (binary) ya da ikiden çok kategori içeren kesikli değişken

¹¹² L.L.HARLOW , The Essence Of Multivariate Thinking : Basic Themes And Methods, LEA Publishers 2005 s152.

olduğunda normallik varsayımı bozulmakta ve doğrusal regresyon analizi uygulanamamaktadır.¹¹³

Lojistik regresyonu doğrusal regresyondan ayıran en belirgin özellik ise lojistik regresyonda sonuç değişkeninin ikili veya çoklu olmasıdır. Lojistik regresyon ve doğrusal regresyon arasındaki bu fark hem parametrik model seçimine, hem de varsayımlara yansımaktadır.

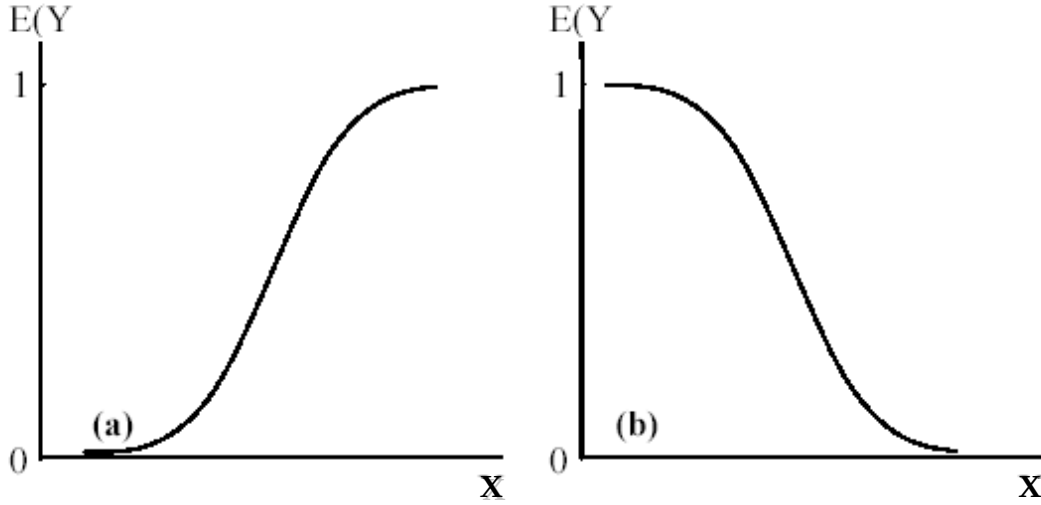
Lojistik regresyonda da, doğrusal regresyon analizinde olduğu gibi bazı değişken değerlerine dayanarak tahmin yapılmaya çalışılır. Ancak bu iki yöntem arasında önemli fark bulunmaktadır.

- Doğrusal regresyon analizinde tahmin edilecek olan bağımlı değişken sürekli iken, lojistik regresyon analizinde bağımlı değişken kesikli bir değer almaktadır.
- Doğrusal regresyon analizinde bağımlı değişkenin değeri, lojistik regresyon analizinde ise bağımlı değişkenin alabileceği değerlerden birinin gerçekleşme olasılığı tahmin edilir.
- Doğrusal regresyon analizinde bağımsız değişkenin çoklu normal dağılım göstermesi şartı aranırken, lojistik regresyon analizinde böyle bir şart yoktur.

Lojistik regresyon modelleri zayıf ölçekle ölçülmüş değişkenler arasındaki ilişkinin şeklini ortaya koyan modellerdir. Yapılan bir çok çalışmada bağımlı değişken sadece iki sonuca sahiptir. Genellikle üzerinde durulan olayın gerçekleşmesi 1 gerçekleşmemesi ise 0 ile gösterilir. Hem teorik hem de deneysel incelemeler bağımlı değişken iki sonuçlu iken cevap fonksiyonunun şeklinin S veya ters S şeklinde olacağını bilinmektedir. Bağımlı değişken, aşağıdaki şekilde de görüldüğü gibi bitiş noktaları dışında yaklaşık olarak doğrusaldır.

¹¹³ H.TATLIDİL , Uygulamalı Çok Değişkenli İstatistiksel Analiz , Hacettepe ünv. 1996 s289.

Bu cevap fonksiyonları 0 ile 1 değerlerinde X ve Y eksenlerine asimptottur.



Şekil 2.21. S ve ters S fonksiyonu

Kaynak : M.AYTAÇ , Matematiksel İstatistik ,Ezgi Yayınevi 1999 s309

Yukarıdaki şekilde gösterilen cevap fonksiyonları, lojistik cevap fonksiyonları olarak bilinmektedir. Lojistik fonksiyonun 0 ile 1 arasında bir değişim aralığına sahip olması lojistik fonksiyonun tercih edilmesindeki ilk önemli nedenidir. Lojistik model, ortaya çıkacak riski 0 ile 1 arasında herhangi bir değer olarak tahmin etmeye yarar. Başka bir deyişle 1'in üstünde veya 0'ın altında bir risk olmaz. Bu durum her model için her zaman doğru olmamaktadır.¹¹⁴

Araştırmacı bazen bağımsız değişkenler üzerinde denetime sahiptir. Böyle bir imkan söz konusu olduğunda Xi değerlerine karşılık gelen hücrelerdeki birim sayısının asgari 30 olması regresyonun verilere uygunluğunu önemli ölçüde yükseltir.

Lojistik regresyon fonksiyonu,

¹¹⁴ K.ÖZDAMAR , Paket Programlar ile İstatiksel Veri Analizi , Kaan Kitabevi 1999 s475.

$$\pi(x) = \frac{\exp(\beta_0 + \beta_1 X)}{1 + \exp(\beta_0 + \beta_1 X)}$$

şeklindedir. Bu ifadenin bir diğer şekli ise,

$$\pi(x) = [1 + \exp(-\beta_0 - \beta_1 X)]^{-1}$$

olarak yazılabilir. $\pi(x) = E(Y/x)$ değeri şartlı ortalama olarak bilinir. Şartlı ortalamanın, modelde yer alan parametrelerle $(\beta_0 + \beta_1)$ doğrusal hale dönüştürülmesi için, transformasyona olduğu gibi tutulması gerekir. Bu transformasyona Logit transformasyon adı verilir. Bu formül aşağıda gösterilmektedir.

$$g(x) = \ln \left[\frac{\pi(x)}{1 - \pi(x)} \right] = \beta_0 + \beta_1 x$$

Transformasyon değişkeni $g(x)$, modeldeki parametrelerle doğrusaldır, süreklidir ve $-\infty, +\infty$ aralığında değişen değerler alır. $\pi(x)$ arttıkça $g(x)$ 'te artar ve eğer $\pi(x) < 0.5$ ise $g(x)$ negatif, $\pi(x) > 0.5$ ise $g(x)$ pozitif değerler alır.

Modelin sonuç değişkeninin sınırlarını genişletmek için uygulanan Logit transformasyonunun bazı özellikleri şöyle sıralanabilir.¹¹⁵

- p arttıkça $\text{logit}(p)$ de artmaktadır.
- p , 0 ile 1 arasında iken $\text{logit}(p)$ reel sayılar doğrusu üzerinde değerler almaktadır.
- $p < 0.5$ olduğunda $\text{logit}(p) < 0$ ve $p > 0.5$ olduğunda $\text{logit}(p) > 0$ olur.

Doğrusal regresyon modelinde bağımlı değişkene ait bir gözlem $y = E(Y/x) + \varepsilon$

¹¹⁵ H.TATLIDİL, Uygulamalı Çok Değişkenli İstatistiksel Analiz, Hacettepe üniv. 1996 s293.

şeklinde gösterilebilir. ε hata terimi olarak isimlendirilir ve gözlemin koşullu olasılıktan ne kadar saptığını gösterir. ε 'nin ortalamasının sıfır ve varyansının ise bağımsız değişkenin her düzeyinde sabit olacak şekilde normal dağılım göstereceği genel bir varsayımdır. Bu varsayım bağımlı değişken iki düzey içerdiği zaman geçerli değildir.

Bu tür durumlarda x verildiğinde sonuç değişkeninin değeri $y = \pi(x) + \varepsilon$ ile gösterilir. Ve ε 'nin mümkün olan iki değerden başka değer alamayacağı varsayılır. Eğer $y = 1$ ise, $\pi(x)$ olasılıkla $\varepsilon = 1 - \pi(x)$ değerini alır ve eğer $y = 0$ ise, $1 - \pi(x)$ olasılıkla $\varepsilon = -\pi(x)$ değerini almaktadır. Böylece ε , sıfır ortalamalı ve $\pi(x)[1 - \pi(x)]$ 'a eşit varyanslı binomiyal bir dağılım göstermiş olur.

Lojistik modelde parametrelerin tahminine bakıldığında , parametre tahmin etmek için çeşitli yöntemler ortaya atılmıştır.Bu çalışmada parametrelerin tahmin edilmesinde en çok olabilirlik (maximum likelihood) tahmin yöntemi kullanılacaktır.

Genel olarak en çok olabilirlik yöntemi, gözlenen veri kümesini elde etmenin olasılığını maksimum yapan bilinmeyen parametrelerin değerlerini verir. Bu metodu uygulamak için öncelikle, en çok olabilirlik fonksiyonunun oluşturulması gerekmektedir.Bu fonksiyon gözlenen verilerin olasılıklarını, bilinmeyen parametrelerin bir fonksiyonu olarak açıklar. Bu parametrelerin en çok olabilirlik tahmin edicileri, fonksiyonu maksimum yapan değerleri bulacak şekilde seçilir.

Böylece sonuçta elde edilen tahminleyiciler, gözlenen verilerle çok yakın değerlere sahiptir. Eğer y , 0 ve 1 olarak kodlandıysa, bu durumda 1 numaralı eşitlikte verilen $\pi(x)$ ifadesi, verilen x değeri için y 'nin 1'e eşit olma koşullu olasılığını vermektedir. Bu olasılık $\pi(x) = P(y = 1/x)$ sembolüyle gösterilir. Buradan hareketle, $[1-\pi(x)]$ ifadesi de, y 'nin 0 degerini alma koşullu olasılığını göstermektedir.¹¹⁶

Bu olasılık da $[1-\pi(x)] = P(y = 0/x)$ şeklinde gösterilir. (x_i, y_i) çifti için $y_i = 1$

¹¹⁶ D.LAROSE , Data Mining Methods and Models , Wiley Publisherss 2006 s155 .

olduğunda olabilirlik (likelihood) fonksiyonuna katkısı $\pi(x_i)$ iken $y_i = 0$ olduğunda olabilirlik fonksiyonuna katkısı $1 - \pi(x_i)$ kadar olmaktadır. (x_i, y_i) çiftinin olabilirlik fonksiyonuna katkısını hesaplama formülü aşağıda verilmiştir.

$$\zeta(x_i) = \pi(x_i)^{y_i} [1 - \pi(x_i)]^{1-y_i}$$

Gözlemlerin birbirinden bağımsız oldukları varsayıldığı için, olabilirlik fonksiyonu yukarıda bulunan formül terimlerin çarpılmasıyla elde edilir.

$$l(\beta) = \prod \zeta(x_i)$$

En çok olabilirliğin temel ilkesinde β kestiriminin yukarıdaki ifadeyi maksimum yaptığı vurgulanmaktadır. Matematiksel olarak bu formülün logaritmasıyla çalışmak daha kolay olacağından log-olabilirlik fonksiyonu aşağıdaki gibi elde edilir.

$$L(\beta) = \ln[l(\beta)] = \sum_{i=1}^n \{y_i \ln[\pi(x_i)] + (1 - y_i) \ln[1 - \pi(x_i)]\}$$

$L(\beta)$ 'yi maksimum yapan β değerlerini bulmak için, $L(\beta)$ 'nin β_0 ve β_1 'e göre türevi alınarak sıfıra eşitlenir. Elde edilecek eşitlikler aşağıda belirtilmiştir.

$$\sum_{i=1}^n [y_i - \pi(x_i)] = 0$$

$$\sum_{i=1}^n x_i [y_i - \pi(x_i)] = 0$$

şeklindeir. Bu eşitlikler olabilirlik eşitlikleri (likelihood equations) olarak adlandırılır.

