Veri Madenciliği

ARGE, 1991 yılından bu yana müşterilerine **yönetim danışmanlığı** hizmetleri sunmaktadır. Bugün gelişmiş metodolojileri, çeşitli sektörlerde elde ettiği deneyimleri ve güçlü danışman ekibi ile **ARGE** Türkiye'nin önde gelen kuruluşları arasında yer almaktadır.

Rakamlar, **ARGE**'nin yeni işlerinin %60'ının eski müşterilerinden geldiğini göstermektedir. Bu istatistik, müşteri mutluluğunun bir göstergesi olarak önem verdiğimiz bir performans göstergesidir.

Gücünü, müşterilerinin artan performanslarından alan **ARGE**'nin müşteri profilinde kendi sektörlerinde önder firmalar bulunmaktadır. Bizce içinde bulunulan konum ne kadar iyi olursa olsun, bununla yetinmemek ve daha iyiye ulaşmak için gayret göstermek, lider olmanın çok önemli ve içgüdüsel bir özelliğidir.

En iyiyi başarmak için hiç vazgeçmediğimiz şu temel prensipleri uygulamaktayız:

- Müşterilerimizin işlerine artı değer katmak,
- Yaratıcı yaklaşımlar ve uygulanabilir çözümler üretmek,
- Üstlendiğimiz her işi en iyi şekilde gerçekleştirmek için hiçbir fedakârlıktan kaçınmamak,
- Müşterilerimizin gizliliğine daima özen göstermek.

Hizmetlerimizi yapılandırırken en önde tuttuğumuz unsur "uygulama"dır. Bu yaklaşım verilen hizmetlerin sadece öneri ve raporlar ile kısıtlı olmasından çok farklı olup gerçek bir katma değer yaratmaktadır.

ARGE'nin görüşüne göre, iş performansını doğrudan etkileyen dört temel unsur bulunmaktadır. Bunlar iş stratejisi, bu stratejiyi gerçekleştirmek için gerekli olan iş yapma yöntemleri, insan kaynakları ve teknoloji yönetimidir. Değer yönetimi alanında özel bir metodoloji ile çalışan **ARGE**, çalışma başında müşterisinin şirketinin değerini ölçer ve çalışmalarını bu değeri yükseltmeye odaklar.

Bu çerçevede **ARGE**'nin sunduğu danışmanlık hizmetleri dört ana başlık altında toplanmaktadır: *Strateji, Yönetimde Kalite, Kurumsallaşma, Geleceği Şekillendirme.*

Bu dört ana başlık altında verilen hizmetler şu şekildedir:

Strateji	Yönetimde Kalite	Kurumsallaşma	Geleceği Şekillendirme
Strateji Geliştirme	İş Etkinliği Değerlendirmesi	İnsan Kaynakları Yönetim Sistemleri	Toplumsal Katkıyı Yapılandırma
Strateji Uygulama (Balanced Scorecard)	Toplam Kalite Yönetimi	Kurumsal Yönetişim	İşbirlikleri Geliştirme
Stratejik İşbirlikleri ve Birleşme sonrası Yapılanma	Süreç Verimliliği	Aile Şirketlerinde Kurumsallaşma	Toplumsal Yönetişim
Senaryo Planlama	Değer Yönetimi	Entelektüel Sermaye Yönetimi	STK Etkinliği Geliştirilme
Ülke Stratejileri	Yeniden Yapılanma	İnsan Kaynakları SistemDeğerlenidrmesi	Sosyal Destek Projeleri

ARGE danışmanları, uzmanlık konularında dünyadaki gelişmeleri takip etmek için senede bir ay eğitim alırlar. Sosyal sorumluluğunun bilincinde olan bir kurum olarak, çalışanlarının haftada bir gün gönüllü kuruluşlarda çalışarak deneyimlerini toplumsal sorunların çözümünde kullanmalarını teşvik eder.

ARGE, Avrupa Parlamentosu'nda kurumsal **sosyal sorumluluk** projeleriyle geleceği şekillendiren **en iyi üç şirket** arasında değerlendirilmiştir (2002).

ARGE

www.arge.com

"Başarınız için yanınızdayız"

Veri Madenciliği

ARGE

Veri Madenciliği

- Veri Madenciliği
  Veriden Bilgiye, Masraftan Değere
  Dr. Yılmaz Argüden

- Veri Madenciliği
  Burak Erşahin

ARGE

VERİ MADENCİLİĞİ

Veriden Bilgiye, Masraftan Değere

Dr. Yılmaz ARGÜDEN
Burak ERŞAHİN

ARGE Danışmanlık

Bu kitap kurumlarda verinin kullanımını artırarak, yönetim kalitesinin geliştirilmesine destek olmak üzere ARGE Danışmanlık A.Ş. tarafından hazırlanmış ve Alkim Kağıt San. ve Tic. A.Ş.'nin destekleriyle yayınlanmıştır

Veri Madenciliği başlıklı kitap KalDer tarafından organize edilen 17. Ulusal Kalite Kongresi katılımcılarına ARGE Danışmanlık'ın hediyesi olarak sunulmaktadır.

ARGE Danışmanlık Yayınları No: 10
Veri Madenciliği
Veriden Bilgiye,Masraftan Değere

Yazan ve Derleyen
Dr. Yılmaz ARGÜDEN
Burak ERŞAHİN

ISBN: 978-975-93641-9-9
1. Basım Kasım 2008

ARGE Danışmanlık A.Ş.
Fazılkaptanoğlu Caddesi
Seba İş Merkezi No: 3 Kat: 9
Seyrantepe, 34418, İSTANBUL
Tel: (0212) 283 59 60
Faks: (0212) 283 59 64
www.arge.com

Bu kitap 100 gr Alkim mat kuşe kağıda basılmıştır

İçindekiler

ÖNSÖZ

Bilgi güçtür. Bilgiyi üretebilen, kullanabilen bireyler, şirketler ve toplumlar daha hızlı gelişir, gelirlerini, kârlılıklarını ve refah düzeylerini artırırlar. Bilgi çağında değer yaratmanın yolu fiziksel varlıklardan çok, bilgi kaynaklarını etkin kullanmaktan geçiyor. Bu nedenle, bilgi yönetimi için birçok yöntem ve araç geliştiriliyor.

Günümüzde gelişen bilgi teknolojileri sayesinde her geçen gün daha çok veri sayısal olarak toplanıyor, saklanıyor ve hepsinden önemlisi kullanılıyor. Veri bilgiye dönüştürülüp, kullanılıncaya kadar değer ifade etmez. Değerli olan verilerin irdelenip, bilgiye dönüştürülmesi ve karar için kullanılabilmesidir.

Büyük miktarda verinin çeşitli yöntemler ile analiz edilmesi ve çıkan sonuçların bir uzman gözüyle yorumlanmasıyla geçmiş verilerden gelecek tahminleri yapmaya yarayacak bilgiler edinilmesi işlemine veri madenciliği deniyor. Ülkemizde henüz emekleme aşamasında olan veri madenciliği konusundaki yetkinliklerimizi artırmanın, kurumlarımızı doğru kararlara yöneltmesi ve başarı şanslarını artırması açısından önemli olduğuna inanıyoruz.

Yönetim kalitesini geliştirme misyonuyla ve yarattığı katma değer ve toplumsal katkılarıyla örnek olma vizyonuyla çalışan ARGE Danışmanlık, her sene yeni bir yönetim aracını ve/veya anlayışını ülkeye kazandırmaya çalışıyor.

ARGE Danışmanlık kuruluşundan bu yana yönetim bilimi konusundaki deneyimlerini gerek iyi örnekler oluşturulmasına destek olarak, gerekse 'Balanced Scorecard', 'Kurumsal Sosyal Sorumluluk', 'Entelektüel Sermaye', 'Değer Yönetimi', 'Değişim Yönetimi', 'İtibar Yönetimi', 'Kurumsal Yönetişim' gibi yayınlarla Türk yönetim dünyasıyla paylaşarak yönetim alanında birçok yeniliğin Türkiye'de uygulanmasına da öncülük yapıyor.

Bu kapsamda her sene yeni bir kitapçığı Ulusal Kalite Kongresi'nde Türk yönetim dünyasının kullanımına sunuyoruz. Bu sene de veri madenciliği konusunu Türk yönetim dünyasının gündemine taşıyarak, bu konudaki deneyimlerimizi "Veri Madenciliği: Veriden, Bilgiye – Masraftan, Değere" başlıklı kitap aracılığıyla sizlerle paylaşmaktan memnuniyet duyuyoruz.

Ülkemizdeki kurumların bilgiye dayalı kararlar ile her geçen gün daha yüksek değer yaratması dileğimizle....

ARGE Danışmanlık A.Ş.

VERİ MADENCİLİĞİ: Veriden Bilgiye, Masraftan Değere

Dr. Yılmaz ARGÜDEN

İnsanoğlu yaşam kalitesini geliştirmek için bilgi ve bilimi kullanabilmesiyle kendisini ayırt ediyor. İnsan topluluklarında da bilgi ve bilimi daha etkin olarak üretebilen ve kullanabilenler göreceli olarak daha yüksek yaşam standartlarına kavuşabiliyorlar. Bu nedenle, ülkelerin bilgi düzeyi, bilginin mülkiyet hakları, analiz yetkinliklerini geliştiren eğitim sistemleri, ve bilgiye dayalı karar verme kültürüne yakınlıkları, gelişmişlik düzeyleri açısından belirleyici oluyor.

Teknolojik gelişmeler dünyada gerçekleşen bir çok işlemin elektronik olarak kayıt altına alınmasını, bu kayıtların kolayca saklanabilmesini ve gerektiğinde erişilebilmesini hem kolaylaştırıyor, hem de bu işlemlerin her geçen gün daha ucuza mal edilmesini sağlıyor. Ancak, ilişkisel veri tabanlarında saklanan birçok veriden kararlar için anlamlı çıkarımlar yapabilmek bu verilerin bilinçli uzmanlarca analiz edilmesini gerektiriyor. Üstelik veri miktarı arttıkça bunların analiz edilmesi de özel araçlar ve yöntemlerin kullanımını zorunlu hale getiriyor.

Veriyi hızlı toplayan ve bilgiye dönüştürerek hızlı kullananlar rekabetçi avantaj elde ederler. Veri madenciliği büyük miktarda veri içinden gelecekle ilgili tahmin yapmamızı sağlayacak ilişki ve kuralların aranmasıdır.