Lineer regresyon analizinde β 'ya göre türevinden elde edilen olabilirlik eşitlikleri, bilinmeyen parametreleri içeren doğrusal ifadelerdir, bu nedenle kolayca çözümlenebilmektedir.

Lojistik regresyon için yukarıda elde edilen eşitlikler β_0 ve β_1 'de lineer değildirler. Bundan dolayı bu eşitliklerin çözümlenmesi için özel yöntemlere ihtiyaç vardır. Bu denklemlerin çözümleri genelleştirilmiş ağırlıklı en küçük kareler yöntemi ile elde edildiği gösterilmiştir.

Lojistik regresyonda gözlenen ve beklenen değerlerin karşılaştırılması log olabilirlik fonksiyonu ile yapılmaktadır. Bu test Önem testi olarak adlandırılır.¹¹⁷

$$D = -2 \ln \left[\frac{\text{Şu andaki modelin olabilirliği}}{\text{Doymuş modelin olabilirliği}} \right]$$

Yukarıdaki formülde parantez içerisinde verilen ifade olabilirlik oranı “likelihood ratio” olarak adlandırılır. $(-2\ln)$ katının alınması, matematiksel olduğu kadar dağılımı bilinen bir değer elde etmektir. Bu değer hipotez testi amacıyla kullanılmaktadır. Böyle bir teste olabilirlik oran testi adı verilmektedir. Formüller yerine konulduğunda eşitlik aşağıdaki şeklini almaktadır.

$$D = -2 \sum_{i=1}^n \left\{ y_i \ln \left(\frac{\hat{\pi}_i}{y_i} \right) + (1 - y_i) \ln \left(\frac{1 - \hat{\pi}_i}{1 - y_i} \right) \right\}$$

Bağımsız bir değişkenin önemine karar vermek için, denklemde bağımsız değişkenin olduğu ve olmadığı durumlardaki D değerleri karşılaştırılır. Bağımsız değişkeni kapsamamasından dolayı ortaya çıkan D'deki değişim aşağıdaki gibidir.

$$G = D(\text{Değişkensiz Model için}) - D(\text{Değişkenli Model için})$$

¹¹⁷ H.TATLIDİL, Uygulamalı Çok Değişkenli İstatistiksel Analiz, Hacettepe ün. 1996 s299.

Hesaplanan bu istatistikde, doğrusal regresyonda kullanılan F testindeki pay kısmı ile aynı rolü üstlenir. G'yi hesaplamak için farkı alınacak D değerlerinin her ikisi için de doymuş modelin olabilirlikleri ortak olduğundan G istatistiği aşağıdaki şekli almaktadır.¹¹⁸

$$G = -2 \ln \left[\frac{\text{Değişkensiz modelin olabilirliği}}{\text{Değişkenli modelin olabilirliği}} \right]$$

Tek bağımsız değişkenli özel durumlarda, değişkenin modelde olmadığı zamanda ki β_0 'ın en çok olabilirlik tahmini $\ln(n_1/n_0)$ 'dır. ($n_1 = \sum y_i$ ve $n_0 = \sum (1-y_i)$). Tahmin

değeri n_1/n_0 sabittir. G istatistiği şu şekilde hesaplanır.

$$G = -2 \ln \left[\frac{\left(\frac{n_1}{n} \right)^{n_1} \left(\frac{n_0}{n} \right)^{n_0}}{\prod_{i=1}^n \hat{\pi}_i^{y_i} (1 - \hat{\pi}_i)^{(1-y_i)}} \right]$$

ya da

$$G = 2 \left\{ \sum_{i=1}^n [y_i \ln(\hat{\pi}_i) + (1-y_i) \ln(1-\hat{\pi}_i)] - [n_1 \ln(n_1) + n_0 \ln(n_0) - n \ln(n)] \right\}$$

dır. $\beta_1 = 0$ hipotezi altında, G istatistiği 1 serbestlik derecesiyle χ^2 dağılımı göstermektedir.

Tüm değişkenleri içeren model ile kestirilen modele ilişkin olabilirlik oran değerlerinin farkına dayanan ölçütlerin ki-kare dağılacığı düşüncesinden hareketle kurulan modelin geçerliliği sınanmaktadır.

¹¹⁸ D.LAROSE , Data Mining Methods and Models , Wiley Publisherss 2006 s160.

Çoklu lojistik regresyon analizine bakıldığında , bağımsız değişkenler değişik ölçüm biçimlerinde olabilmektedir. Kesikli ve nominal ölçekli bağımsız değişkenleri modele dahil etmek için dizayn değişkenleri kullanılması gerekir. Öncelikle modeldeki tüm bağımsız değişkenlerin her birinin en az aralık ölçekli olduğu varsayılmaktadır.

$X' = (x_1, x_2, x_3, \dots, x_p)$ vektörü ile gösterilsin. Sonuç değişkeninin mevcut olduğu ($Y=1$) zaman ki koşullu olasılık, $P(Y = 1/x) = \pi(x)$ 'e eşit olacaktır. Çoklu lojistik regresyon modelinin logiti aşağıdaki denklem ile gösterilir.¹¹⁹

$$g(x) = \beta_0 + \beta_1 X_1 + \dots + \beta_p X_p$$

Buradanda ,

$$\pi(x) = \frac{e^{g(x)}}{1 + e^{g(x)}}$$

Formülü elde edilir.

Bağımsız değişkenler kesikli, nominal ölçekli ise, o zaman bu değişkenler yerine dizayn (kukla) değişkenlerinin bu değişkenleri temsil etmesi için kullanılması gerekir. Genel olarak nominal değişken k kategoriye sahipse, o zaman k-1 dizayn değişkenine ihtiyaç vardır. J. Bağımsız değişken (x_j), k_j kategoriye sahip olsun. $K_j - 1$ dizayn değişkeni D_{ju} olarak ve katsayıları da β_{ju} , $u = 1, 2, \dots, k_j - 1$ olarak belirtilirse, j. değişken kesikli olan p değişkenli model için logit aşağıdaki gibi olur.¹²⁰

$$g(x) = \beta_0 + \beta_1 X_1 + \dots + \sum_{u=1}^{k_j-1} \beta_{ju} D_{ju} + \beta_p X_p$$

Birbirinden bağımsız n tane (x_i, y_i), $i=1, 2, \dots, n$ gözlem çiftinin olduğu olduğunu düşündüğümüzde tek değişkenli modelde olduğu gibi modelin kurulması için tahmin vektörünün $\beta' = (\beta_1, \beta_2, \beta_3, \dots, \beta_p)$ elde edilmesi gerekir. Çok değişkenli durumda,

¹¹⁹ M. KANTARDZIC ,Data Mining Concepts, Models, Methods, and Algorithms , John Wiley & Sons Publishers 2003 s106.

¹²⁰ H.TATLIDİL , Uygulamalı Çok Değişkenli İstatistiksel Analiz , Hacettepe ün. 1996 s304.

tek deęişkenli durumda olduęu gibi tahmin metodu en çok olabilirlik metodu olacaktır.

$\hat{\beta}$ en çok olabilirlik tahmin teorisi, log olabilirlik fonksiyonunun ikinci dereceden türevlerinden oluşan matrizen tahmin deęerlerinin elde edilmektedir. Logaritmik olabilirlik fonksiyonunun $\beta_0, \beta_1, \dots, \beta_{p-1}$ parametrelerine göre ikinci dereceden kısmi türevlerinin matrisini G ile gösterilir. G matrisi,

$$G = (g_{ij})_{p \times p} \quad i = 0, 1, 2, \dots, p-1; j = 0, 1, 2, \dots, p-1$$

şeklinde gösterilir.

$$g_{00} = \frac{\partial^2 \log_e L(\beta)}{\partial \beta_0^2}, \quad g_{01} = \frac{\partial^2 \log_e L(\beta)}{\partial \beta_0 \partial \beta_1}, \dots,$$

deęerleri elde edilir. Bu matris, Hessiyan matrisi olarak adlandırılır. Hessiyan matrisdeki ikinci derece kısmi türevleri, $b = i$ olarak; yani, en çok olabilirlik tahminleri olarak görmek gerekir. En yüksek ihtimal tahmini için kullanıldığında Lojistik regresyondan tahmin edilen regresyon katsayılarının tahmini yaklaşık varyans kovaryans matrisi şu eşitlikten elde edilir.

$$s^2\{b\} = [(-g_{ij})_{\beta=b}]^{-1}$$

Örnek hacmi yeterince büyük olduğunda , basit veya çoklu lojistik regresyon modellerindeki regresyon katsayılarının anlamlı olup olmadığını test ederken aşağıdaki deęere baęlı olarak karar verilir.

$$\frac{b_k - \beta_k}{s\{b_k\}} \approx Z, \quad k = 0, 1, 2, \dots, p-1$$

Yukarıdaki formüldeki Z deęeri standart normal deęerdir. $S\{b_k\}$ deęeri ise b_k 'nın tahmini standart sapmasıdır. Çoęu kez çoklu lojistik regresyon modelindeki X deęişkenlerinin alt grupları ile ilişkili regresyon katsayılarının önemli olup olmadığı araştırılır. Kullanılacak test prosedürü en çok olabilirlik tahmininin genelleştirilmiş

bir şeklidir. Büyük örnekler durumunda uygulanabilen bu test olabilirlik oranı testi olarak adlandırılır. Genel model olarak aşağıdaki model kullanılır.

$$\pi = \left[1 + \exp(-\beta'X) \right]^{-1}$$

Bu modelde ,

$$\beta'X = \beta_0 + \beta_1X_1 + \dots + \beta_{p-1}X_{p-1}$$

olurlar. Model için bulunacak en çok olabilirlik tahminlerini b_F ile gösterilir. Olabilirlik fonksiyonunu $L(\beta)$ ile gösterdiğimizizde $\beta = b_F$ olmaktadır. Genel modelde olabilirlik fonksiyonunun bu değerini $L(F)$ ile gösterilir. Test edilecek hipotezler şunlardır.

$$H_0: \beta_q = \beta_{q+1} = \dots = \beta_{p-1} = 0$$

H_1 : En az bir β_k değeri sıfırdan farklıdır.

En son p-q katsayılarını test etmek için model düzeltilir. Kısaltılmış model,

$$\pi = \left[1 + \exp(-\beta_R'X) \right]^{-1}$$

şeklindedir. Bu modelde

$$\beta_R' = \beta_0 + \beta_1X_1 + \dots + \beta_{q-1}X_{q-1}$$

olarak yazılır. Şimdi kısaltılmış model için maksimum olabilirlik tahminlerinin elde edişi şu şekildedir. Maksimum olabilirlik tahminleri b_R ile gösterilir. $\beta_R = b_R$ olduğunda q adet parametre ihtiva eden kısaltılmış model için olabilirlik tahmini tarif edebiliriz. Olabilirlik fonksiyonunun bu değeri L° ile gösterilir. L° değeri $L(F)$ değerini hiçbir zaman geçemez. Bu sebeple L° değeri $L(F)$ değerine yaklaştığında ilave parametreler olabilirliği fazlaca artırmayacakları için H_0 hipotezinin doğru

olduđuna karar verilir. L° deęeri $L(F)$ deęerinden yeterince kk olursa H_1 hipotezinin doęru olduđuna karar verilir.

Test istatistięi χ^2 gsterildięinde,¹²¹

$$\chi^2 = 2 \log_e \left[\frac{L(R)}{L(F)} \right] = -2 [\log_e L(R) - \log_e L(F)]$$

olur. rnek hacmi yeterince byk olduęunda H_0 hipotezi doęru ise χ^2 istatistięi yaklařık olarak $\chi^2_{(1-\alpha; p-q)}$ řeklinde daęılım gsterir. Serbestlik derecesi, $v = (n - q) - (n - p)$ řeklindedir.

Bylece $\chi^2 \leq \chi^2_{(1-\alpha; p-q)}$ olduęunda H_0 kabul edilirken $\chi^2 > \chi^2_{(1-\alpha; p-q)}$ olduęunda H_1 kabul edilir.