Veri madenciliği, özel ve kamu sektörü kuruluşlarında birçok şekilde kullanılabilmektedir. Bunlardan bazıları aşağıdaki gibi sıralanabilir:

- Bir süpermarket müşterilerinin satın alım eğilimlerini irdeleyerek, promosyonlarını belli müşterilere yönlendirmesi, aynı kaynakla daha çok satış gerçekleştirmesine yardımcı olabilir.
- Bankalar kredi kararlarında kredi isteyenlerin özelliklerini ve davranışlarını irdeleyerek batık kredi oranını azaltabilir.
- Havayolları sürekli müşterilerinin davranış biçimlerini irdeleyerek daha etkin fiyatlandırma ile kârlılıklarını artırabilirler.
- Bir telefon şirketi müşteri davranışlarından öğrendikleri ile yeni hizmetler geliştirerek, müşteri bağlılığını ve kârlılığını artırabilir.
- Maliye Bakanlığı Gelir İdaresi, şirketler için risk modelleri kurarak vergi incelemelerini daha etkin yönlendirip, vergi kaçaklarını azaltabilir.
- Hastaların teşhis ve tedavi maliyetleri irdelenerek hastalık riskinin ilk aşamada tespiti, kontrolü ve kaynak planlama açısından faydalı olur.

Ancak, bu faydaları sağlayabilmek için veri madenciliği konusuna yatırım yapmak gerekiyor.

Büyük veri tabanlarının analiz odaklı olarak kullanılmasının önünde çeşitli engeller var: (i) Veri tabanlarındaki gözlemlerin birçoğunda bilgilerin eksik veya yanlış olması, (ii) Bazı verilerin kişisel değerlendirmelere dayandırılması ve bu nedenle gözlemler arasında tutarlı karşılaştırma yapılmasının güç olması, (iii) Veri toplama süreçlerinin bütünü resmetmeyi engelleyecek ve bilinçli olmayan seçicilikler içermesi (selection bias), (iv) Veri tabanı yapısının analiz odaklı olmaması, (v) Analizi mânalı hale getirecek

bilgi eksikliklerini tamamlamanın maliyetli (veya imkansız) olması, (vi) Yetkin analistlerin kullanılmaması, (vii) Analiz teknikleri konusunda bilgi sahibi olanlarla, irdelenecek karar hakkında bilgi sahibi olanlar arasındaki iletişimin sağlıklı modelleme yapacak düzeyde olmaması gibi...

Bu nedenle, büyük veri tabanlarından faydalanılarak bilgi üretme sürecinde dikkat edilmesi gereken unsurlar var. Öncelikle analizlerin güvenilir verilere dayandırılmasını, yapılacak herhangi bir analizin başkaları tarafından da tekrarlanabilir olmasını ve verinin cevaplandırabileceği sorulara odaklanılmasını sağlamak gerekiyor. Bu nedenle, veri madenciliği yapacak analistin ilk adımı veri tabanındaki verilerin hangi süreç ile ve nasıl toplandığını çok iyi anlamaktır. Bazı durumlarda çok bilgi var sanılırken, birçok gözlemde aynı bilginin kaydedilmiş olması, aslında bazı boyutlarda veri tabanının sığ olduğunu, bu boyutlardaki analizlerin çok az veriye dayandırılacağını gösterir. Yine verilerin toplanma sürecindeki istemsiz seçicilikler analizde ve daha da önemlisi analiz sonuçlarının nerelerde uygulanabilir olduğunu belirlemek açısından büyük önem taşır.

Ayrıca, analistlerin eğitim süreçlerinde genellikle temizlenmiş, örnek küçük veri tabanlarının kullanılması, onların büyük veri tabanlarını incelerken bazı önemli adımları atlamalarına neden olabilir. Bu nedenle, veri madenciliğinde ikinci adım veri tabanının içerdiği verileri iyi anlamaktır. Bunun için her gözlemde bulunan verilerin dağılımı, ilişkilendirilmesi planlanan boyutların örneklemesi ve veri tabanındaki gözlem birimi ile modellemeye temel oluşturacak gözlem birimi arasındaki farklılıkları giderme kuralları konusunda detaylı irdeleme yapılmalıdır. Bir veri tabanının iyi anlamak için yapılacak ilk yatırım, analizlerin ve sonuçların kullanılabilir olmasında büyük önem taşır.

Üçüncü önemli adım ise analiz sonuçlarının kullanılması planlanan kararları ve çevre şartlarını iyi anlamaktır. Çünkü veri tabanının hangi soruları yanıtlaması istendiğini anlamak doğru modelleme yapabilmenin temelidir. Korelasyon, neden-sonuç ilişkisi değildir. Doğru modelleme yapacak teorik bilgiyi edinmeksizin, sadece teknik olarak korelasyonlara dayandırılan çıkarımlar, analistlerin güvenilirliğini zedeleyebilir veya karar vericileri yanlış kararlara yönlendirebilir.

Dördüncü adım, teorik modeli kurmak ve veri tabanı kullanılarak test edilecek hipotezleri oluşturmaktır. Kurulan modelin, geçerliliğini test etmek, bu nedenle veri tabanının bir kısmını kullandıktan sonra model ile yapılacak öngörülerin veri tabanındaki diğer veriler kullanılarak doğrulanması, karar vericilerin modele duydukları güveni ve modelin gerçek hayatta kullanılma olasılığını artırır.

Özetle, veri madenciliği veriden bilgi üreterek, ortalama kararlar yerine özgün kararlar verilmesini destekleyen, satışları, kârlılığı, yenilikçiliği ve kaynak kullanımında etkinliği artıran önemli bir yönetim aracıdır.

Bu nedenle, veri madenciliği konusuna yapılacak yatırımları özendirmek, eğitim sistemimizin bu konuda yetkin-

likleri geliştirmesini sağlayacak adımları atmak ülkemizdeki kurumların başarısını artıracaktır. Bu şekilde yönetim kalitesini artıran kurumlar aynı zamanda toplumsal refah düzeyimizin artmasını ve toplumsal düzeninin korunmasını sağlarlar.

VERİ MADENCİLİĞİ

Burak ERŞAHİN

Giriş – Anlaşılabilirlik

Dünya ile ilgili en anlaşılmaz şey, herşeyin tamamen anlaşılabilir olmasıdır. - Albert Einstein

Veri madenciliği dünyanın anlaşılabilirliğine önemli ölçüde destek olan bir kavramdır. Gelişen bilgi toplama, depolama ve işleme yetkinlikleri, giderek artan bir şekilde mevcut verilerin incelenerek anlamlı sonuçlar elde edilmesine olanak sağlamaktadır.

Artık hangi genlerin hangi hastalığa neden olduğunu, hangi müşterilerin kredisini geri ödeyemeyeceğini, hangi koşullarda yağmur yağacağını, hangi filmin yüksek gişe hasılatı yapacağını veya müşterilerin bir sonraki alışverişlerinde hangi ürünleri alacaklarını bilebilmek çok şaşırtcı değildir. Asıl şaşırtıcı olan bu sonuçlara ulaşabilecek kurumların, verilerini topluyor olmalarına rağmen bu verileri anlamlı bilgilere dönüştürmüyor olmalarıdır.

Bu çalışma veri madenciliği konusunda yazılmış mevcut literatüre bir alternatif değildir. Çalışmanın temel amacı; konuya giriş yapmak, veriler ile neler yapılabildiğini aktarmak ve bireyleri karar alırken veri kullanmaları konusunda teşvik etmek ve böylelikle karar kalitesini artırmaktır.

Çalışmanın ana konusu olan üç temel kavramı şu şekilde tanımlayabiliriz.

- Veri; sayılar, metinler, sesler ve görüntülerdir.
- Veri tabanı; sistematik erişim imkanı olan, yönetilebilir, güncellenebilir, taşınabilir, aralarında tanımlı ilişkiler bulunabilen verilerdir.
- Bilgi: öğrenerek, deneyerek, araştırarak elde edilen, karar almaya yardımcı olan her türlü sonuçtur.

Veri Madenciliği Nedir?
Verinin bilgiye dönüştürülmesi

Veritabanlarında Bilgi Keşfi, Bilgi Çıkarımı, Veri Analizi, Veri Arkeolojisi, Bilgi Hasatı, İş Zekâsı, Data Mining, Information Harvesting, Knowledge Discovery in Databases, Data Pattern Processing, Database Mining, Data Archaeology, Knowledge Mining, Data Dredging, Knowledge Extraction, Siftware, Desen Algılama, Pattern Recognition

Veri madenciliği, büyük hacimli veri yığınları içerisinden karar alabilmek için potansiyel olarak faydalı olabilecek, uygulanabilir ve anlamlı bilgilerin çıkarılmasına verilen addır. Veri madenciliği geniş anlamda veri analiz teknikleri bütünüdür ve tek başına bir çözüm değildir. Mevcut problemleri çözmek, kritik kararları almak veya geleceğe yönelik tahminleri yapmak için gerekli olan bilgileri elde etmeye yarayan bir araçtır. Ortaya çıkarılması hedeflenen bilgiler; üstü kapalı, çok net olmayan, önceden bilinmeyen, daha önce keşfedilmemiş ancak potansiyel olarak kullanışlı anlamlı ve kritik bilgilerdir.

Veri madenciliği döngüsü, veri yığınlarını elden geçirmekle başlayarak, analiz sonucunda ortaya çıkan sonuçların uzman gözüyle yorumlanması ile tamamlanır. Veri madenciliği çalışmalarının

[1] The most incomprehensible thing about the world is that it is at all comprehensible

alt yapısının önemli bir bölümünü istatistik ve veritabanı uygulamaları oluşturmaktadır.

Veri madenciliği, büyük boyutlu veri ambarlarının meydana çıkmasının bir sonucudur. 1960'larda veriler elektronik ortamda toplanmaya ve geçmiş veriler bilgisayarlar ile analiz edilmeye başlanmıştır. 1980'lerde bağıntılı (relational) veritabanları ve SQL ile verilerin dinamik ve anlık analiz edilmesine olanak sağlanmıştır. 1990'lara gelindiğinde toplanmakta olan verinin hacmi çok büyük boyutlara ulaşmış ve verilerin depolanması için veri ambarları kullanılmaya başlanmıştır. Veri madenciliği toplanan bu büyük veri kütlelerinin değerlendirilmesi için istatistik ve yapay zeka tekniklerinin kullanılması sonucunda ortaya çıkmıştır.

Teknolojik gelişmeler, ham verilerin yeni fırsatlar üretmek üzere yönetim ve pazar ihtiyaçlarına yanıt verecek bilgiye dönüştürülmesini kolaylaştırmış ve bir anlamda kurumları veri madenciliği üzerinde çalışmaya mecbur bırakmıştır.

- Ölçüm cihazlarının çeşitlenmesi ve otomatik veri toplama araçlarının gelişmesi sonucunda toplanan verilerin türleri ve sayısı artmıştır.
- Veritabanları ve veritabanı teknolojisinin gelişmesi sonucunda veri depolarında çok miktarda verinin depolanması sağlanmıştır.
- Bilgisayar ve veri işleme teknolojisinin gelişmesi sonucunda toplanan verilerin hızlı biçimde çözümlenmesine olanak sağlanmıştır.

Veri madenciliği; veritabanı teknolojisi, makine öğrenmesi, desen tanıma, istatistik, görselleştirme gibi birçok farklı disiplinden yararlanmaktadır.

Neden Veri Madenciliği

Veri ile değil, bilgi ile çalışma avantajı

Her alanda, verilen kararların doğruluğu, kararı veren kişinin yeteneklerine ve deneyimine olduğu kadar sahip olduğu bilginin yeterliliğine de bağlıdır. Bu nedenle artık "bilgi", mal ve hizmetin yanında üçüncü üretim faktörü olarak değerlendirilmektedir. Bilginin yeterli olması, bilgiyi oluşturan verilerin doğru depolanması, doğru işlenmesi ve doğru yorumlanmasına bağlıdır. Buna ek olarak karar vericiler doğru kararları alabilmek için mümkün olduğunca çok veriyi depolamaya çaba göstermektedirler.

Ancak verilerin toplanması, bir oyuna giriş bileti almaktan çok farklı değildir. Asıl zorluk devamlı çoğalmakta olan ham veriyi, anlamlı ve kullanılabilir bir bilgiye dönüştürebilmektir. Bilgi sistemleri birçok açık olmayan ve geleneksel yöntemlerle anlaşılamayan bilgileri içermektedir.

Veri madenciliği, özellikle kar ve pazar payı elde edebilmek için yoğun rekabetin yaşandığı pazarlama alanında ön plana çıkmaktadır. Hangi müşteri, hangi ürünü, ne zaman satın alabilir, kimler tedarikçilerinden vazgeçmekte ve bu tür müşterileri vazgeçirmek / geri kazanmak için neler yapılabilir, ürünün değerini yitirmesine hangi değişkenler neden olmaktadır, vb.

[2] Wikipedia

soruların cevapları veri yığınlarının altındadır ve cevapları bulabilmek için veri madenciliği çözümleri gereklidir.

Veri Madenciliği ile şirketler önceden bilinmeyen bilgileri ortaya çıkararak karar verme süreçlerini iyileştirirler. Veri madenciliği teknikleri kullanarak; maliyetleri azaltmak, gelirleri artırmak, verimliliği artırmak, yeni fırsatları ortaya çıkarmak, yeni keşifler yapmak, emek yoğun faaliyetleri otomatikleştirmek, sahtekarlıkları belirlemek ve müşteri deneyimini geliştirmek mümkündür.

Özetle, veri madenciliği iki gereksinimden ortaya çıkmaktadır.

- Toplanan çok miktarda verinin işlenme ihtiyacı
- Artan rekabette doğru karar verebilme yetkinliğini artırmak ihtiyacı

Veri Madenciliği Ne Değildir?

İdeal durumda tüm kurumlar faaliyetleri sonucunda elde ettikleri verileri değerlendirerek, kullanılabilir sonuçlar elde etmeyi hedeflemelidirler. Ancak uygulamalara baktığımızda kurumların önemli bir kısmının verileri toplamanın ötesine geçmedikleri gözlenmektedir. Gelişim çizgisine bakıldığında verilerin toplanması (ve doğru şekilde toplanması) başlangıç noktasıdır. Elde edilen verilerden yapılacak sorgulamalar ve detaylı analizler ile elde edilen sonuçları veri madenciliği olarak değerlendirmemek gereklidir. Bir ölçüde bunlar da veri madenciliğidir ancak daha doğru tanımı veri düzenlemeciliği olarak adlandırılabilir.

Veri madenciliği; veri toplamak, mevcut verilerden sorgulamalar yapmak veya gelişmiş analiz teknikleri kullanmanın ötesinde bir noktadır.

- Bir restoran zincirinde; hangi şubelerin ne kadar ciro yaptığı, hangi ürünlerin hangi noktalarda daha fazla satıldığı, hangi saatlerde yoğunluk yaşandığı, gibi analizler veya
- Bir satış şirketinde; hangi müşterilerin devamlılık gösterdikleri, hangi bölgelerde performans düşüklüğü yaşadıklarını belirlemek veri madenciliği değildir.
- Gelir ile yaş ilişkisinin incelendiği bir değişken, bir sonuç ve az sayıda veriden oluşan bir modeli tanımlayarak, yaşa göre gelir tahmini yapmak da veri madenciliği değildir. Yüz değişkenin olduğu, değişkenler arasında sadece rakamsal değerlerin değil, sıralı (örnek: yüksek-orta-düşük) veya sırasız (örnek: evli-bekar-dul) kategorilerin olduğu, milyon tane verinin olduğu ancak doğru algoritmalar ve güçlü bir bilgisayar ile sonuca ulaşmanın mümkün olduğu modelleri kurmak veri madenciliğidir. Algoritmalar yukarıdaki örnekteki lineer regresyondan daha karmaşık olmakla birlikte, kavram aynıdır, mevcut verileri kullanarak tahmin veya tanımlama yapmak.

Veri Madenciliğine Örnek

Bir banka müşterilerine yeni bir ürün sunacaktır. Bu ürün çağrı merkezinden yapılacak müşteri aramaları ile telefon üzerinden sunulacaktır. Çağrı merkezi ile yapılan anlaşma gereği her bir müşteriye ulaşmanın bedeli 3 YTL olarak tanımlan-

mıştır. Ürünün satılması halinde bankanın elde etmeyi beklediği kar ortalama 100 YTL'dir. Bankanın müşteri portföyünde 2 milyon müşteri bulunmaktadır. Daha önceki satış kampanyalarından müşterilerin %2,5 oranında olumlu yanıt verdikleri gözlenmiştir.

Bu veriler doğrultusunda değerlendirme yapıldığında ürün sunma kararını almak mümkün değildir çünkü kampanyadan 1 milyon YTL zarar edilmesi öngörülmektedir.

Bu noktada veri madenciliği teknikleri ile bir değerlendirme yapıldığında kapmayanın yapılması kararı alınmıştır.

Müşterilerin %1'lik kısmı için bir deneme kampanyası yapılmıştır. Bu kampanyada 2.000.000 müşterinin %1'i olan 20.000 müşteriye ulaşılmış ve 500 müşteriden olumlu yanıt alınmıştır. Bu deneme kampanyasında elde edilen veriler, istatistiksel teknikler ve müşterilerin bilgileri birlikte kullanılarak gerçek kampanya için kullanılabilecek bazı önemli sonuçlara ulaşılmıştır.

Bu değerlendirme sonucunda müşterilerin sadece %50'si ile temas kurarak, ürünü kabul edeceklerin %70'ine ulaşmak veya müşterilerin sadece %40'ı ile temas kurarak, ürünü kabul edeceklerin %60'ına ulaşmanın mümkün olacağı sonucuna ulaşılmıştır.

Gelir	**Müşteri Sayısı x Yanıt Oranı x Beklenen Kar**	**2.000.000 x 0,025 x 100 = 5.000.000 YTL**
Maliyet	**Müşteri Sayısı x Birim Ulaşım Maliyeti**	**2.000.000 x 3 = 6.000.000 YTL**
Kar	**Gelir – Maliyet**	**5.000.000 – 6.000.000 = -1.000.000 YTL**

Bu durumda;

Gelir	**1.980.000 x %2,5 x %70 x 100 = 3.465.000 YTL**	**1.980.000 x %2,5 x %60 x 100 = 2.970.000 YTL**
Maliyet	**1.980.000 x %50 x 3 = 2.970.000 YTL**	**1.980.000 x %40 x 3 = 2.376.000 YTL**
Kar	**3.465.000 – 2.970.000 = 495.000 YTL**	**2.970.000 – 2.376.000 = 594.000 YTL**

Veri Madencisi Kimdir?

Analiz ve iş bilgisinin bir araya gelmesi

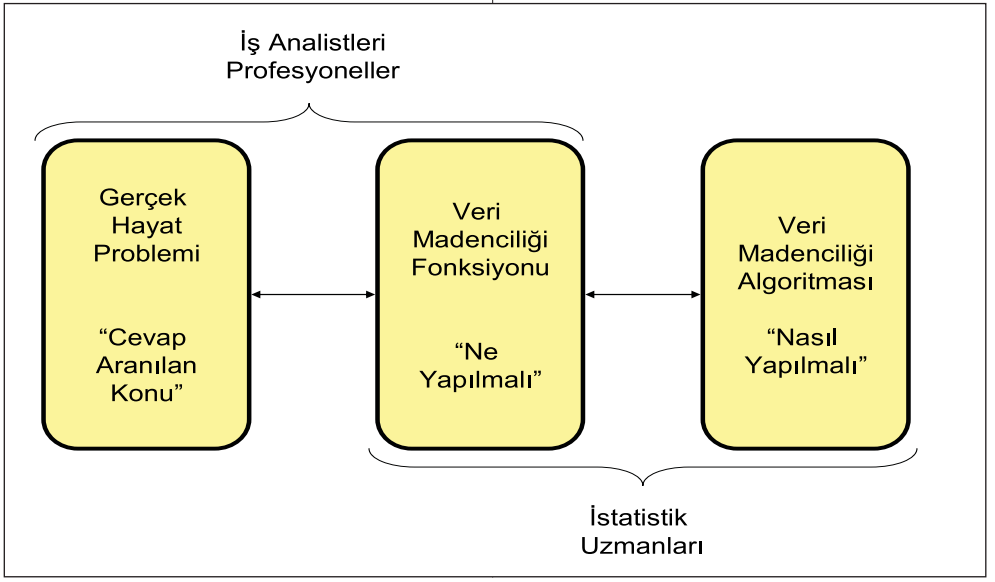
Cevap aranılan soru veya çözülecek problem için kurulan bir modelin başarılı olabilmesi sadece metodolojilerin derinlemesine biliniyor olmasına bağlı değildir. Veriyi ve pazarı tanımak, kurumun iş hedeflerini biliyor olmak, modelin altyapısını oluşturan metodolojilerden çok daha önemlidir.

Her alanda olduğu gibi veri madenciliğinde de teknoloji ile deneyimin birleşimi en doğru sonuca ulaştırmaktadır. Deneyimin elde edilen sonuçlar üzerindeki etkisi oldukça yüksektir.

Veri madenciliği bilincinin artması ile birlikte, bu tür çalışmalara ağırlık vermek isteyen şirketlerin büyük bölümü iki önemli hata yapmaktadırlar.

- Çalışmaları gerçekleştirmek için teknik konulara hakim istatistik uzmanları veya teknik analistleri işe alarak, modelleri kurgulamalarını istemek: Bu kategorideki uzmanlar teknik konularda çok yetkin olmalarına rağmen, gerekli iş kavrayışına yeterince sahip olmamaları nedeniyle arzu edilen sonuçlara çoğunlukla ulaşılamamaktadır.
- Sofistike veri madenciliği yazılımları satın almak: Konu ile ilgili çok detaylı, tüm metodolojileri içeren yazılımlar mevcuttur ancak yazılımlardan faydalı sonuçlar alabilmek için doğru modeli kurgulamak ve doğru girdileri sunmak gereklidir. Bu düşünce sürecinden geçmeden yazılımdan faydalı sonuçlar elde etmek mümkün değildir.