Regresyon katsayılarının nemli olup olmadıęını test etmede kullanılabilecek ikinci test Wald testidir. Wald testine ait test istatistiginin daęılımını standart normal daęılıma yaklařır. Her deęiřken iin listedeki standart hatalar kullanılarak Z testi yapılır. Wald testi, rnek hacminin byk olması durumunda anlam kazanmaktadır.

Eęim parametresinin en yksek ihtimal tahmincisi standart hatasının tahmini deęeri ile mukayese edilir. $\beta_1 = \hat{\beta}_1$ iken test istatistięinin daęılımını standart normal daęılıma uygundur. Bu teste ait test istatistik forml řu řekildedir.

$$W = \frac{\hat{\beta}_1}{SE(\hat{\beta}_1)}$$

Kurulan modelin uyum iyilięi testi Hosmer-Lemeshow'un hem onlu risk grupları hem de sabit kesim noktası yntemine gre hesaplanmaktadır. Uyum iyilięine karar vermek iin onlu risk grupları yntemine gre hesaplanmak řu řekildedir.

¹²¹ D.LAROSE , Data Mining Methods and Models , Wiley Publisherss 2006 s172-173.

$$\hat{C}_g^* = \sum_{k=0}^1 \sum_{l=1}^{10} \frac{(o_{kl} - e_{kl})^2}{e_{kl}}$$

Hosmer-Lemeshow $\hat{C}_g^*$ istatistiği, t-2 serbestlik dereceli ki-kare dağılımı göstermektedir.

Kestirilen modelin uyum iyiliği testi sabit kesim noktası yöntemiyle hesaplanmak istendiğinde ise, Hosmer-Lemeshow istatistigi kullanılmaktadır.

$$\hat{H}_g^* = \sum_{k=0}^1 \sum_{l=1}^{10} \frac{(o'_{kl} - e'_{kl})^2}{e'_{kl}}$$

Hosmer-Lemeshow $\hat{H}_g^*$ istatistiği, t-2 serbestlik dereceli ki-kare dağılımı göstermektedir.

Lojistik regresyonda parametreler şunlardır.

- n : İncelenen birey sayısı,
- $Y = (Y_1, Y_2, Y_3, \dots, Y_n)'$: açıklayıcı değişken vektörü,
- $\beta = (\beta_1, \beta_2, \beta_3, \dots, \beta_p)'$: parametre vektörü,
- $\varepsilon = (\varepsilon_1, \varepsilon_2, \varepsilon_3, \dots, \varepsilon_n)$: hata terimleri vektörü,
- p : değişken sayısı,
- X : açıklayıcı değişken matrisidir.

Açıklayıcı değişkenler matrisi X şu şekilde tanımlanır.

$$X = \begin{bmatrix} 1 & X_{1,1} & X_{2,1} & \dots & X_{p,1} \\ 1 & X_{1,2} & X_{2,2} & \dots & X_{p,2} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & X_{1,n} & X_{2,n} & \dots & X_{p,n} \end{bmatrix}$$

Lojistik regresyon analizinin uygulamadaki adımlarına bakıldığında şu aşamalardan geçmektedir.¹²²

- Önsel grup üyelikleri belirlenir.
- Modele girecek değişkenler belirlenir. Bu amaçla önsel bilgidен ya da istatistiksel tekniklerden yararlanılabilir.
- Modelin parametreleri Newton-Raphson yöntemi ile tahmin edilir. Ardından modelin tümünün anlamlılığı olabilirlik oranı ile test edilir. Model anlamlı değilse analize son verilir. Eğer model anlamlı bulunursa diğer aşamaya geçilir.
- Tahmin edilen model parametrelerinin tek tek anlamlılığı incelenir. Bu amaçla olabilirlik oranı ya da Wald istatistiği kullanılabilir. Her katsayının anlamlılığı incelendikten sonra, teklik oranları incelenerek, açıklayıcı değişkenlerin bağımlı değişken üzerindeki etkileri yorumlanabilir.
- Tahmin edilen model parametreleri kullanılarak, her bir gözlemin hangi gruptan geldiği tahmin edilir.
- Modelin uyum iyiliğini incelemek amacıyla doğru sınıflandırma yüzdesi ve yapay R^2 ölçütleri kullanılır. Modelin uyum iyiliği kabul edilebilir düzeyde ise 5. aşamadaki grup tahminleri kullanılabilir. Aksi halde 2. aşamaya geçilerek modele girecek değişkenler yeniden gözden geçirilir ve işlemler tekrar edilir.

¹²² K.ÖZDAMAR , Paket Programlar ile İstatistiksel Veri Analizi , Kaan Kitabevi 1999 s475-479.

2.5.4.Çok Boyutlu Ölçekleme

Çok boyutlu ölçekleme , veri boyutunu indirgeme amaçlı kullanılan bir tekniktir.Faktör analizinin alternatifi olduğu söylenebilir.Ölçülebilen ve ölçülemeyen ölçekleme yöntemleri olarak iki ayrılmaktadır.Verilerlerinin birbirine olan uzaklıklarına bakılarak indirgeme yapılan yöntem psikoloji ve sosyal bilimlerde yaygın olarak kullanılmaktadır.

Çok boyutlu ölçekleme analizi, n tane nesne ya da birim arasındaki p değişkene göre belirlenen uzaklıklara dayalı olarak nesnelerin k boyutlu ($k < p$) bir uzayda gösterimini elde etmeyi amaçlayan, böylece nesneler arasındaki ilişkileri belirlemeye yarayan bir yöntemdir.¹²³

Çok boyutlu ölçekleme analizi, hem görsel bir haritalama imkanı vermekte hem de gözlemler ile değişkenler arasındaki ilişkileri görsel olarak ortaya çıkarmaktadır .

Çok boyutlu ölçekleme, n tane nesne arasındaki uzaklık değerlerini kullanarak bu nesnelerin çok boyutlu uzaydaki konumlarını, ilişki yapısını veren resmini ortaya koymayı amaçlamaktadır. Bu analizde X veri matrisi yerine n tane bireyin uzaklıklardan elde edilen $n \times n$ boyutlu D uzaklıklar matrisi kullanılmaktadır. Uzaklıklar matrisinin simetrikliği nedeniyle, işlemler $\frac{n(n-1)}{2}$ tane uzaklık değeri kullanılarak sürdürülmektedir .

Çok boyutlu ölçekleme, nesneler arasındaki ilişkilerin bilinmediği, fakat aralarındaki uzaklıkların hesaplanabildiği durumlarda uzaklıklardan yararlanarak nesneler arasındaki ilişkileri ortaya koymaya yardımcı olan bir istatistiksel tekniktir. Uzaklıklar veya farklılıklar yardımıyla nesnelerin geometrik konumlarının belirlenmesi, şekillendirilmeleri önemli bir konudur. Bu amaçla yapılan çalışmalarda genellikle elde edilen şekillerin çok boyut içermesi sebebiyle bu ölçeklemelere çok boyutlu ölçekleme adı verilmiştir.

¹²³ K.KURTULUŞ ,Pazarlama Alıştırmaları ,Avcıol Yayını ,1998 s436-445.

Genellikle metrik ve metrik olmayan olarak iki çeşit çok boyutlu ölçekleme vardır. Metrik ölçmede veri en az mesafeli seviyede olmalıdır. Metrik olmayan modelde veriler en azından sıralama biçiminde olmalıdır.¹²⁴

Çok boyutlu ölçekleme analizi boyut indirgeme amacıyla da kullanılmaktadır. Aynı zamanda n tane nesne ya da birim arasındaki farklılıkları subjektif bir sıralamaya sokmak amacıyla da kullanılabilir. Çok boyutlu ölçekleme yöntemi uzaklıklar matrisinden faydalananarak çözüm yapar.

Bu nedenle analizde kullanılacak veri türüne uygun olarak uzaklık matrisi hesaplamak gereklidir. Çok boyutlu ölçekleme yönteminde hesaplanan uzaklık matrisine farklılık matrisi denir.

Çok boyutlu ölçekleme analizi belli bir dağılım varsayımı gerektirmeyen bir yöntemdir. Fakat buna karşın bu yöntemin sağlaması gereken bazı varsayımlar vardır. Bunlar;¹²⁵

- Çok boyutlu ölçekleme, n tane nesne yada birim arasındaki uzaklıkları kullanır. Bu uzaklıklar simetrik ve yansımalıdır.
- Veriler sınıflama veya sıralama düzeyinde ölçülerek nesneler arası uzaklıklar hesaplanmış ise , değerlerin sıraya dizilmesi ve monotonik regresyona göre konfigürasyon uzaklıklarının belirlenmesi gerekir.
- Çok boyutlu ölçekleme içinde yer alan yöntemlerin uygulanabilmesi için yöntemin gerektirdiği veri tipini doğru olarak belirlemek gerekir. Örneğin; sınıflamalı, sıralı, eşit aralıklı veya orantılıdır

¹²⁴ H.TATLIDİL , Uygulamalı Çok Değişkenli İstatistiksel Analiz , Hacettepe ün. 1996 s353.

¹²⁵ N.H.TIMM ,Applied Multivariate Analysis, Springer Publishers 2002 s541-544.

- Çok boyutlu ölçekleme ile analiz edilecek veriler farklılıklar belirtiyor ise, farklılıklar matrisi nicel değerler içermeli ve tüm farklılık ölçüleri aynı ölçümleme ile hesaplanmış olmalıdır.

BÖLÜM 3 SİGORTACILIK SEKTÖRÜNDE MÜŞTERİ İLİŞKİLERİ YÖNETİMİ YAKLAŞIMIYLA BİR VERİ MADENCİLİĞİ UYGULAMASI

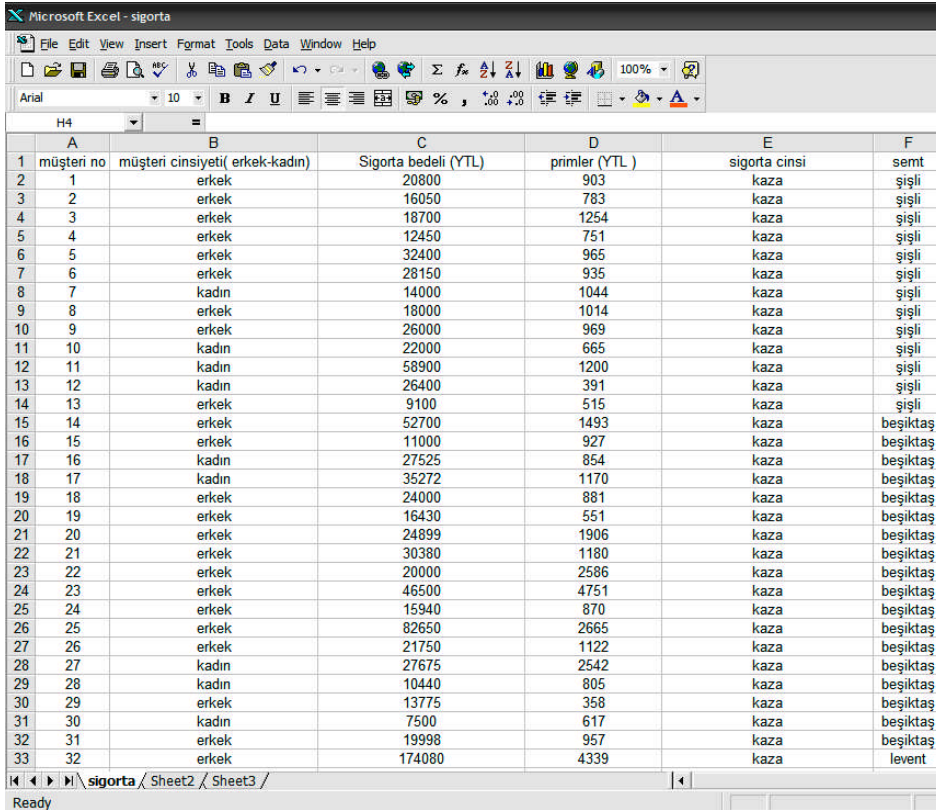
3.1.ARAŞTIRMANIN AMACI

Bu uygulamada x sigorta şirketinin veri tabanından alınan , yangın sigortası , konut sigortası ve de kaza sigortası verileri veri madenciliği analizi için kullanılmaktadır.Sigorta satın alan müşterilerin davranış kalıplarını ortaya çıkartarak müşterileri gruplamak ve bu müşterilere uygun bir satış kampanyası hazırlanması için bu veri seti kullanılmaktadır.