Her iki yaklaşımda da; hedefi oluşturma, veriyi elde etme, veriyi hazırlama, modeli uygulama, sonuçları değerlendirme gibi önemli alanlarda bilgi eksikliği söz konusu olabilir. Bu alanların herhangi birinde yapılacak hata



çok maliyetli olabileceği gibi tamamen yanlış sonuçlara da götürebilir.

İstatistiksel araçları çok iyi bilen en iyi teknik analistlere sahip olmak kadar bunu gerçek dünyanın problemlerine nasıl uyarlayacaklarını bilmek de önemlidir. Bu aşamada veri madenciliğinin 3 farklı boyutuna bakmakta, ilişkileri ve gereksinimleri anlamak açısından fayda vardır.

- Yanıtlanacak soru nedir? / Neye cevap aranmaktadır?
- Cevap aranan konuyu hangi veri madenciliği fonksiyonu ile çözümlemek gerekir?
- İlgili veri madenciliği fonksiyonu için hangi algoritma ile model oluşturmak uygun olur?

Cevap aranılan sorunun tanımlanması ve uygun fonksiyonun seçilmesi aşamasında faaliyetten sorumlu olan profesyonellerin daha etkin rol alması, seçilen fonksiyona uygun algoritmanın belirlenmesi ve işletilmesi aşamasında istatistik uzmanlarının daha etkin rol alması gerekir.

Veri Madenciliği Süreci (Döngüsü)

Veri madenciliği süreci en basit şekli ile dört adımdan oluşmaktadır.

1. **Veri Seçmek:** Mevcut olan ve elde edilebilecek verilerin oluşturduğu veri havuzundan çalışma için kullanılacak verilerin seçilmesi
2. **Veri İşlemek:** Farklı kaynaklardan alınan verilerin birleştirilmesi, hatalı verilerin çıkarılması, vb. ile seçilmiş verilerin kullanılabilir hale getirilmesi
3. **Veri Eğilimlerini / Desenlerini Belirlemek:** İşlenmiş verilerin veri madenciliği fonksiyonları ve algoritmaları ile değerlendirilerek verilerden anlamlı eğilimlerin, desenlerin (pattern) çıkarılması
4. **Bilgiye Ulaşmak:** Verilerden çıkarılan anlamlı eğilimler ve desenlerin yorumlanarak bilgi elde edilmesi

Daha detaylı süreç tanımı ise veri madenciliğinin uluslararası düzeyde standardı olarak kabul edilmiş, CRISP-DM (CRoss Industry Process for Data Mining) ile yapılmaktadır. Veri madenciliği projelerinin hızlı, daha verimli ve daha az maliyetli gerçekleştirilmesi için geliştirilmiş olan bu süreç altı adımdan oluşmaktadır.

1. **İşi ve İş Ortamını Anlama (Business Understanding):**
İlk adım veri madenciliği çalışmasının hangi amaç için yapılacağının net olarak tanımlanmasıdır. Amaç; cevap aranılan sorunun üzerine odaklanmalı, net biçimde ifade edilmeli ve sonuç değerlendirme kriterleri tanımlanmalıdır. Çalışma sonunda doğru cevaplanmış birçok yanlış soru elde edilmek istenmiyorsa, çalışmanın cevap aranılan soru ile uyumlu olması güvence altına alınmalıdır.

a. İş Hedeflerini Algılamak: Çalışmanın temel amacının belirlenmesi ve bu amacın mümkün olduğunca ikincil amaçlardan ayrıştırılarak net olarak tanımlanması aşamasıdır. Çalışma sonuçlarının değerlendirme kriterlerinin belirlenmesi de

bu aşamada yapılması gereken diğer bir konudur.

b. *Durumu Değerlendirmek:* Veri madenciliğinin temel amacı verim artırmaktır. Bu amaç elde edilecek sonuçlar kadar sürecin kendisi için de geçerlidir. Çalışma sonucunda elde edilecek faydayı değerlendirmek (yanlış kararların maliyetleri ve doğru kararların getirilerine ilişkin öngörüler) önemli bir gerekliliktir. Bu aşamada çalışma için gerekli kaynaklar, tahmini maliyet, mevcut kısıtlar, olası riskler, vb. değerlendirilerek elde edilecek faydanın boyutu ile karşılaştırılır.

2. Veriyi Anlama (Data Understanding): İkinci adım ilk verilerin toplanması, mevcut verilerin uygunluğunun değerlendirilmesi, modeli oluşturmak için gerekli farklı veri ihtiyaçlarının tespit edilmesi, sahip olunan kayıt sayısının yeterliliği gibi veri kalite ve yeterliliğine yönelik düşünce sürecinden geçilmesi aşamasıdır. Hedef çalışmada kullanılacak verilere aşinalık kazanmaktır. Veriyi anlamak ile işi anlamak iç içe geçmiş alt süreçlerdir. İşi anladıkça farklı verilere bakmak veya verilerin gösterdiklerini anlamak, verilere baktıkça iş ile ilgili farklı bakış açıları kazanmak mümkündür. Bu döngü kendi içinde devam ettikçe çalışmada kullanılacak verilerin netlik kazanır.

a. *Başlangıç Verilerini Toplamak:* Proje kaynaklarında tanımlanmış olan başlangıç verilerinin toplanması aşamasıdır.

b. *Veriyi Tanımlamak:* Toplanan verinin tanımlanması ve ihtiyaçları karşılama yeterliliğinin değerlendirilmesi aşamasıdır.

c. *Veriyi Keşfetmek:* Başlangıç aşamasında toplanan veriler ile başlangıç hipotezlerinin oluşturulması, limitli bir şekilde veriden çıkarımlar yapılması aşamasıdır. Bu aşamada sonuca yönelik bilgilerin elde edilmesinden daha çok çalışmanın gerçekleştirilebilmesi için veri anlamında eksikliklerin tespit edilmesi amaçlanır.

d. *Verinin Kalitesini Belirlemek:* Veri tam mı, doğru mu, hatalar içeriyor mu, hatalar içeriyorsa ne tür hatalar içeriyor, veride eksik bölümler var mı şeklindeki sorular ile verinin kalitesinin tespit edilmesi aşamasıdır.

3. Veri Hazırlama (Data Preparation): Bu aşama başlangıç verilerinin, çalışmalara temel oluşturacak final verilere dönüştürülmesi aşamasıdır. Bu çalışmanın adımlarının belirgin bir sırası veya tekrar sayısı yoktur. Modelin kurulması aşamasında ortaya çıkacak sorunlar, bu aşamaya sık sık geri dönülmesine ve verilerin yeniden düzenlenmesine neden olabilmektedir. Bu aşama karar vericinin veri keşfi sürecinin toplamı içerisindeki enerji ve zamanının % 50'sinden fazlasını harcamasına neden olmaktadır.

[3] www.crisp-dm.org

a. *Veri Setini Tanımlamak:* Modelin kurulacağı, tanımlanan soru için gerekli olduğu düşünülen veri setinin (veriler ve bu verilerin toplanacağı veri kaynakları) belirlenmesi aşamasıdır. Bazı durumlarda kurum içinde büyük boyutlarda veri toplanıyor olmasına karşın, toplanmakta olan veriler tanımlanan soru için yeterli olmayıp, başka veri gruplarından eklemeler yapmak gerekebilir. Müşterilerinin yaptığı her işlemi kaydeden, hesaplarının düzeyini, kredi başvuru detaylarını tamamen bilen bir bankanın veri madenciliği çalışmaları için tüm verilere sahip olduğu düşünülebilir. Ancak yapılacak bir pazarlama kampanyasında müşterilerin bireysel ilgi alanları gibi demografik göstergelere göre değerlendirme yapılmak isteniyorsa bu verileri bankanın veri tabanlarındaki operasyonel verilerden sağlamak mümkün değildir. Çalışmalar öncesinde bu veriler doğrudan müşteriler ile temas kurularak ya da bir veri sağlayıcısından temin edilerek tamamlanmalıdır.

b. *Veriyi Seçmek:* Yapılacak analizde kullanılacak verilerin belirlenmesi aşamasıdır. Değerlendirme sırasında verinin hedefler ile ilişkisine, kalitesine, teknik limitlere dikkat etmek gerekir. Verilerin değişken sayısı kadar kullanılan kayıt sayısı da önemlidir. Gereğinden az veri, çalışmayı eksik bırakabileceği gibi, gereğinden fazla veri, veri kirliliğine ve sürecin uzamasına neden olabilecektir.

c. *Veriyi Temizlemek:* Gürültülü ve tutarsız verileri çıkararak verinin kalitesini artırma aşamasıdır. Yanlış girişden veya istisnalardan kaynaklanan verilerin, değerlendirmeden çıkarılması tercih edilir. Bazı durumlarda çok büyük veri tabanı ile çalışmak yerine örnekleme yapılması uygun olabilir (seçilen örneklerin tüm popülasyonu temsil düzeyi önemlidir.) Verileri temizlemek sadece çıkarmak gibi düşünülmemelidir, bazı durumlarda eksik verilerin tamamlamak için modellemeler yapmak da söz konusudur.

d. *Veriyi Kurmak:* Verileri tanımlayan mevcut değişkenlerde modifikasyonlar yaparak model için daha kullanılır değişken setleri oluşturma aşamasıdır. Örneğin müşterilere yapılan satışların aylık olarak kaydedildiği bir yapıda, müşterilerin ortalama sipariş verme sıklığını tanımlamak, mevcut değişkenlerden yeni bir değişken tanımlamaktır. Herhangi bir ayda alım yapmamış müşterilerin listede yer almaması analizlerde yanlış sonuçlar doğurabileceğinden alım yapmayan müşteriler için sıfır miktarlı kayıtlar eklemek yine bu aşamada sık yapılan uygulamalardan biridir.

e. *Veri Birleştirmek:* Veri madenciliğinde kullanılacak verilerin farklı kaynaklardan toplanması, doğal olarak veri uyumsuzluklarına neden olmaktadır. Bu uyumsuzlukların başlıcaları farklı zamanlara ait olmaları, güncelleme hataları, veri formatlarının

farklı olması, kodlama farklılıkları (örneğin bir veri tabanında cinsiyet özelliğinin e/k, diğer bir veri tabanında 0/1 olarak kodlanması), farklı ölçü birimleri ve varsayım farklılıklarıdır. Bu adımda farklı kaynaklardan toplanan verilerin uyumsuzluklar mümkün olduğu ölçüde giderilerek, tek bir veri tabanında toplanması amaçlanır. Ancak burada çok dikkatli ve titiz davranmak gereklidir. Dikkatin en çok gerektiği aşamalardandır. Bu aşamada yapılacak bir hata, ileriki aşamalarda daha büyük sorunlar yaratacaktır. Doğru sonuç alınacak veri madenciliği çalışmaları ancak doğru verilerin üzerine kurulabileceği için, toplanan verilerin ne ölçüde uyumlu oldukları bu adımda incelenerek değerlendirilmelidir.

f. *Veri Formatlamak:* Veri seti oluşturulduktan sonra kullanılacak modele göre anlam değişikliği yapmayacak format düzenlemelerinin yapılması aşamasıdır.