3.2.ARAŞTIRMANIN KAPSAMI

Bu araştırmada kullanılan veri seti 111 adet veriden oluşurken değişkenleri ise müşteri numarası , müşteri cinsiyeti , sigorta bedeli , primler , sigorta cinsi ve de semt değişkenidir. Uygulamada kullanılan veri seti şu şekildedir.

Tablo 3.1.Verit seti tablosu



	A	B	C	D	E	F
1	müşteri no	müşteri cinsiyeti(erkek-kadın)	Sigorta bedeli (YTL)	primler (YTL)	sigorta cinsi	semt
2	1	erkek	20800	903	kaza	şişli
3	2	erkek	16050	783	kaza	şişli
4	3	erkek	18700	1254	kaza	şişli
5	4	erkek	12450	751	kaza	şişli
6	5	erkek	32400	965	kaza	şişli
7	6	erkek	28150	935	kaza	şişli
8	7	kadın	14000	1044	kaza	şişli
9	8	erkek	18000	1014	kaza	şişli
10	9	erkek	26000	969	kaza	şişli
11	10	kadın	22000	665	kaza	şişli
12	11	kadın	58900	1200	kaza	şişli
13	12	kadın	26400	391	kaza	şişli
14	13	erkek	9100	515	kaza	şişli
15	14	erkek	52700	1493	kaza	beşiktaş
16	15	erkek	11000	927	kaza	beşiktaş
17	16	kadın	27525	854	kaza	beşiktaş
18	17	kadın	35272	1170	kaza	beşiktaş
19	18	erkek	24000	881	kaza	beşiktaş
20	19	erkek	16430	551	kaza	beşiktaş
21	20	erkek	24899	1906	kaza	beşiktaş
22	21	erkek	30380	1180	kaza	beşiktaş
23	22	erkek	20000	2586	kaza	beşiktaş
24	23	erkek	46500	4751	kaza	beşiktaş
25	24	erkek	15940	870	kaza	beşiktaş
26	25	erkek	82650	2665	kaza	beşiktaş
27	26	erkek	21750	1122	kaza	beşiktaş
28	27	kadın	27675	2542	kaza	beşiktaş
29	28	kadın	10440	805	kaza	beşiktaş
30	29	erkek	13775	358	kaza	beşiktaş
31	30	kadın	7500	617	kaza	beşiktaş
32	31	erkek	19998	957	kaza	beşiktaş
33	32	erkek	174080	4339	kaza	levent

Müşteri numarası değişkeni

Bu değişken sürekli bir değişken olup müşteri sıralamasını belirtmektedir.

Müşteri cinsiyeti değişkeni

Bu değişken sigorta satın alan müşterilerin cinsiyetlerini göstermektedir. Kesikli verilerden oluşan bu değişken kadın-erkek değerlerini alabilmektedir. Bu değişkenin özet bilgileri şu şekildedir.

Tablo 3.2. Cinsiyet değişkeninin özet bilgileri

Results					
Attribute	Gini	Distribution			
müşteri cinsiyeti(erkek-kadın)	0,4503	Values	Count	Percent	Histogram
		erkek	73	65,77 %	
		kadın	38	34,23 %	

Yukarıdaki tabloya göre sigorta satın alan müşterilerin % 65,77 'si erkek müşteri iken % 34,23 'ü kadın müşteridir.

Sigorta bedeli (YTL) değişkeni

Bu değişken sigortalanacak olan malın belirlenmiş olan bedel değerlerini içermektedir. Bu değişkenin dağılımı şu şekildedir.

Tablo 3.3. Sigorta bedeli değişkeninin özet bilgileri

Results					
Attribute	Min	Max	Average	Std-dev	Std-dev/avg
Sigorta bedeli (YTL)	624	400000	40689,9099	50355,6361	1,2375

Yukarıdaki tabloya göre bu değişken en az 624 YTL en fazla 400000 YTL değer alabilmektedir. Ortalama değeri 40689,90 YTL iken standart sapması 50355,64 'dür.

Primler (YTL) değişkeni

Bu değişken sigorta satın alan müşterilerin toplam ödeyecek oldukları prim değerlerini göstermektedir. Bu değişkenin özet bilgileri şöyledir.

Tablo 3.4. Primler değişkeninin özet bilgileri

Results					
Attribute	Min	Max	Average	Std-dev	Std-dev/avg
primler (YTL)	25	4751	790,8108	801,0788	1,0130

Yukarıdaki tabloya göre bu değişken en az 25 YTL en fazla 4751 YTL değer alabilmektedir. Ortalama değeri 790 ,81 YTL iken standart sapması 801,08 ‘ dir.

Sigorta cinsi değişkeni

Bu değişken sigorta satın alan müşterilerin hangi sigortayı tercih ettiklerini göstermektedir. Kesikli verilerden oluşan bu değişken yangın-kaza-konut değerlerini alabilmektedir. Bu değişkenin özet bilgileri şu şekildedir.

Tablo 3.5. Sigorta cinsi değişkeninin özet bilgileri

Results					
Attribute	Gini	Distribution			
		Values	Count	Percent	Histogram
sigorta cinsi	0,6607	kaza	40	36,04 %	
		konut	41	36,94 %	
		yangın	30	27,03 %	

Yukarıdaki tabloya göre bu değişkenin % 36,04 ‘ü kaza sigortalarından , % 36,94 ‘ü konut sigortalarından % 27,03 ‘ü ise yangın sigortalarından oluşmaktadır.

Semt değişkeni

Bu deęişken sigortalayı alan müşterinin hangi semtte oturduęunu gösteren bir deęişkendir. Kesikli verilerden oluşan bu deęişken Şişli-Beşiktaş-Levent deęerlerini alabilmektedir.Bu deęişkenin özet bilgileri şu şekildedir. Bu deęişkenin özet bilgileri şu şekildedir.

Tablo 3.6.Semt deęişkeninin özet bilgileri

Results					
Attribute	Gini	Distribution			
		Values	Count	Percent	Histogram
semt	0,6587	sisli	37	33,33 %	
		besiktas	44	39,64 %	
		levent	30	27,03 %	

Yukarıdaki tabloya göre müşterilerin % 33,33 ‘ü Şişli , % 39,64 ‘ü Beşiktaş %27,03 ‘ü ise Levent semtlerinde oturmaktadırlar.

Bu veri setine uygulanacak olan analizler şunlardır.

- Birliktelik kuralları analizi
 - Sepet analizi ,
- Sınıflandırma analizleri
 - C&RT analizi ,
 - C4.5 analizi ,
- Kümeleme analizleri
 - K-ortalamlar(k-means) analizi ,
 - Kanonikel diskriminant analizi ,

Yapılacak olan analizlerin ilki olan sepet analizi , bir müşterinin bir sigortayı satın aldıktan sonra bir başka sigortayı satın alma olasılıęını ortaya koymaktadır.Bu analiz belirli müşteri gruplarına belli bir poliçenin yanında

başka bir poliçesinde taktif edilmesine olanak sağlayarak karlılığı arttırmaya yardımcı olmaktadır.

İkinci olarak uygulanacak olan analizler C&RT analizi ve C4.5 analizidir.Bu analizler müşterileri belirli kriterlere göre ayırıştırarak müşteri grupları oluşturmak için kullanılmaktadır.

Üçüncü olarak uygulanacak olan analizler K-ortalamalar(k-means) analizi ve Kanonikel diskriminant analizidir.Bu analizler , benzerliklerine göre müşterileri gruplayarak kümeler oluşturmaktadır.Bu kümeleme yöntemleri kullanılarak müşteri gruplarına ortak bir pazarlama kampanyası yürütülmesine olanak sağlamaktadır.

3.3.ARAŞTIRMA İLE İLGİLİ UYGULAMA

3.3.1.Birliktelik Kuralları Analizi

Birliktelik kuralları analizinden en yaygın olarak kullanılan analiz yöntemi Sepet analizidir.Bu analizin temeli koşullu olasılık varsayımlarına dayanmaktadır.Bu analizde kullanılan formüller $P(X \cap Y)$ destek kriteri ve $P(X/Y)$ güven kriteridir.Güven kriterini hesaplamak için şu formül kullanılır.

$$P(X/Y) = \frac{P(X \cap Y)}{P(Y)}$$

Kullanılacak olan değişkenler aşağıda açıklanmıştır.

- E : Sigorta satın alan erkek müşteri sayısı ,
- K : Sigorta satın alan kadın müşteri sayısı ,
- Y : Yangın sigortası satın alan müşteri sayısı ,
- Ka : Kaza sigortası satın alan müşteri sayısı ,
- Ko : Konut sigortası satın alan müşteri sayısı ,
- S : Şişli semtindeki müşteri sayısı ,

L : Levent semtindeki müşteri sayısı ,

B: Beşiktaş semtindeki müşteri sayısıdır.

Bu analizde ilk olarak kadın ve erkek müşterilerin yangın sigortası satın alma olasılığını hesaplamak için kullanılacaktır.Bunun için kullanılacak olan tablo aşağıda verilmiştir.

Tablo 3.7.Yangın sigortası içindeki kadın-erkek oranı tablosu

sigorta cinsi=yangın			
Examples		[27,0 %] 30	
Att - Desc	Test value	Group	Overral
Continuous attributes : Mean (StdDev)			
Discrete attributes : [Recall] Accuracy			
müşteri cinsiyeti(erkek-kadın)=erkek	1,0	[30,1 %] 73,3 %	65,8 %
müşteri cinsiyeti(erkek-kadın)=kadın	-1,0	[21,1 %] 26,7 %	34,2 %

Yukarıdaki tablo kullanılarak elde edilen formül şu şekildedir.

$$P(Y / E) = \frac{P(Y \cap E)}{P(E)}$$

Bu formül sigorta satın alan bir erkek müşterinin yangın sigortası alma olasılığını gösterir.

$$P(Y \cap E) = \%73,3 * 30 = 21,99$$

$$P(E) = \%65,8 * 111 = 73,038$$

$$P(Y / E) = \frac{21,99}{73,038} = 0,3010$$

Sigorta satın alan bir erkek müşterinin yangın sigortası alması olasılığı % 30,10 ‘dur.Aynı analizi sigorta satın alan bir kadın müşteri için yaparsak şu sonuç elde edilir.

$$P(Y \cap K) = \%26,7 * 30 = 8,01$$

$$P(K) = \%34,2 * 111 = 37,962$$

$$P(Y / K) = \frac{8,01}{37,962} = 0,211$$

Sigorta satın alan bir kadın müşterinin yangın sigortası alması olasılığı % 21,1‘dir.

Bu analizde ikinci olarak kadın ve erkek müşterilerin kaza sigortası satın alma olasılığı hesaplanacaktır.Bunun için kullanılacak olan tablo aşağıda verilmiştir.

Tablo 3.8.Kaza sigortası içindeki kadın-erkek oranı tablosu

sigorta cinsi=kaza			
Examples		[36,0 %] 40	
Att - Desc	Test value	Group	Overral
Continuous attributes : Mean (StdDev)			
Discrete attributes : [Recall] Accuracy			
müpteri cinsiyeti(erkek-kadın)=erkek	1,5	[41,1 %] 75,0 %	65,8 %
müpteri cinsiyeti(erkek-kadın)=kadın	-1,5	[26,3 %] 25,0 %	34,2 %

Sigorta satın alan bir erkek müşterinin kaza sigortası alma olasılığını şu şekilde hesaplanır.

$$P(Ka \cap E) = \%75 * 40 = 30$$

$$P(E) = \%65,8 * 111 = 73,038$$

$$P(Ka / E) = \frac{30}{73,038} = 0,4107$$

Sigorta satın alan bir erkek müşterinin kaza sigortası alması olasılığı % 41,07 'dir.Aynı analizi sigorta satın alan bir kadın müşteri için yaparsak şu sonuç elde edilir.

$$P(K_a \cap K) = \%25 * 40 = 10$$

$$P(K) = \%34,2 * 111 = 37,962$$

$$P(K_a / K) = \frac{10}{37,962} = 0,2634$$

Sigorta satın alan bir kadın müşterinin kaza sigortası alması olasılığı % 26,34'dür.