4. Modelleme (Modeling): Benzer veri madenciliği problemleri için birden çok çözüm tekniği olabilmektedir. Bazı teknikler verilerde spesifik ihtiyaçlar duyarlar. Bu nedenle veri hazırlama ve model kurma aşamaları, en iyi olduğu düşünülen modele varılıncaya kadar yinelenen bir süreçtir. (Veri Madenciliği Fonksiyonları / Algoritmaları bölümünde daha detaylı incelenmiştir.)

a. *Model Tekniğini Seçmek:* Kullanılacak veri madenciliği fonksiyonun ve algoritmasının belirlenmesi aşamasıdır. (Veri Madenciliği Fonksiyonları / Algoritmaları bölümünde daha detaylı incelenmiştir.) Genel olarak verilerin oluşturulma aşamasından itibaren bu konuda bir öngörünün oluşmuş olması gerekir.

b. *Model Test Tasarımı Yapmak:* Modeli işletip sonuçları elde etmeye başlamadan önce, modelin kalitesini ve geçerliliğini test etmek gereklidir. Örneğin öngörü fonksiyonlarından sınıflandırma fonksiyonunda hata oranlarını kalite göstergesi olarak kullanılır. Veriler hazırlandıktan sonra, ilk aşamada verinin bir kısmı modelin öğrenilmesi, diğer kısmı ise modelin geçerliliğinin test edilmesi için ayrılır. Modelin öğrenilmesi, öğrenim kümesi kullanılarak gerçekleştirildikten sonra, test kümesi ile modelin doğruluk derecesi belirlenir. Modelde yanlış olarak sınıflanan olay sayısının, tüm olay sayısına bölünmesi ile hata oranı, doğru olarak sınıflanan olay sayısının tüm olay sayısına bölünmesi ile ise doğruluk oranı hesaplanır.

Sınırlı miktarda veriye sahip olunması durumunda, kullanılabilecek bir yöntem, çapraz geçerlilik testidir. Bu yöntemde veri kümesi rasgele iki eşit parçaya ayrılır. İlk aşamada bir parça üzerinde model eğitimi ve diğer parça üzerinde test işlemi; ikinci aşamada ise ikinci parça üzerinde model öğrenimi ve birinci parça üzerinde test işlemi yapılarak elde edilen hata oranlarının ortalaması kullanılır.

Model kuruluşu çalışmalarının sonucuna bağlı olarak, aynı teknikle farklı parametrelerin kullanıldığı veya

başka algoritma ve araçların denendiği değişik modeller kurulabilir. Model kuruluş çalışmalarına başlamadan önce, hangi tekniğin en uygun olduğuna karar verebilmek güçtür. Bu nedenle farklı modeller kurarak, doğruluk derecelerine göre en uygun modeli bulmak üzere denemeler yapılmasında yarar bulunmaktadır.

Önemli bir diğer değerlendirme kriteri modelin anlaşılabilirliğidir. Bazı uygulamalarda doğruluk oranlarındaki küçük artışlar çok önemli olsa da, bir çok kuruluş uygulamasında ilgili kararın niçin verildiğinin yorumlanabilmesi çok daha büyük önem taşıyabilir.

Kaldıraç oranı ve grafiği, bir modelin sağladığı faydanın değerlendirilmesinde kullanılan önemli bir yardımcıdır. Örneğin kredi kartını muhtemelen iade edecek müşterilerin belirlenmesi amacını taşıyan bir uygulamada, kullanılan modelin belirlediği 100 kişinin 35'i gerçekten bir süre sonra kredi kartını iade ediyorsa ve tesadüfi olarak seçilen 100 müşterinin aynı zaman diliminde sadece 5'i kredi kartını iade ediyorsa kaldıraç oranı 7 olarak hesaplanmaktadır.

Ancak kurulan modelin doğruluk derecesi ne denli yüksek olursa olsun, gerçek dünyayı tam anlamı ile modellediğini garanti edebilmek mümkün değildir.

c. *Modeli Kurmak:* Model için kullanılacak algoritmanın/yöntemin/tekniğin hazırlanan veri üzerinde çalıştırılması aşamasıdır. Kurulan ve geçerliliği kabul edilen model doğrudan bir uygulama olabileceği gibi, bir başka uygulamanın alt parçası olarak da kullanılabilir. Kurulan modeller risk analizi, kredi değerlendirme, dolandırıcılık tespiti gibi işletme uygulamalarında doğrudan kullanılabileceği gibi, promosyon planlaması simülasyonuna entegre edilebilirler.

d. *Modeli Değerlendirmek:* Başarı kriterleri, daha önceki tecrübeler ve test sonuçlarına göre modelin değerlendirilmesi aşamasıdır. Tüm projenin değerlendirilmesinden çok modelin teknik değerlendirilmesi amaçlanır.

5. Değerlendirme (Evaluation):

Bu aşamaya gelindiğinde kurulmuş bir model vardır. Bu aşama, modelin nihai olarak sunulmasından önce modeli yoğun olarak değerlendirilmesi ve iş hedefleri ile uyumlu olup olmadığının kontrol edilmesini amaçlar. Kapsanmamış (açıkta kalmış) konu olup olmadığı değerlendirilmelidir.

a. *Sonuçları Değerlendirmek:* Ön değerlendirme aşamaları modelin geçerliliği ve uygunluğu konusunda değerlendirme sunarken, bu aşamada modelin iş hedeflerini ne ölçüde karşıladığı değerlendirilir. Eğer zaman ve bütçe varsa gerçek veriler ile modelin test edilmesi tercih edilir. Modelleme sırasında ortaya çıkan ancak ana hedefler ile ilişkisi olmayan diğer ek faydaların da bu aşamada tanımlanması uygun olur.

b. *Süreci Değerlendirmek:* Kalite güvence aşamasıdır. Modelin iş hedeflerini karşılamaya yeterli olduğu kararını aldıktan sonra, modelin doğru kurulup kurulmadığı, sadece eldeki verilerden mi yararlanıldığı, gelecekte kullanılabilecek farklı verilerin neler olabileceği gibi konularda değerlendirmeler yapılmalıdır.

c. *Gelecek Adımları Planlamak:* Projenin geldiği noktanın yeterli olup olmadığı, ek çalışma gerekliliğinin değerlendirilmesi aşamasıdır. Atılacak başka adımlar nelerdir, bunu gerçekleştirebilecek bütçemiz var mı, eğer devam etmek mantıklı ise nereden devam etmeliyiz gibi konularda değerlendirmeler yapılmalıdır.

6. Yayma (Deployment): Modelin tamamlanmış olması projenin nihai sonucu değildir. Modelin amacı veriler hakkında bilinenleri artırmak dahi olsa, elde edilen veri kullanılacak biçimde organize edilmeli ve sunulmalıdır. Genellikle gerçek verilerden örneklerin sunulması şeklinde olur.

a. *Yayma Planını Oluşturmak:* Sonuçları değerlendirilerek, yayma stratejisinin oluşturulması aşamasıdır.

b. Takip ve Bakımı Planlamak: Zaman içerisinde bütün sistemlerin özelliklerinde ve dolayısıyla ürettikleri verilerde ortaya çıkan değişiklikler, kurulan modellerin sürekli olarak izlenmesini ve gerekiyorsa yeniden düzenlenmesini gerektirecektir.
Uzun süre yanlış veri kullanarak çalışmanın önüne geçmek için bakım çok önemlidir.

c. *Final Raporu Hazırlamak:* Yapılan çalışmanın başkaları tarafından da tekrarlanabilirliğini sağlamak ve sonuçlarını karar vericilere aktarabilmek üzere hazırlanan rapordur. Çalışmanın üçüncü taraflarca denetlenebilmesini ve güvenilirliğini sağlamak açısından önem taşır.

d. Projeyi Değerlendirmek: Yapılan çalışmaya dayandırılan kararların ve sonuçların belli bir zaman sonrasında beklentilerle karşılaştırılması ve gerektiğine çalışmanın yenilenmesi aşamalarını içerir.

Sektörel Uygulama Örnekleri

CNN'nin yayın akışının her dakikasının aldığı ratingden, IMDB web sayfasının kaç ziyaretçi aldığına, İstanbul'a eylül aylarında düşen yağmur miktarından, Markette en çok satılan süt markasına, Av Köpekleri Takım Yıldızı'nın (Canes Venatici) parlaklık düzeylerindeki değişimden, sedan otomobil satın alanların yaşlarına, pandaların doğurganlık oranından, Everest'e tırmanan dağcı sayısına kadar çevremizde yaşananların çok önemli bir kısmı sürekli olarak kaydedilmektedir.

Sayılardan oluşan tüm bu kayıtlar ancak doğru şekilde bakıldığında bir anlam ifade etmektedirler. Bu nedenle veri madenciliği en geniş anlamı ile yaşadığımız ve kaydettiğimiz olaylara anlam katmaktır.

Veri madenciliğinin uygulama alanlarını bilimsel ve iş dünyası olarak ikiye ayırmak mümkündür. Bilimsel çalışmalarda veri madenciliği kullanımının ardında yatan sebepler; gelişmiş veri toplama yöntemleri (otomatik istasyonlar, uydu ve uzaktan algılama sistemleri, teleskop taramaları, gen çözümlemeleri, vb.) ile işlenmek üzere ham olarak çok büyük boyutlarda veri toplanması, geleneksel tekniklerin ham verileri işlemede yetersiz kalması ve hipotezler oluşturma, sınıflandırma, karar alma gibi bilimsel çalışma adımlarında bilim insanlarına destek olmasıdır.

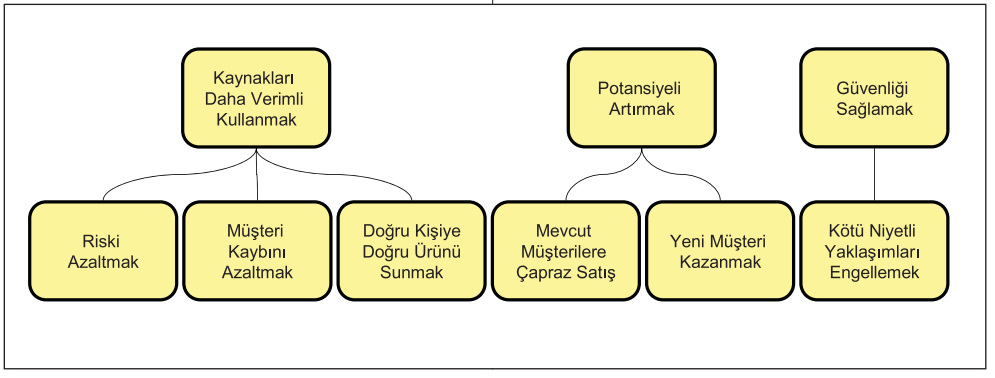
İş dünyasında veri madenciliği uygulamalarının kullanılmasının temel nedeni; müşteriyi tanıyarak (müşteri gibi düşünerek) müşteri memnuniyeti sağlamak ve bu şekilde rekabet ortamında hızlı ve doğru kararları alabilmektir.