Bu analizde üçüncü olarak kadın ve erkek müşterilerin konut sigortası satın alma olasılığı hesaplanacaktır.Bunun için kullanılacak olan tablo aşağıda verilmiştir.

Tablo 3.9.Konut sigortası içindeki kadın-erkek oranı tablosu

sigorta cinsi=konut			
Examples		[36,9 %] 41	
Att - Desc	Test value	Group	Overall
Continuous attributes : Mean (StdDev)			
Discrete attributes : [Recall] Accuracy			
müpteri cinsiyeti(erkek-kadın)=kadın	2,5	[52,6 %] 48,8 %	34,2 %
müpteri cinsiyeti(erkek-kadın)=erkek	-2,5	[28,8 %] 51,2 %	65,8 %

Sigorta satın alan bir erkek müşterinin konut sigortası alma olasılığını şu şekilde hesaplanır.

$$P(K_o \cap E) = \%48,8 * 41 = 20$$

$$P(E) = \%65,8 * 111 = 73,038$$

$$P(K_o / E) = \frac{20}{73,038} = 0,2739$$

Sigorta satın alan bir erkek müşterinin kaza sigortası alması olasılığı % 27,39'dır. Aynı analizi sigorta satın alan bir kadın müşteri için yaparsak şu sonuç elde edilir.

$$P(K_0 \cap K) = \%51,2 * 41 = 20,992$$

$$P(K) = \%34,2 * 111 = 37,962$$

$$P(K_0 / K) = \frac{20,992}{37,962} = 0,5530$$

Sigorta satın alan bir kadın müşterinin konut sigortası alması olasılığı % 55,30'dur.

3.3.2. Sınıflandırma Analizleri

Sınıflandırma analizi veri setini belirli kriterlere göre ayrıştırarak veri setini doğru yorumlamamıza yardımcı olmaktadır. Burada uygulanacak olan analizler C&RT analizi ve C4.5 analizidir. Bu analizlerin sonuç değerleri karar ağacı şeklinde belirtilmektedir. Bu analizler aşağıda veri setine uygulanmaktadır.

C&RT Analizi

Veri setine , müşterilerin yerleşim yerleri esas alınarak yapılan C&RT analizi şu şekildedir.

Tablo 3.10. C&RT analizi sonuç matrisi ve hata oranları

Error rate			0,4595				
Values prediction			Confusion matrix				
Value	Recall	1-Precision		sisli	besiktas	levent	Sum
sisli	0,4595	0,3929	sisli	17	20	0	37
besiktas	0,9318	0,4938	besiktas	3	41	0	44
levent	0,0667	0,0000	levent	8	20	2	30
			Sum	28	81	2	111

Yukarıdaki tablo C&RT analizinin oluşturduğu matris değerleridir. Bu analizin yaprak sayısını belirlemek için kullanılan tablo aşağıda verilmiştir.

Tablo 3.11. C&RT analizinin yaprak analiz sonuçları

N°	# Leaves	Err (growing set)	Err (pruning set)
5	1	0,5541	0,7027
4	2	0,5270	0,7027
3	9	0,3784	0,6216
2	12	0,3243	0,6216
1	21	0,2027	0,5676

Burada oluşan hata değerlerinden , C&RT analizi için kullanılacak olan yaprak sayısı belirlenmektedir.Bunun için genel hata oranı ile üretilen veri setinin (growing set) hata değeri karşılaştırılır. Analiz sonucunda elde edilen karar ağacı şu şekildedir.

Tablo 3.12. C&RT analizi sonucunda elde edilen karar ağacı diyagramı

- Sigorta bedeli (YTL) < 1410,0000 then semt = levent (100,00 % of 2 examples)
- Sigorta bedeli (YTL) >= 1410,0000
 - primler (YTL) < 1373,5000
 - Sigorta bedeli (YTL) < 2525,0000 then semt = besiktas (100,00 % of 3 examples)
 - Sigorta bedeli (YTL) >= 2525,0000
 - primler (YTL) < 101,5000 then semt = sisli (100,00 % of 2 examples)
 - primler (YTL) >= 101,5000
 - primler (YTL) < 1190,0000
 - primler (YTL) < 1111,5000
 - primler (YTL) < 963,0000
 - Sigorta bedeli (YTL) < 10270,0000 then semt = sisli (66,67 % of 6 examples)
 - Sigorta bedeli (YTL) >= 10270,0000 then semt = besiktas (47,62 % of 42 examples)
 - primler (YTL) >= 963,0000 then semt = sisli (57,14 % of 7 examples)
 - primler (YTL) >= 1111,5000 then semt = besiktas (100,00 % of 4 examples)
 - primler (YTL) >= 1190,0000 then semt = sisli (100,00 % of 2 examples)
- primler (YTL) >= 1373,5000 then semt = besiktas (83,33 % of 6 examples)

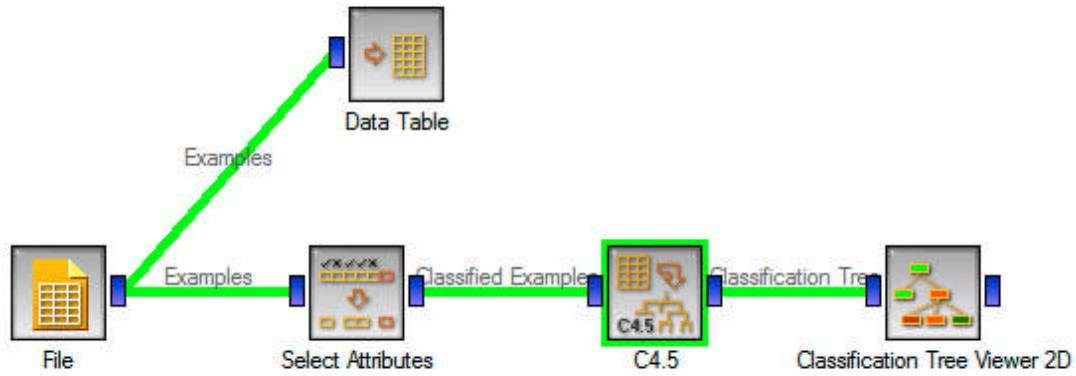
Yapılan analize göre veri seti iki ana gruba ayrılmıştır.ilk grup sigorta bedeli 1410 YTL ‘den küçük müşteriler , ikinci grup ise sigorta bedeli 1410 YTL’den büyük olanlar olarak ayrılmıştır.Buna göre ;

- Sigorta bedeli 1410 YTL ‘den küçük olan bir müşterinin tamamı Levent semtinde oturmaktadır.
- Sigorta bedeli 1410 YTL ve de büyük olan bir müşterinin için ;

- Sigorta bedeli 1410 YTL ile 2525 YTL arasında ve sigorta primleri 1373 YTL ‘den küçük olan müşterilerin % 100 Beşiktaş semtinde oturmaktadır.
- Sigorta bedeli 2525 YTL ‘den büyük ve sigorta primleri 101 YTL ‘den küçük olan müşterilerin % 100 Şişli semtinde oturmaktadır.
- Sigorta bedeli 2525 YTL ile 10270 YTL arasında ve sigorta primleri 101 YTL ile 963 YTL arasında olan müşterilerin % 66,67 ‘si Şişli semtinde oturmaktadır.
- Sigorta bedeli 10270 YTL ve daha büyük değerler için ve sigorta primleri 101 YTL ile 963 YTL arasında olan müşterilerin % 47,62 ‘si Beşiktaş semtinde oturmaktadır.
- Sigorta bedeli 2525 YTL ‘den büyük ve sigorta primleri 963 YTL ile 1111 YTL arasında olan müşterilerin % 57,14 Şişli semtinde oturmaktadır.
- Sigorta bedeli 2525 YTL ‘den büyük ve sigorta primleri 1111 YTL ile 1190 YTL arasında olan müşterilerin % 100 Beşiktaş semtinde oturmaktadır.
- Sigorta bedeli 2525 YTL ‘den büyük ve sigorta primleri 1190 YTL ile 1373 YTL arasında olan müşterilerin % 100 Şişli semtinde oturmaktadır.
- Sigorta bedeli 2525 YTL ‘den büyük ve sigorta primleri 1373 YTL ve daha büyük olan müşterilerin % 83,33 Beşiktaş semtinde oturmaktadır.

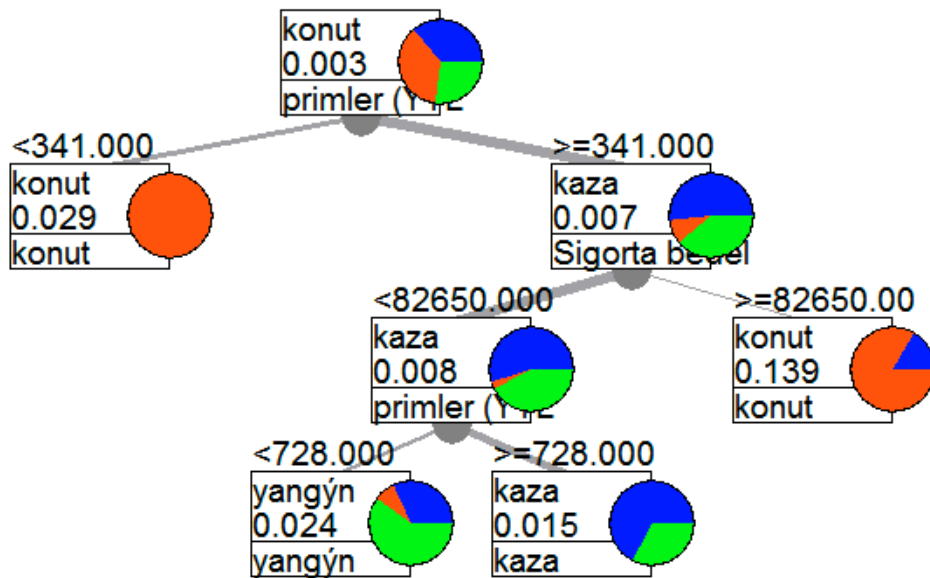
C4.5 Analizi

Bu analiz , en iyi karar ağacı analizidir.Özellikle veri setinde belirli kurallar oluşturmak için kullanılır.Verdi setine uygulanan analiz süreci aşağıda belirtilmiştir.



Şekil 3.1. C 4.5 analizi uygulama süreci

Uygulanan C4.5 analiz sonucu elde edilen karar ağacı elde aşağıdaki şekilde gösterilmektedir.



Şekil 3.2. C 4.5 analizi sonucu elde edilen karar ağacı diyagramı

Elde edilen karar ağacının belirlenmesinde sigorta cinsi değişkeni esas alınmıştır.Buna göre ;

- Prim değerleri 341 YTL ‘den az olan müşterilerin % 30,6’sı konut sigortalarını tercih etmiştir.
- Prim değerleri 341 YTL ve daha fazla olan müşterilerin %51,9 ‘u kaza sigortalarını , %39,0 ‘ı yangın sigortalarını ve % 9,1’i konut sigortalarını tercih etmiştir.
 - Prim değerleri 341 YTL ve daha fazlası ve sigorta bedeli 82650 YTL ‘den daha fazla olan müşterilerin % 83,3 ‘ü konut sigortalarını , %16,7’ si kaza sigortalarını tercih etmiştir.
 - Prim değerleri 341 YTL ve daha fazlası ve sigorta bedeli 82650 YTL ‘den daha az olan müşterilerin % 2,8 ‘i konut sigortalarını , %54,9’u kaza sigortalarını ve % 42,3 ‘ü ise yangın sigortalarını tercih etmiştir.
 - Prim değerleri 341 YTL ile 728 YTL arasındaki değerler ve sigorta bedeli 82650 YTL ‘den daha az olan müşterilerin % 8 ‘i konut sigortalarını , % 32 ‘si kaza sigortalarını ve % 60 ‘ı ise yangın sigortalarını tercih etmiştir.
 - Prim değerleri 728 YTL ‘den daha büyük değerler için ve sigorta bedeli 82650 YTL ‘den daha az olan müşterilerin % 67,4 ‘ü kaza sigortalarını ve % 32,6 ‘sı ise yangın sigortalarını tercih etmiştir.

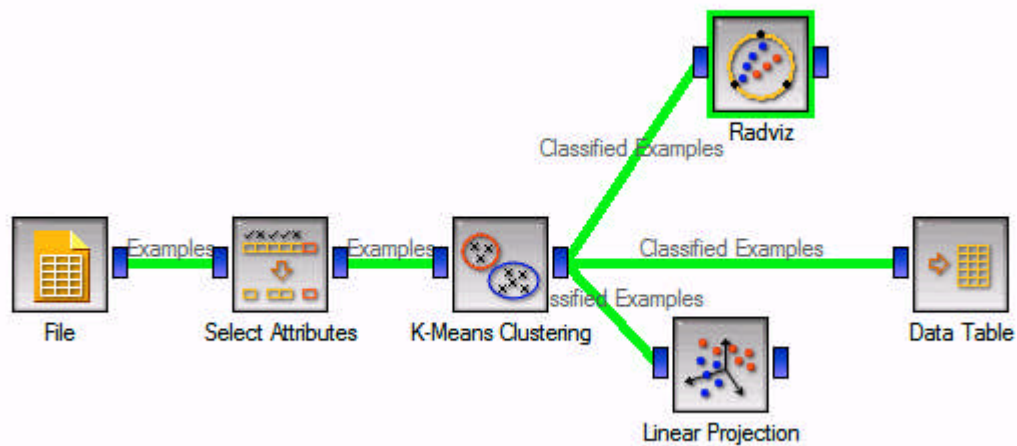
Müşterilerin tanımları çerçevesinde hangi ürünü daha çok aldıklarını gösteren C4.5 analizi , özellikle satış kampanyalarının belirlenmesinde önemli bir rol oynamaktadır.

3.3.2. Kümeleme Analizleri

Kümeleme analizleri , veri setindeki benzerliklerini dikkate alarak verileri belirli gruplara atamamıza olanak sağlayan önemli bir veri madenciliği tekniğidir.Burada kullanılacak olan kümeleme analizleri , K-Ortalamalar (K-MEANS) analizi ve Kanonikel diskriminant analizidir.

K-Ortalamalar (K-MEANS) Analizi

Veri setinin bu yöntemler kümelere ayrılması süreci şu şekildedir.



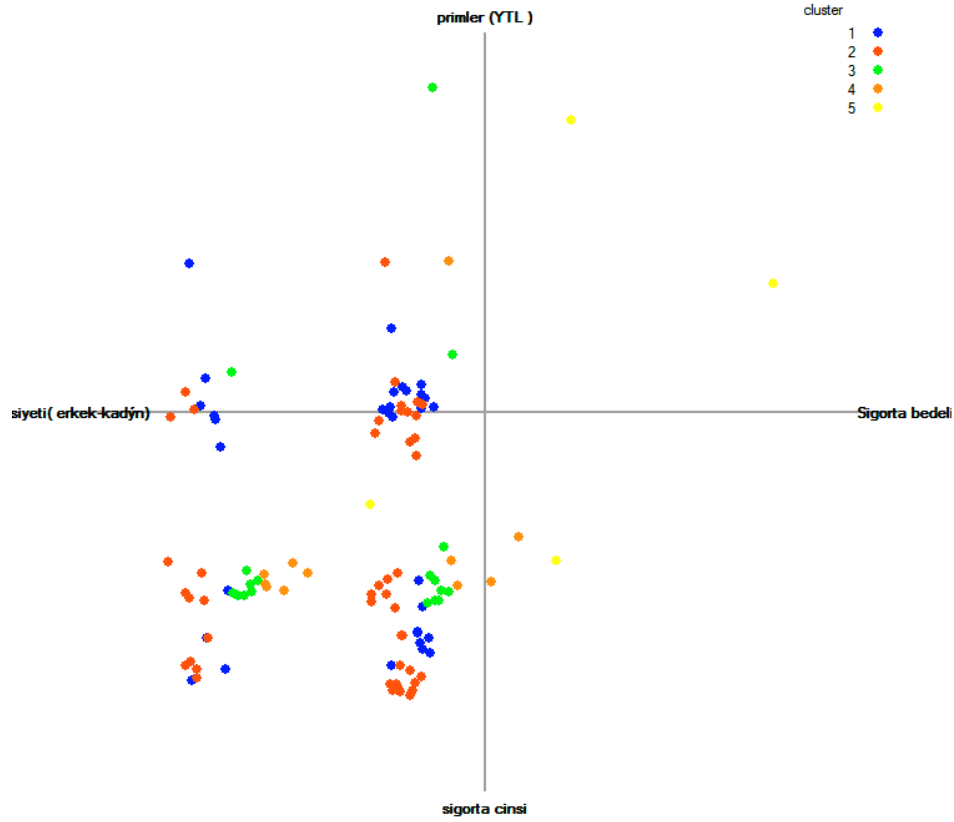
Şekil 3.3. K-ortalamalar analizi uygulama süreci

Bu analiz sürecinin sonucunda , oluşan kümelerin yapıları aşağıdaki tabloda belirtilmiştir.

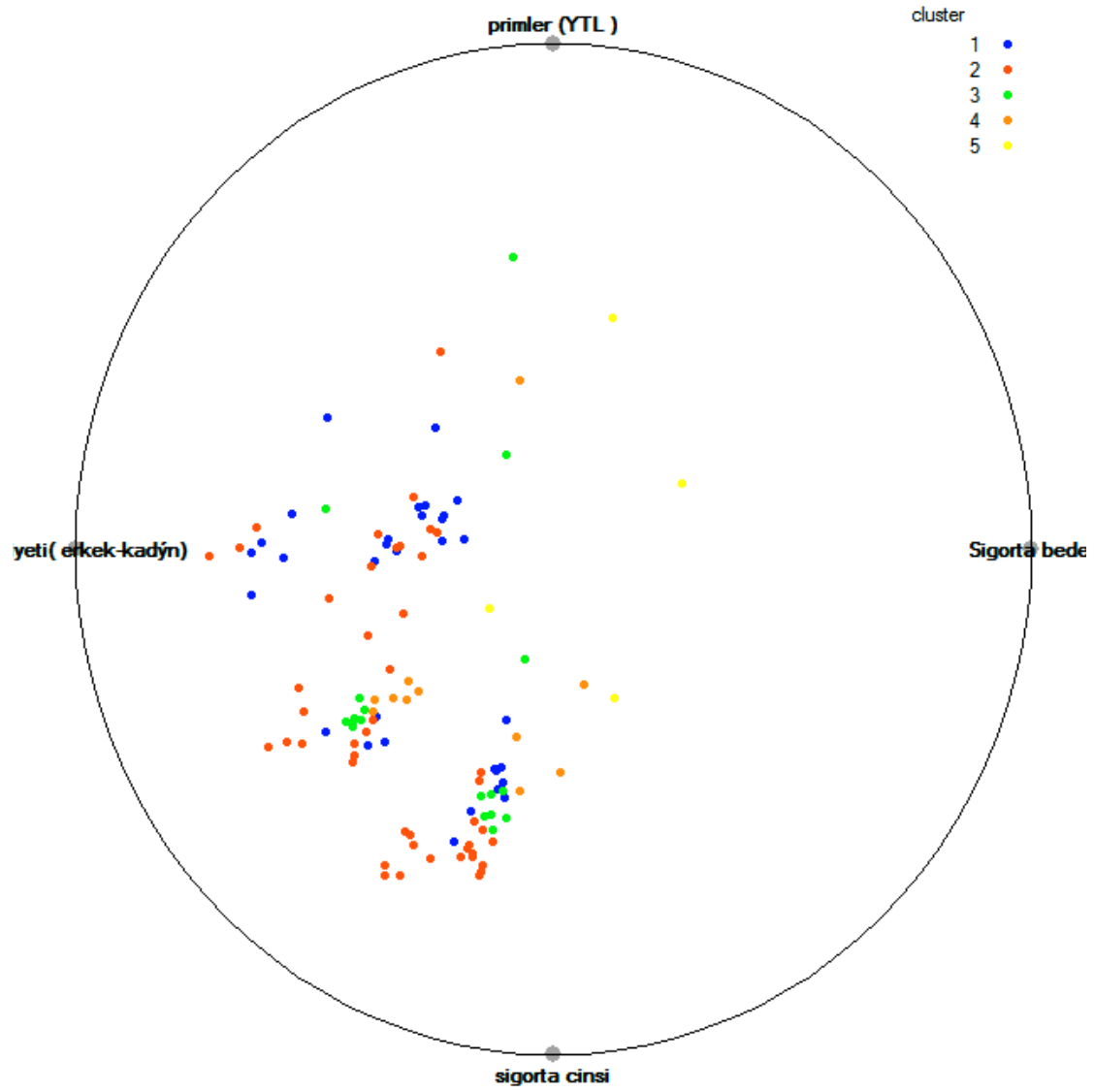
Tablo 3.13. K-ortalamalar analizi sonucu elde edilen kümeler

	ID	Items	Fitness	BIC
1	1	32	0.579	1772.36
2	2	46	0.518	2560.84
3	3	18	0.686	974.74
4	4	11	0.426	570.46
5	5	4	0.146	159.23
6	Total	111	0.540	6377.34

Oluřturulan kümelerin uygunluk derecesine bakıldığında uygunluđu en iyi olan küme 0.686 ile 3.kümedir.İkinci en iyi uygunluđa sahip olan küme ise 0.579 ile 1.kümedir. Oluřturulan kümelerin grafiksel olarak görüntüsü řu řekildedir.



řekil 3.4. K-ortalamlar analizi sonucu elde edilen kümelerin grafiksel görüntüsü



Şekil 3.5. K-ortalamalar analizi sonucu elde edilen kümelerin grafiksel görüntüsü

Yukarıda belirtilen grafikler , değişkenlere göre oluşturulan kümelerin dağılımlarını göstermektedir.

Kümeleri oluşturulan verilerin liste biçimi şu şekildedir.

Tablo3.14. K-ortalamlar analizi sonucu elde edilen kümelerin veri listesi

	Sigorta bedeli (YTL)	primler (YTL)	müşteri cinsiyeti(erkek-kadın)	sigorta cinsi	cluster
1	400000	3938	erkek	konut	5
2	188700	450	erkek	konut	5
3	200000	1344	kadın	konut	5
4	174080	4339	erkek	kaza	5
5	93600	306	erkek	konut	4
6	78750	110	kadın	konut	4
7	110000	338	kadın	konut	4
8	111100	228	kadın	konut	4
9	86355	392	erkek	konut	4
10	98100	262	kadın	konut	4
11	124020	911	erkek	konut	4
12	75750	168	erkek	konut	4
13	80800	195	kadın	konut	4
14	107000	341	kadın	konut	4
15	82650	2665	erkek	kaza	4
16	53600	2214	erkek	yangın	3
17	64350	250	erkek	konut	3
18	66000	262	kadın	konut	3
19	60000	110	erkek	konut	3
20	55550	138	kadın	konut	3
21	60600	94	erkek	konut	3
22	51500	103	erkek	konut	3

Görüldüğü gibi K-ortalamlar analizi veri setinde belirli gruplar oluşturarak oluşturulan gruplara aynı satış kampanyasını uygulamamıza olanak sağlamaktadır.

Kanonik Diskriminant Analizi

Veri setine bu analiz sigorta cinsi esas alarak uygulanmıştır.Buradan elde edilen vektörlerin tablosu şu şekildedir.

Tablo 3.15. Kanonik diskriminant analizi sonuçları

Coefficients	Unstandardized	
Attribute	Root n°1	Root n°2
müşteri no	0,0908	-0,0114
Sigorta bedeli (YTL)	0,0000	0,0000
primler (YTL)	-0,0003	-0,0017

Buradan elde edilen vektörler şunlardır.

$$Z_1 = -4,8 + 0,09M + 0S - 0,0003P$$

$$Z_2 = 0,915 - 0,0114M + 0S - 0,0017P$$

Bu vektörlerin güvenirlik tablosu aşağıdaki belirtilmiştir.