İş dünyasında her alanda veri madenciliği uygulamalarını kullanmak mümkündür ancak sundukları ürün ve servislerle bilgiye dayalı yönetime en fazla ihtiyaç duyan sektörler ve alanlar; Finans (Bankacılık, Sigortacılık), Telekomünikasyon, Pazarlama ve Perakendedir.

İş dünyasında veri madenciliği çalışmalarının büyük bölümü üç temel ihtiyacı karşılamak için kullanılır. Bu ihtiyaçlar; "Kaynakları Daha Verimli Kullanmak", "Potansiyeli Artırmak" ve "Güvenliği Sağlamaktır".

Riski Azaltmak (Riski Önlemek)

Ürün veya hizmet sunumu sonrasında kayıp yaşama potansiyeli yüksek olan sektörlere özgü çalışmalardır. En



çok bankacılık ve sigortacılık sektörlerinde riskleri değerlendirmek için kullanılır. Bankalar müşterilerine kredi verdiklerinde bir finansal risk öngörürler, kurgulanan risk modelleri ile kredi alanların kredilerini geri ödeyememe ihtimallerini tahmin ederler. Krediye karşılık araba veya evin ipotek edildiği durumlarda risk daha düşük olmakla birlikte, kredi kartı bankalar açısından riski en yüksek kredi tipidir.

Sahtekarlık riski de bankalar için önemli bir konudur. Kredi kartı kaybolduğunda bankalar kaybedilme sırasında oluşan zararın bir kısmını üstlenmektedirler. Bu dönemde oluşan zararları azaltmak için sahtekarlığı tespit edici sistemler kurgulanmaktadır. Müşterilerin tipik harcama biçimlerini önceden tanımlayıp, harcama eğilimlerinde oluşan ani değişiklikleri tespit etmek ve bu doğrultuda satınalma işlemlerini onaylamayı durdurmak kullanılan yöntemlerden biridir.

Sigorta endüstrisinde risk müşterinin sigortalattığı üründe zararın oluşması ve müşterinin zararının karşılanmasını talep etmesidir. Tüm sigortacılık ürünlerinde fiyatlandırma için riskin doğru belirlenmesi gereklidir. Fiyatlar üzerinde düzenleyici kurulların etkisi nedeniyle, fiyatlandırma kârlılık açısından çok önem taşımaktadır.

Bankacılık ve sigortacılık gibi tahsilatını daha sonra yapmak üzere ürün ve hizmet sunan bir çok sektör risk taşımaktadır. Telekomünikasyon şirketleri, enerji şirketleri, perakendeciler, vb. ürünün veya hizmetin ödemesini yapmayacak olan müşteri riskini taşıyan benzer sektörlerdir.

- **Yüksek riskli müşteriler ile çalışmamak:** Firma için finansal kayıp oluşturabilecek müşterileri veya müşteri adaylarını belirleyerek bunlar ile çalışılmamasını sağlamak.
- **Kredi taleplerini değerlendirmek:** Mevcut müşterileri verilerinden kredi risk davranış modelleri oluşturarak, yeni başvurularda riskin en aza indirilmesini sağlamak.
- **Kredi geri ödemelerini kontrol altında tutmak (farklı risk politikaları oluşturmak):** Kredi kartı ödemelerini aksatan, gecikmeli olarak yapan veya hiç yapmayanların özelliklerinden yola çıkarak bundan sonra aynı duruma düşebilecek muhtemel kredi sahiplerini saptamak. Kötü ödeme performansı gösteren müşterilerin ortak özelliklerini belirleyerek, benzer özelliklere sahip tüm müşteriler için politikalar geliştirmek.

Müşteri Kaybını Azaltmak (Churn)

Müşterilerin, şirketin ürünlerini almaktan vazgeçerek rakip şirketin ürünlerini tercih etmeleri birçok endüstride giderek büyüyen bir sorundur. Müşterilerin bir firmadan diğer firmaya geçmesinin en önemli sebebi çoğunlukla daha iyi bir teklif almış olmalarıdır. Örneğin bankalar düşük faiz oranları ile rakiplerinin kredi kartı müşterilerinin kendi kredi kartlarını kullanmalarını sağlamaktadırlar. Kredi kartı kullanımı yaygınlaştıkça yeni kredi kartı müşterileri pazarı oldukça küçülmüştür. Bu da kredi kartı sunan bankaları yeni müşteri bula-

bilmek için rakiplerinin müşterilerini elde etmek zorunda bırakmaktadır. Bankalar müşterileri çekebilmek için kısa bir dönem için daha düşük kredi oranları sunmakta, oranlar normale döndüğünde müşterilerin bankada kalacaklarını düşünmektedirler.

Telekomünikasyon sektöründe en önemli sorun müşteri kaybıdır. Şirketler müşterilerinin rakiplerine geçmesini engellemek için çeşitli pazarlama taktikleri uygulamakta, ürünleri sürekli yeni sunular ile çeşitlendirmektedirler.

- **Mevcut müşteriyi elde tutmak:** Kuruluşlar hangi müşterilerini kaybedebileceklerini önceden belirleyebildikleri durumda, bu müşterilerini elde tutma amaçlı stratejiler geliştirebilirler. Kendi müşterisiyken rakibine giden müşterilerle ilgili analizler yaparak rakiplerini tercih eden müşterilerinin özelliklerini elde etmek mümkündür. Bundan yola çıkarak gelecek dönemlerde kaybetme olasılığı olan müşterilerin kimler olabileceği konusunda tahminlerde bulunarak onlara özgü ürünler ve servisler geliştirebilirler. Müşteriyi elde tutmak için yapılacak maliyet her zaman için müşteriyi geri kazanmak için yapılacak maliyetten düşük olacaktır.
- **Kaybedilen müşterileri yeniden kazanmak:** Eski müşterileri kazanmak için kurulmuş modellerdir. Müşterilerden ömür boyu elde edilecek getiri belirlenerek bu müşterilere sunulacakların maliyeti ile karşılaştırılır.

Doğru Kişiye Doğru Ürünü (Yaklaşımı) Sunmak

Mevcut ve potansiyel müşteriler hakkında detaylı bilgiye sahip olmak rekabetçi kalmak için önemli bir gerekliliktir. Farklı müşteri grupları için en uygun ürünleri bulmak, hangi tip müşterilerin ne tür ürün aldıklarını belirlemek, müşteri tabanını gruplara ayırmak, bu grupların karlılıklarını belirlemek ve buna göre farklı seviyelerde hizmet sunmak mümkün olabilir. Ürün veya hizmette hangi özelliklerin ne derecede müşteri memnuniyetini etkilediği, hangi özelliklerinden dolayı müşterin bunları tercih ettiği ortaya çıkarılabilir.

Ürün veya hizmet sunumuna kimin yanıt vereceğini tahmin etmek maliyet düşürmek açısından önemli yöntemlerden biridir. Bir ürün veya hizmet ile ilgili bir kampanya programı oluşturmak için hedef kitlenin seçiminden başlayarak bunun hedef kitleye hangi kanallardan sunulacağı kararına kadar olan süreçte veri madenciliği kullanılabilir. Aynı grubun geçmiş davranışlarına dayandırılabileceği gibi, mantıksal bir alternatif popülasyonun davranışlarına da dayandırılabilir.

- **Kampanya şartlarını düzenlemek:** Düzenlenecek çeşitli kampanyalarda mevcut müşteri kitlesinin seçmek ve bu müşterilerin davranış özelliklerine yönelik yaklaşımlar geliştirmek. Bu şekilde pazarlama veya perakende kampanyalarına cevap alma oranını artırmak, müşteri ilişkileri yönetimi maliyetlerinin azaltmak hedeflenir.

- **Özel kampanyalar düzenlemek:** Potansiyel müşteriler arasından en karlı olabilecekleri belirleyerek onlara özel kampanyalar uygulamak. En masraflı müşterileri daha masrafsız müşteriler haline dönüştürmek. Örneğin en çok bankacılık işlemi yapanlar ortaya çıkarılıp bunlar şube bankacılığı yerine internet bankacılığına yönlendirmek.

- **Müşterilere özgü satış politikaları oluşturmak:** Aynı karakteristikleri (gelir düzeyi, ilgi alanları, harcama alışkanlıkları, vb.) paylaşan "model" müşteri gruplarını bulmak ve satınalma profillerine göre satış şartlarını ve fiyatları belirlemek.

- **Yeni ürün geliştirmek:** Farklı müşteri gruplarının ihtiyaç duydukları özellikleri belirtip, ihtiyaç duymadıkları özellikleri üründen çıkararak müşterilerin beklentilerini karşılayacak şekilde farklı ürünler geliştirmek

Mevcut Müşterilere Çapraz Satış / Ek Satış (Cross-Sell / Up-Sell)

Kârlılığı artırmak için mevcut müşterilere satış yapmak, yeni müşteriler bulup onlara satış yapmaktan daha avantajlı bir yöntemdir. Bu nedenle; çapraz satış modelleri ile müşterilerin bir şirketten aldıkları ürünler dışında aynı şirketten ek ürün alma ihtimallerini, ek satış ile müşterilerin aynı ürünü tekrar veya daha çok alma olasılıklarını tahmin eden modeller oluşturulabilir.

- **Çapraz satış:** Kredi kartı harcamalarına göre müşteri gruplarının belirlenmesi, farklı finansal göstergeler arasında gizli korelasyonların bulunması, hangi müşteri profilinin neyi, ne zaman ve neden tercih ettiğini anlayabilen modeller ile ürün satışları arasındaki bağlantı ve ilişkileri bulmak ve bu bağlantılara dayalı tahminler geliştirerek ek ürünler sunmak. Çapraz satış ile birim müşteriye yapılan satış miktarının artırılması, karsız müşteriler karlı hale getirilmesi sağlanabilir. Perakende alanında pazar sepeti analizi ile birlikte satılan ürünlerin bulunması ve buna göre stratejilerin geliştirilmesi en yaygın uygulamalardandır.