Tablo 3.16. Kanonikel diskriminant analizi sonucu elde edilen vektörler

Root	Eigenvalue	Proportion	Canonical R	Wilks Lambda	CHI-2	d.f.	p-value
1	8,05290	0,87642	0,943153	0,051727	316,9109	6	0,000000
2	1,13550	1,00000	0,729194	0,468275	81,1808	2	0,000000

P test istatistiğine bakıldığında Z_1 ve Z_2 vektörlerinin anlamlı bir vektör olduğu söylenebilmektedir.

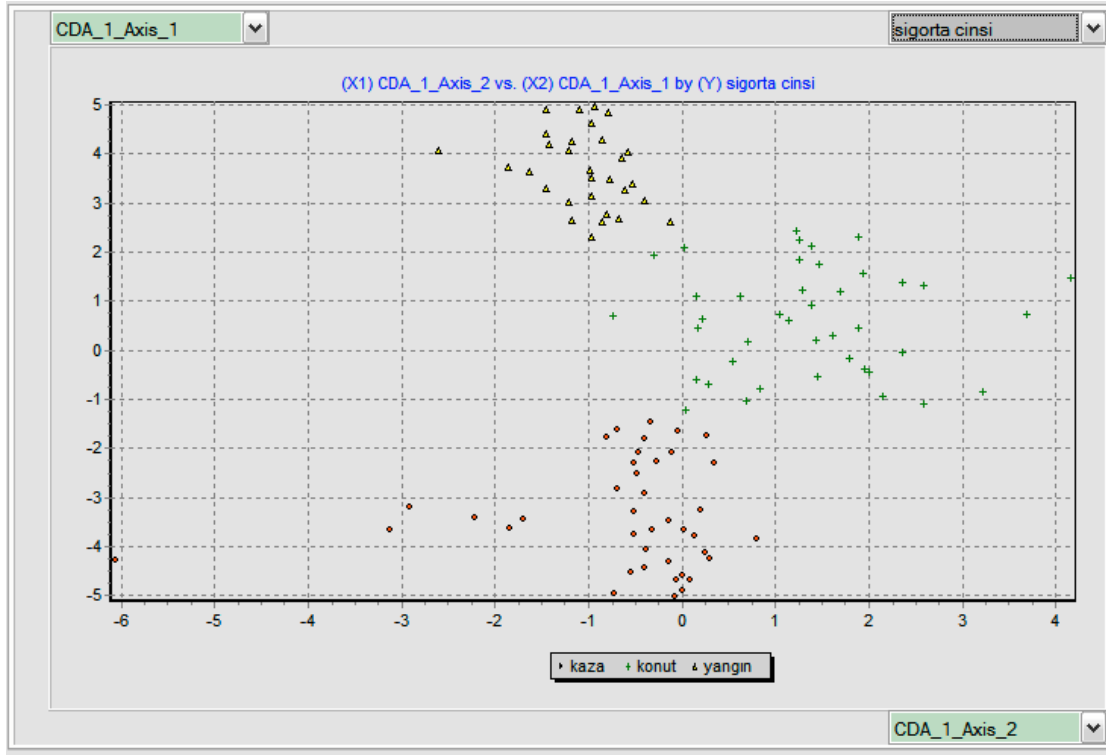
Tablo 3.17. Elde edilen vektörlerle değişkenler arasındaki ilişki

Root	Root n°1			Root n°2		
Descriptors	Total	Within	Between	Total	Within	Between
müpteri no	0,996	0,968	0,999	-0,043	-0,085	-0,033
Sigorta bedeli (YTL)	0,006	0,002	0,014	0,542	0,404	1,000
primler (YTL)	-0,269	-0,102	-0,532	-0,553	-0,431	-0,847

Z_1 vektörü , müşteri numarası değişkeni ile pozitif yönde 0.996 oranında , sigorta bedeli değişkeni ile pozitif yönde 0,006 oranında ve primler değişkeni ile ters yönde 0,269 oranında bir ilişki vardır.

Z_2 vektörü , müşteri numarası değişkeni ile ters yönde 0.043 oranında , sigorta bedeli değişkeni ile pozitif yönde 0,542 oranında ve primler değişkeni ile ters yönde 0,553 oranında bir ilişki vardır.

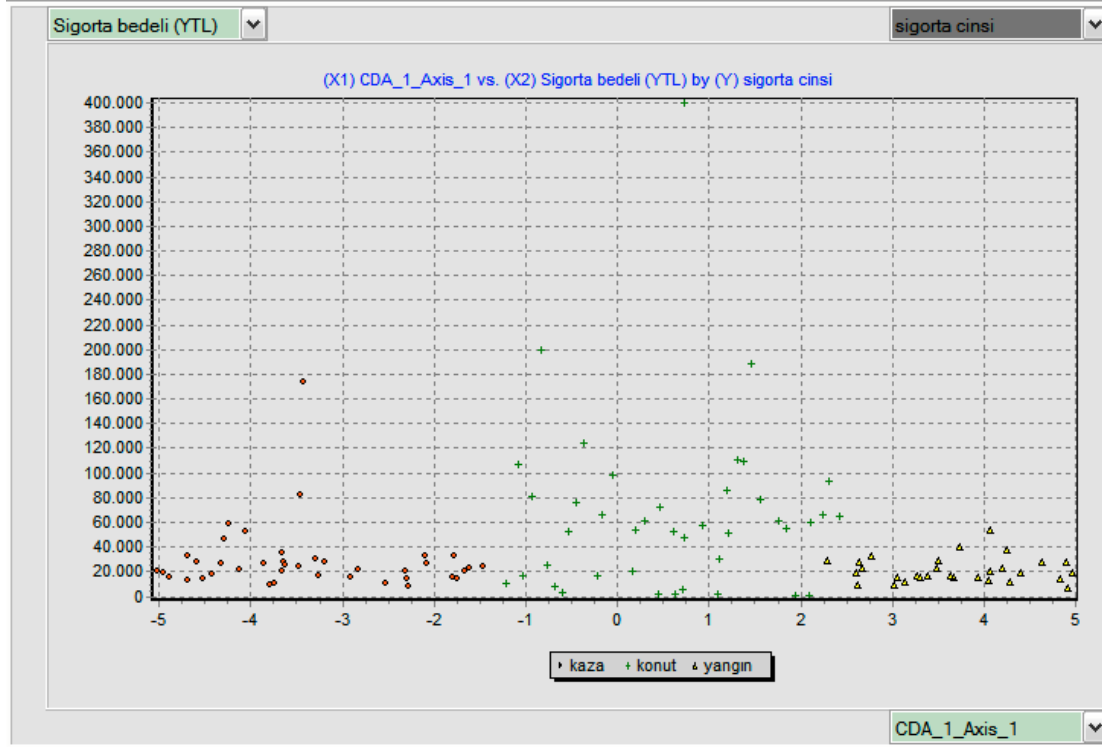
Buradan elde edilen vektörlerin grafiksel görüntüsü şu şekildedir.



Şekil 3.6. Elde edilen vektörlerle arasındaki ilişkiyi gösteren grafik

Bu grafikteki X eksenini , veri setinin gruplanmasına yardımcı olmaktadır. X-ekseninde , -1 ile -3 değerleri arasındaki veriler iyi , -1 ile 1 değerler arasındaki veriler normal ve 1 ve 3 değerleri arasındaki veriler kötü olarak tanımlanmaktadır. Buna göre kaza sigortası değerleri normal gruba girerken yangın sigortası iyi ve de konut sigortası kötü gruba girmektedirler.

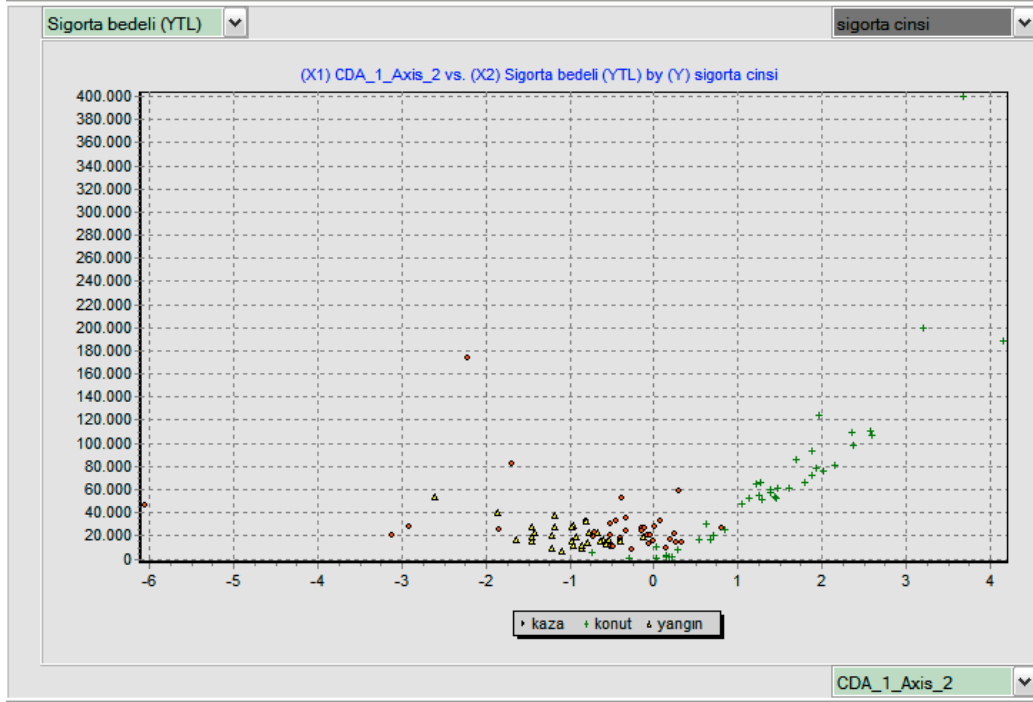
Yukarıdaki grafikteki y ekseninde sigorta bedeli değişkeni ve x eksenine de Z_1 vektörü yerleştirildiğinde grafik aşağıdaki şeklini almaktadır.



Şekil 3.7. Z_1 vektörü ile satın alma miktarı arasındaki ilişkiyi gösteren grafik

Z_1 vektörü üzerindeki dağılımına bakıldığında kaza sigortalarının iyi , konut sigortalarının normal ve de yangın sigortaları kötü gruba girmektedir.

Yukarıdaki grafikteki y ekseninde sigorta bedeli değişkeni ve x eksenine de Z_2 vektörü yerleştirildiğinde grafik aşağıdaki şeklini almaktadır.



Şekil 3.8. Z_2 vektörü ile satın alma miktarı arasındaki ilişkiyi gösteren grafik

Z_2 vektörü üzerindeki dağılımına bakıldığında kaza ve yangın sigortaları normale yaklaşmış iken konut sigortalarında ise bir değişiklik olmamıştır.

111 adet x sigorta şirketinin verilerine yapılan sepet analizinin kadın müşterilerin çoğunluğunun konut sigortalarını tercih ederken erkek müşteriler ise kaza sigortalarını ağırlıklı olarak tercih ettikleri görülmektedir. Türkiye genelinde erkek ve kadın statülerinin yukarıda yapılan analiz sonuçlarını etkilediği yani cinsiyetin tercih edilen sigorta poliçesinin üzerinde yüksek bir etkiye sahip olduğu söylenebilmektedir. Yapılan karar ağacı analiz sonuçlarına bakıldığında 341 YTL primden daha az prim ödeyen müşterilerin konut sigortasını tercih ettikleri ortaya çıkmıştır. K-ortalamlar analizinin sonucunda oluşturulan veri

kümelerinin tamamına yakını % 40 anlamlılık düzeyini aşmıştır.Kanonikel diskiriminant analizi sonucunda ise özellikle konut sigortası verilerini azaltarak veriyi daha anlamlı bir hale getirebileceği ortaya çıkmıştır.

SONUÇ

Bu çalışmada , son yıllarda yaygınlaşan veri madenciliği ve müşteri ilişkileri yönetimi detaylı bir şekilde incelenmiştir.Müşteri ilişkileri yönetimi aslında bir işletme felsefesi olup daha çok satış kampanyalarının başarılı olması için kullanılan bir tekniktir.Verit madenciliği ise satış kampanyalarında kullanılacak olan hedef ya da hedef gruplarının tespit edilmesinde kullanılmaktadır.

Sigorta şirketlerinin gerek müşteri portföylerini arttırmak ve de varolan müşteri portföyünü korumak için müşteri ilişkileri yönetimine ihtiyaç duymaktadır.Yoğun rekabet ortamında poliçe satın alan müşterilerin devamlılığını sağlamak , kar marjını arttırmak için , özetle şirketlerin hayatta kalabilmesi için uygulanması gerek ve şart bir sistemdir.Özellikle verilen poliçelerin müşteriye özel bir hale dönüştürülmesinin faydaları ve de sağladığı kar marjının yüksek olacağı görülebilmektedir.Mal , eşya satan bir firmadan hizmet satan sigorta şirketleri müşteri ilişkileri yönetiminin uygulamalarına çok daha müsaittir.

Veri madenciliği özellikle mevcut müşteri bilgilerini özetleyerek veya verileri arındırarak verileri daha kullanışlı hale getirilmesine olanak sağlamaktadır.İşletmelerin bilgi bombardımanından kurtularak verileri amaçlarına uygun kullanma imkanı vermektedir.

Sigortacılık sektörüne bakıldığında ise gerek veri madenciliği gerekse müşteri ilişkileri yönetimi yeni yeni yaygınlaşmaya başlamıştır.Özellikle banka bağlantılı olan sigorta şirketleri , müşteri ilişkileri yönetimi için pazarlama departmanına bağlı bölümler oluşturmaktadırlar.Bu oluşturulan bölümlerin amacı satış ve satış sonrası destekte müşteri ilişkileri yönetimini ve de veri madenciliğini kullanmaktır.

Gerek müşteri ilişkileri yönetimi (CRM) gerekse veri madenciliği uygulanabilmesi için belirli bir düzeyde teknoloji desteğine ihtiyaç duymaktadırlar. Veri değerleri ne kadar fazla olursa teknolojiye olan ihtiyaç da o kadar artmaktadır.

Veri madenciliğinden elde edilen sonuçların şirketlerin müşterilerini daha iyi tanımalarına olanak sağladığı için müşteri ihtiyaçları daha kesin bir şekilde tespit edilebilmektedir.

Bu çalışmanın incelenen müşteri ilişkileri yönetimi (CRM) sonucunda müşteri ilişkileri yönetiminin bir felsefe olduğu , veri tabanı ve veri analizine bağlı olarak belirli bir uygulama süreci sonucunda elde edildiği , elde edilen değerler kullanarak müşterilere özel ürün oluşturma süreci olduğu ortaya çıkmıştır.

Bu çalışmanın ayrıca veri madenciliği incelenmiştir. Veri madenciliğinin çok teknik bir konu olduğu , uygulamanın yapılabilmesi için teknoloji desteğe ihtiyaç duyulduğu ortaya çıkmıştır. Veri madenciliği , genel olarak istatistik analiz yöntemlerinin çok sayıda veri kümesine uygulanması sonucunda müşteri ilişkileri yönetimi için hedef gruplar oluşturmayı amaçladığı görülebilmektedir.

Bu çalışmanın uygulamasında ise x sigorta şirketinin müşteri verileri kullanılarak sigorta sektöründe veri madenciliği kullanımının sonuçları değerlendirilmiştir. Veri setine uygulanan analizlerin ana başlıklarına bakıldığında bunlar ; birliktelik kuralları analizi , sınıflandırma analizleri ve de kümeleme analizleridir.

Birliktelik kuralları analizi sonucunda erkek müşterilerin çoğunluğun kaza sigortalarını tercih ederken kadın müşteriler ise konut sigortasını tercih ettikleri görülmüştür.

Yapılan sınıflandırma analizi sonucunda ise , 341 YTL değerinden küçük prim ödeyenlerin kaza sigortalarını tercih ettikleri , 341 YTL prim değerinden fazla ödeme yapan ve sigorta bedelleri 82 650 YTL'den büyük olan bir mali sigortalayan müşterilerin çoğunlu konut sigortasını tercih etmiştir. Ayrıca 341 ila

728 YTL arasında prim ödeyip sigorta bedeli 82 650 YTL'den az olan bir malı sigortalayan müşteri ise yangın sigortasını tercih etmektedir.

Kümeleme analizi sonucunda 5 adet küme oluşturulmuştur.En çok veriye sahip olan küme % 51 anlamlılık derecesiyle 46 veriye sahiptir.En anlamlı küme ise % 69 ile 18 veriye sahiptir.

Bu çalışmada görüldüğü üzere veri madenciliği teknikleri müşterileri sınıflandırmak , kümelemek ve de davranış olasılıklarını tahmin etmekte oldukça güçlü analizleri içermektedir.Yapılan çalışmada özellikle bankacılık sektöründe yaygın olarak kullanılan müşteri ilişkileri yönetimi ve veri madenciliği teknikleri sigortacılık sektöründe de uygulanmaya çalışılmıştır.Ürün-müşteri , şirket-müşteri arasındaki ilişkileri ortaya konularak müşterilerin tercihlerine göre poliçe satışında artış sağlanmaya çalışılmıştır.

KAYNAKÇA

1. http://www.microsoft.com/turkiye/dynamics/crm/crm_nedir.msp
2. <http://www.biyomed.com/pages/makaleler/makale49.htm>
3. http://www.apluspst.com/bilgi_teknoloji.php
4. http://www.kobifinans.com.tr/bilgi_merkezi/020305/14318
5. www.tepum.com.tr/Etkinlikler/crm_nedir.pps
6. 65.110.73.19/UploadsNew/Gallery/Presentations/CRM/CRM-distribution.pps
7. http://www.ixirteknoloji.com.tr/crm_nedir.html
8. 65.110.73.19/UploadsNew/Gallery/Seminars/ICT3/III_CRM_Awareness-Handouts.pdf
9. www.srdc.metu.edu.tr/webpage/documents/Kosgeb/CRM_KOSGEB.ppt
10. <http://www.sauemk.com/makale.html>
11. http://www.erpcrm.com/crm_anasf/crm_nedir.htm
12. http://blog.inspark.com/blog/2006/10/mteri_liklileri.html
13. <http://www.herkesmusteri.com/scrm/crmbasarisiz.aspx>
14. A. PAYNE , Handbook Of CRM: Achieving Excellence in Customer Management , Butterworth-Heinemann Publishers 2005.
15. P. Bligh, D. Turk , CRM Unplugged Releasing CRM's Strategic Value , Wiley Publishers 2004.
16. O.C.Gel , CRM Yolculuğu ,Sistem Yayıncılık 2004.
17. A.KIRIM ,Strateji Ve Birebir Pazarlama CRM , Sistem Yayıncılık , 2007.
18. Y.ZENGİN, Değer Yaratan Müşteri İlişkileri Yönetimi ,Yüksek Lisans Tezi, Marmara Üniv. ,2006
19. R.S. SWIFT ,Accelerating Customer Relationships ,Prentice Hall , 2001.
20. T.T.BİLGİN , Veri Madenciliğinde Kavramı Ve Analiz Yöntemi Uygulamaları , Yüksek Lisans Tezi, Marmara Üniv. ,2003 .

21. M. DEMİRALAY ,Hiyerarşik Kümeleme Metotları İle Veri Madenciliği , Yüksek Lisans Tezi, Marmara Üniv. ,2005.
22. H.TATLIDİL , Uygulamalı Çok Değişkenli İstatistiksel Analiz , Hacettepe üniv. 1996.
23. K.KURTULUŞ ,Pazarlama Alıştırmaları ,Avcıol Yayını ,1998.
24. H.ARICI , İstatiksel Yöntemler ve Uygulamaları , Hacettepe üniv. Yayınları 1998.
25. Ö.SERPER ,Uygulamalı istatistik 1 , Filiz Kitapbevi 1996.
26. K.ÖZDAMAR , SPSS ile Bioistatistik , Kaan Kitabevi 2001.
27. D.GUJARATI , Temel Ekonometri , Literatür Yayınevi 1999.
28. K.ÖZDAMAR , Paket Programlar ile İstatiksel Veri Analizi , Kaan Kitabevi 1999.
29. N.ORHUNBİLGE , Uygulamalı Regresyon ve Kolerasyon Analiz , İÜ. Yayınları 1996.
30. N.ÇÖMLEKÇİ , Deney Tasarımı ve Çözümlemesi , Anadolu üniv. Yayınları 1988.
31. M.AYTAÇ , Matematiksel İstatistik , Ezgi Kitabevi 1999.
32. M.SPIEGEL, İstatistik , Bilim Tekik Yayınevi 1995.
33. Ö.SERPER ,Uygulamalı istatistik 2 , Ezgi Kitapbevi 2000.
34. F.NEWEELL , CRM Neden Başarılı Olmuyor, Sistem Yayınları 2004.
35. E.TAŞKIN , Müşteri İlişkileri Eğitimi ,Papatya Yayınları 2005.
36. Y.ODABAŞ , Satışta ve Pazarlamada Müşteri İlişki Yönetimi , Sistem Yayınları 2005.
37. O.GEL , Büyük Müşteri Yönetimi , Sistem Yayınları 2004.
38. W.CHU, Foundations and Advances in Data Mining ,Springer Publisherss 2005.
39. D.RUAN , Intelligent Data Mining Techniques and Applications ,Springer Publisherss 2005.
40. K.ANDERSON , Customer Relationship Management , McGRaw-Hill Publisherss 2002.
41. F. TOURNIAIRE , Just Enough CRM , Prentice Hall PTR Publisherss 2003.
42. C .TODMAN , Designing a Data Warehouse: Supporting Customer Relationship Management , Prentice Hall PTR Publisherss 2000.
43. N. WOODCOCK , The Customer Management Scorecard: Managing CRM for Profit, Kogan Page Publisherss 2003.

44. M. KANTARDZIC ,Data Mining Concepts, Models, Methods, and Algorithms , John Wiley & Sons Publishers 2003 .
45. O.RUD , Data Mining Cookbook , Wiley Computer Publishers 2001.
46. S.MITRA , Data Mining Multimedia , Soft Computing and Bioinformatics , Wiley Publishers 2003 .
47. N.YE , The Handbook Of Data Mining , Lawrence Erlbaum Associates Publishers 2003 .
48. D . HAND , Principles Of Data Mining , The MIT Press 2001.
49. J.WANG , Data Mining Opportunities and Challenges , IDEA GROUP Publishers 2003 .
50. M .BERRY , Data Mining Techniques , Wiley Publishers 2004 .
51. T. JOHNSON , Exploratory Data Mining And Data Cleaning ,Wiley Publishers 2003
52. J. BIGUS , Data Mining With Neural Networks ,McGraw Hill Publishers , 1996 .
53. D.PYLE , Business Modeling And Data Mining , Morgan Kaufmann Publishers 2003.
54. F.GUILLET , Quality Measures in Data Mining , Springer Publishers 2007.
55. D.LAROSE , Data Mining Methods and Models , Wiley Publishers 2006 .
56. L.LOFTIS, Building The Customer-Centric Enterprise , Wiley Publishers 2001.
57. A.Berson , Building Data Mining Application For CRM ,McGraw Hill Publishers 1999.
58. http://www.bendevar.com/v3/makale_326.html
59. http://www.kobifinans.com.tr/yazici.php?Article=8652&Where=bilgi_merkezi&Category=020305
60. http://www.init.com.tr/news_articles_tr.asp?haber_id=12
61. R.S.TSAY ,Analysis Of Financial Time Series ,Wiley publishers 2005.
62. B.KEDEM ,Regression Models For Time Series Analysis , Wiley Publishers 2002.
63. W.Hardle , Multivariate Statistics:Exercises And Solutions ,Springer Publishers 2007.
64. N.H.TIMM ,Applied Multivariate Analysis, Springer Publishers 2002.
65. L.L.HARLOW , The Essence Of Multivariate Thinking : Basic Themes And Methods, LEA Publishers 2005 .