- **Ek satış:** En iyi müşterileri veya müşteri gruplarını bulmak, bulunan bu müşteri gruplarının ihtiyaçları belirleyerek kişiselleştirilmiş ürün ve hizmetler geliştirmek, bu şekilde müşterilerin vazgeçemeyeceği ürün sunularını oluşturmak. Örneğin, yeniden sigorta poliçesi talep edecek müşterilerin tahmin edilmesi

Yeni Müşteri Kazanmak

Firmaların temel amaçlarından biri sürekli olarak yeni müşteriler kazanmaktır. Ancak tüm müşteriler eşit ölçüde kârlı değildir. Firmalar hedef kitlelerini ilk aşamada gelir, yaş, vb. gibi bazı temel ölçütlere göre seçerler ancak seçilen tüm potansiyeller kazanılsalar bile firmaya faydalı olmayabilirler. Firma ile uzun süre

çalışmayabilir, tüm ihtiyaçlarını firmadan almayabilir, sürekli olarak farklı teklifleri değerlendirerek başka firmalara geçebilir, alım sıklıkları düşük olabilir veya sadece düşük karlılığı olan ürünleri satın alıyor olabilirler. Tüm bu sebeplerle uzun vadede müşterilerden elde edilmek istenen sonuçlara ulaşılamayabilir. Müşterileri elde etmek için cazip teklifler sunmak yüksek müşteri edinme maliyeti yarattığından, doğru potansiyele sahip müşteriye odaklanmak, müşteriden ömür boyu elde edilecek değeri belirlemek önemlidir.

Doğru müşteriyi elde etmek için kritik yöntemlerden biri veri madenciliğidir. 3. taraflardan elde edilen başlangıç verilerinden segmentasyon ve sınıflandırmalar ile çeşitli müşteri segmentleri oluşturulabilir. Bu segmentlerden hangilerinin söz konusu ürün veya hizmeti alabileceği tespit edilir. Bu bilgiyi elde etmek için mevcut ve geçmişte alım yapmış müşterilerin alım bilgileri ve özelliklerinin yer aldığı bilgileri kullanılabilir. Alım yapma potansiyeli olan müşteriler belirlendikten sonra, hangi müşterilerin karlı olabileceği belirlenmelidir. Bu aşama da geçmiş verilere gereksinim duyar. Sınıflandırma yöntemleri ile karlı müşteriler belirlenerek ulaşılmaya çalışılır.

- **Müşterilerden ömür boyu elde edilecek getirileri belirlemek (Lifetime Value):** Müşterilerin firma ile ilişkileri boyunca yaratacakları katma değerleri tahmin etmek ve bu katma değer kategorilerine göre seçim yapmak, yaklaşım biçimleri belirlemek.
- **Kampanya şartlarını düzenlemek:** Düzenlenecek çeşitli kampanyalarda hedef müşteri kitlesinin seçmek ve bu müşterilerin davranış özelliklerine yönelik yaklaşımlar geliştirmek. Bu şekilde potansiyel müşterilere yapılan pazarlama veya perakende kampanyalarına cevap alma oranını artırmak, yeni müşteri edinme maliyetlerini azaltmak hedeflenir.

Standart dışı Davranışları Belirlemek / Güvenlik (Fraud Detection)

Para ile ilişkili tüm alanlarda sahtekarlık (fraud) riski vardır. Sağlık, finans sektörü ve vergilendirme en çok sahtekarlık denemesine rastlanan alanlardır. Genellikle geçmiş verilere göre sahtekarlık desenleri örnekleri çıkarılabilir ve bu örneklere benzer davranışlar tespit edilerek sahtekarlık denemelerinin önüne geçilebilir. Araştırılması gereken her olay maliyet yükü getirdiğinden veya verilen hizmeti aksattığından, sahtekarlık ihtimallerini belirlerken araştırılması gereken olay sayısını minimumda tutmak esastır. (İstatistik uygulamalarındaki 1. tür ve 2. tür hata tipleri bu konuda önemlidir.)

Kara para aklama, kayıp ve kaçakları engelleme, sigorta dolandırıcılıklarının tespiti, kredi kartı dolandırıcılıklarının tespiti, kaçak enerji kullananların profillerini tespit ederek olası kaçak enerji kullanıcılarını tahmin etmek, telefon görüşmelerinin (aranan yer, arama süresi, aranan zaman, vb.) modellenmesi ve

beklenen değerlerden sapmanın olduğu konuşmaların tespiti, sağlık sigortası uygulamalarında gereksiz veya birbiriyle ilgisiz sağlık testlerinin tespiti, vergi ile ilgili yolsuzlukları ve izlerini belirlemek, ağ saldırının tespit edilmesi vb. şeklinde farklı konularda veri madenciliği uygulamaları kullanılabilir.

Diğer Konular

Yukarıda belirtilen başlıklar veri madenciliği çalışmalarının genel olarak çözüm ürettiği konulardır. Ancak tüm sektörlerde, verinin olduğu her yerde, mutlaka bir veri madenciliği çalışması ihtiyacı oluşabilmektedir.

- Hastanelerde hastanın hastanede kalma süresinin azaltılması
- Hastanelerde hasta sonuçlarının iyileştirilmesi
- Hisse senetlerinde zaman serileri analizleri ile gelecek değerlerin belirlenmesi
- Devletin kurumlara vereceği destek programlarında verilecek desteğin doğru miktarda ve doğru hedefleri olan kuruluşlara verilmesinin sağlanması
- Emniyet birimlerinin hangi profildeki insanların ne tür suçlara eğilimi olduklarını belirleyerek, suç engelleme politikalarının oluşturulması
- E-ticaret / e-devlet uygulamalarında web sayfa tasarımını en iyi kullanılabilir hale getirilmesi
- Karayollarının belirli yollardaki kaza oranlarını düşürülmesini sağlaması
- Üniversitelere eğitimi tamamlayacak ve başarılı olabilecek öğrencilerin alınması
- Telekom şirketlerinde ağ performansının yönetimi
- Arşivde belirli bir dokümana benzer dokümanların bulunması
- …

"Arşivimde (veya internet üzerinde) bu dokümana benzer hangi dokümanlar var?" gibi soruların yanıtlarını bulacak şekilde "Doküman Madenciliği" (Text Mining) konusu da veri madenciliğinin bir alt kolu olarak gelişmektedir. Bu çalışmalarda amaç dokümanlar arasında ayrıca elle bir ayrım gerekmeden benzerlik hesaplayabilmektir. Bu genelde otomatik olarak çıkarılan anahtar sözcüklerin tekrar sayısı ile yapılır. Metin Madenciliğinin; haber, email, yazılı dökümanlar, arşiv ve internet üzerinde uygulamaları vardır.

Veri Madenciliğinin Fonksiyonları

Tahmin / Öngörü (Supervised) Fonksiyonlar

Geçmiş verilerden yararlanarak, gelecek ile ilgili bir sonucu tahmin etmek için kullanılan fonksiyonlardır. Yeni bir nesnenin niteliklerini inceleme ve bu nesneyi önceden tanımlanmış bir sınıfa atamaktır. Modellemelerinde olası sonucu öngörmeye yarayan faktörler ve sonuç yer alır. Model kurulurken geçmiş deneyimlerde, faktörlerin aldığı değerlere göre elde edilen sonuçlar girdi olarak kullanılır. Beklenen sonuç; "Katılır-Katılmaz" şeklinde kategorik değer veya rakamsal değerdir.

Tahmin edilen sonuçların kalitesi (ne kadar iyi tahmin edildiği) tahmin edilen sonuç kadar önemlidir. Çoğunlukla tahmin edilen sonuç ile birlikte, bu sonucun kalitesine yönelik; güvenlik aralığı, olasılığı, vb. değerleri belirlenir.

Sınıflandırma (Classification)

"Genç kadınlar küçük araba satın alır, yaşlı, zengin erkekler büyük, lüks araba satın alır."

En temel veri madenciliği fonksiyonlarından biri olarak kategorik sonuçları tahmin etmek için kullanılır. Modeli kurabilmek için, sonuçları önceden bilinen durumlar ve bu durumlarda ilgili faktörlerin aldığı değerler gereklidir. Bu değerler "eğitim verisi" olarak adlandırılır. Elde edilmesi beklenen sonuç "müşteri %80 ihtimal ile bu kampanyaya olumlu yanıt verecek" şeklinde belirli bir olasılık ile birlikte sunulur. Sonuçlar "Hizmeti Bırakır-Hizmeti Bırakmaz" şeklinde iki alternatifli olabileceği gibi "Kesin Tercih Eder-Tercih Eder-Yanıt Vermez-Tercih Etmez-Kesinlikle Tercih Etmez" şeklinde çoklu alternatifli de olabilir. Bir deneme kümesi modelin doğruluğunu belirlemek için kullanılır. Genellikle verilen veri kümesi öğrenme ve deneme kümesi olarak ikiye ayrılır. Öğrenme kümesi modeli oluşturulmasında, deneme kümesi modelin doğrulanmasında kullanılır. Örneğin bir otomobil satıcısı şirket geçmiş müşteri hareketlerinin analizi ile yukarıdaki gibi iki kural bulursa genç kadınların okuduğu bir dergiye reklam verirken küçük modelinin reklamını verir.

- **Uygulama Alanları:** Potansiyel müşteriler için düzenlenen kampanyalara dönüşler, mevcut müşterilerin belirli bir hizmeti almaktan vazgeçme olasılıkları, kredi başvurularının risk seviyeleri, çeşitli belirtilere göre hastalık ihtimalleri, vb.
- **Örnek Model:** Satışlarını artırmak için kampanya düzenlemek isteyen bir otomobil firması, kampanyasına katılma ihtimali olan potansiyel alıcıları belirlemek için daha önceden satış yapmış olduğu müşterilerinin verilerini (sonuçlarını) kullanarak, hangi özelliklere sahip adayların kampanyaya katılabileceğini belirli bir olasılık aralığında tahmin edebilir. Bu şekilde; ihtiyacı kadar veri satın alarak (eğer adayların verisini dışarıdan alıyorsa) ve sadece alma potansiyeli yüksek olan adaylara ulaşmaya çalışarak tasarruf sağlamaktadır.

  Aşağıdaki örnekte adayın gelir düzeyi, mesleği, yaşı, çocuk sayısı, kullandığı mevcut aracın modeli, sınıfı,

Durumlar	Girdi Faktörleri							Sonuç
	Mevcut Aracın Markası	Mevcut Aracın Sınıfı	Mevcut Aracın Yaşı	Çocuk Sayısı	Gelir Düzeyi	Yaşı	Mesleği	Kampanyaya Yanıt (Evet – Hayır)
Aday 001	Ford	B	6	2	40.000	60	Emekli	Hayır
Aday 002	Renault	B	2	1	120.000	40	Serbest Meslek	Hayır
Aday 003		A	5	0	60.000	35	Muhasebe Uz.	Evet

yaşı, gibi faktörler göz önüne alınarak bir model tasarlanmıştır.

- **Yöntemler / Algoritmalar:** Yapay Sinir Ağları (Neural Networks), Bayes Sınıflandırması (Bayesian Classification), En Yakın Komşu (Nearest Neighbour), Karar Destek Makineleri (Support Vector Machines), Zaman Serisi Analizi (Time Series Analysis), Karar Ağaçları (Decision Trees), Lojistik Regresyon (Logistic Regression)

Regresyon / Eğri Uydurma (Regression)

"Ev sahibi olan, evli, aynı iş yerinde beş yıldan fazladır çalışan, geçmiş kredilerinde geç ödemesi bir ayı geçmemiş bir erkeğin kredi skoru 825'dir."

Süreklilik gösteren değerleri tahmin etmek için kullanılan fonksiyonlardır. Regresyon ile amaç girdiler ile çıktıyı ilişkilendirecek modeli oluşturup, en iyi tahmine ulaşmaktır. Sonuç "bağımlı değişken", girdiler "bağımsız değişken" olarak adlandırılır. Sonucun alacağı değer genellikle bir güvenlik aralığı içinde belirtilir. Girdiler, çözülecek probleme göre bir veya birden fazla olabilir. Örneğin; bir inşaat firması konut satışlarının, faaliyet gösterdiği bölgede elde edilen toplam gelir ile ilişkili olduğunu düşünüyorsa, sadece bölgesel gelire dayalı bir model oluşturarak, bölgesel gelirdeki değişime göre satacağı ev sayısını tahmin etme yoluna gidebilir. Ancak gerçek hayatta çözülecek problemlerin hemen hepsinde doğru tahmine ulaşmak için birden fazla girdiden faydalanmak gereklidir. Bu noktada önemli olan konu girdilerin sonucun doğru tahmin edilmesine yaptıkları katkıdır. Bazı durumlarda sonuca katkısı limitli olan girdileri modelden çıkarmak, daha etkin bir model oluşturmak için önemli bir gerekliliktir.

- **Uygulama Alanları:** Finansal tahminler, zaman serisi tahminleri, biomedikal ve ilaç reaksiyonları, konut fiyatı değerlendirmeleri, müşterinin yaşam çevrimi boyunca yarattığı değer, atmosferdeki CO2 oranı, vb.
- **Örnek Model:** Bir dergiye ilk kez reklam vermeye başlayacak olan bir şirket daha önce reklam vermiş olduğu dergilerin sayfa maliyetlerini

Durumlar	Girdi Faktörleri			Sonuç
	Okuyucu Sayısı	Bayan Okur Payı	Ortalama Yıllık Gelir (YTL)	Sayfa Maliyeti (YTL)
Cosmopolitan	24.000	% 70	100.000	10.000
Capital	20.000	% 30	50.000	20.000
Esquire	9.000	% 5	45.000	5.000

kullanarak, çalışılmaya başlanılacak olan derginin vermiş olduğu fiyatın uygunluk seviyesini belirli bir güven aralığı içinde değerlendirebilir. Yada daha sonra yapacağı kampanyalarda çalışmakta olduğu dergilerin verecekleri fiyatların ne kadar makul olduğunu önceden öngörebilir.

Aşağıdaki örnekte derginin okuyucu sayısı, bayan okuyucuların payı, okuyucuların ortalama yıllık kazancı, gibi faktörler göz önüne alınarak bir model tasarlanmıştır.

- **Yöntemler / Algoritmalar:** Yapay Sinir Ağları (Neural Networks), Karar Destek Makineleri (Support Vector Machines), Karar Ağaçları (Decision Trees), Lineer Regresyon (Linear Regression)

Tanımlama (Unsupervised) Fonksiyonlar

Fonksiyonların amacı belirli bir hedefi tahmin etmek değildir. Amaç veri setinde yer alan veriler arasındaki ilişkileri, bağlantıları ve davranışları bulmaktır. Var olan verileri yorumlayarak davranış biçimleri ile ilgili tespitler yapmayı ve bu davranış biçimini gösteren alt veri setlerinin özelliklerini tanımlamayı hedefler. Tanımı bilmek; tekrarlanan bir faaliyette veya tanımı bilinen yeni bir verinin yapıya katılmasında ne şekilde hareket edileceği konusunda karar almaya destek olur.

Kümeleme/Gruplama/ Demetleme/ Öbekleme (Clustering)

Müşterilerin büyük bir kısmı düzenli olarak pazartesi akşamları kredi kartıyla alışveriş yaparlar.

Veriyi birbirlerine benzeyen elemanlardan oluşan sınıflara (kümelere) ayırarak, heterojen bir veri grubundan, homojen alt veri grupları elde edilmesi işlemidir.

Kümeleme fonksiyonu genellikle bölümleme sorunlarını çözmekte kullanılır. Kümelemenin temel hedefleri arasında; geniş veri yığınları için tanımlayıcı veriler belirleyerek, işlenecek veri hacmini daraltmak, veri yığınlarındaki doğal kümeleri ortaya çıkarak aynı kümede olması gereken verileri belirlemek, belirlenmiş kümelerin dışında kalan istisna durumları

tanımlamak sayılabilir. Başlangıç aşamasında verilerin hangi kümelere ayrılacağı veya kümelemenin hangi değişken özelliklerine göre yapılacağı bilinmemekte, konunun uzmanı olan bir kişi tarafından kümelerin neler olacağı tahmin edilmektedir.

Kümeleme algoritmaları; küme içinde benzerliğin maksimize (küme içi uzaklıkların minimize edilmesi) edilmesi, kümeler arası benzerliğin minimize (kümeler arası uzaklıkların maksimize edilmesi) edilmesi kavramına dayanır. Sonuçta elde edilen farklı kümelere ait elemanlar arasında benzerlik azdır.

Kümeleme fonksiyonu ile sınıflandırma fonksiyonu arasındaki en önemli fark, kümelemenin önceden tanımlanmış girdilere dayanmıyor olmasıdır. Sınıflandırma fonksiyonunda tanımlı girdiler ve bunların geçmişte aldıkları değerler temel modeli oluştururken, kümeleme fonksiyonunda önceden tanımlanmış girdiler ve örnekler yoktur. Veriler kendi içlerindeki benzerliklere göre gruplanırlar. Benzerliği tanımlayacak boyutlar ve özellikler modeli kuran tarafından öngörülür.

Kümeleme fonksiyonu bazı durumlar başka bir veri madenciliği fonksiyonun öncesinde kullanılabilir. Hangi promosyon kampanyasına müşteriler en iyi tepkiyi verirler diye değerlendirmek yerine öncelikli olarak müşterilerin belirli kümelere yarılması bunun ardından her küme için en iyi promosyon kampanyasının ne olacağı belirlenebilir.

Müşterileri kümelemek için genellikle karlılık ve pazar potansiyeli boyutları kullanılır. Perakende sektöründe müşterilerin; söz konusu firmadaki alım alışkanlıkları ve tüm mağazalardaki alım alışkanlıklarına göre kümelenmeleri ve en yüksek potansiyelli kümeye odaklanılması sıkça rastlanan bir uygulamadır.

- **Uygulama Alanları:** Benzer hücreleri tanımlamak, benzer davranışlar gösteren perakende müşterilerini tanımlamak, gen ve protein analizleri, ürün gruplaması, hastalık belirtileri, metin madenciliği
- **Örnek Model:** İki boyutlu bir örnekte kümeleme fonksiyonunu algılamak oldukça kolaydır. Yaş ve gelir düzeyleri belirtilmiş 40 kişiden oluşan bir grubu, grafik yardımı ile kümelerine ayırmak mümkündür. Yaş ve gelir düzeyi değerlerinin histograma yerleştirilmesi ve en yoğun durumların merkez olarak belirlenmesi en basit anlamda bir kümeleme işlemidir. Bu örnekte veri madenciliği yöntemleri kullanılmadan kümeler oluşturulmuştur. Ancak onlarca değişken olduğunda verileri kolayca kümelemek mümkün değildir, bu aşamada kümeleme fonksiyonuna özgü algoritmaları kullanmak gereklidir.
- **Yöntemler / Algoritmalar:** Bölme yöntemleri (Partitioning methods), Hiyerarşik yöntemler (Hierarchical methods), Yoğunluk tabanlı yöntemler (Density-based methods), Grid tabanlı yöntemler (Grid-based methods), Model tabanlı yöntemler (Model-based methods)

Birliktelik Analizi / Bağıntı / Eşleme / İlişki Kuralları (Association Rules)

"Çocuk bezi alan müşterilerin 30%'u süt de alır."

Büyük veri kümeleri içinde farklı veriler arasındaki birliktelik ilişkilerini bulma işlemidir. Birliktelik analizi, belirli bir veri kümesinde yüksek sıklıkta birlikte görülen özellik değerlerine ait ilişkisel kuralların keşfidir. Sonuçta elde edilen birliktelik kuralları (A → B) şeklinde sunulur. şirketlerin karar alma işlemlerini daha verimli hale getirmektedir.

En klasik örneği sepet analizidir. (basket analysis) Bu analizde müşterilerin beraber satın aldığı ürünlerin analizi yapılır. Amaç ürünler arasındaki pozitif veya negatif korelasyonları bularak müşterilerin satınalma alışkanlıklarını ortaya çıkarmaktır. Çocuk bezi alan müşterilerin mama da satın alacağını veya deterjan satın alanların yumuşatıcı da alacağını tahmin edebiliriz ancak manuel olmayan bir analiz bütün olasılıkları göz önüne alır ve kolay düşünülemeyecek, "mama" ve "yumuşatıcı" gibi bağıntıları da bulur. Bu verilere sahip olan marketler, birlikte satılan ürünleri yakın raflara koyarak, katalogda birlikte satılan ürünlerin birlikte görülmesini sağlayarak veya müşteriler için cazip ürün paketleri oluşturarak satışları artırabilirler.

- **Uygulama Alanları:** Birlikte hareket eden verilerin bulunması ile verimlik sağlanacak her alanda kullanılabilir. Süpermarkette birlikte satılan ürünler, otomobilde sunulacak ekstra özellikler, depolarda birbirine yakın konumlandırılması gereken ürünler, alışveriş merkezinde olması gereken mağazalar, vb.
- **Örnek Model:** Bir A ürününü satın alan müşteriler aynı zamanda B ürününü da satın alıyorlarsa, bu durum A → B [destek = %2, güven = %60] şeklinde ifade edilir. Buradaki destek ve güven değerleri, birliktelik kuralının ilginçlik ölçüleridir. "Destek" tanımlanan kuralın sıklığını ve "güven" tanımlanan kuralın kabul edilebilirliğini gösterir. %2 oranındaki bir destek değeri, analiz edilen tüm alışverişlerden %2'sinde A ile B ürünlerinin birlikte satıldığını belirtir. %60 oranındaki güven değeri ise A ürününü satın alan müşterilerinin %60'ının aynı alışverişte B ürününü de satın aldığını ortaya koyar . Kullanıcı tarafından minimum destek eşik değeri ve minimum güven değeri belirlenir ve bu değerleri aşan birliktelik kuralları dikkate alınır.
  Büyük veri tabanlarında birliktelik kuralları bulunurken, iki aşamalı bir süreç işletilir. İlk aşamada sık tekrarlanan öğeler bulunur: Bu öğelerin her biri en az, önceden belirlenen minimum destek sayısı kadar sık tekrarlanırlar. İkinci aşamada sık tekrarlanan öğeler arasından güçlü birliktelik kuralları oluşturulur:
- **Yöntemler / Algoritmalar:** Apriori

Sıralı Dizi Analizi (Sequence Analysis / Sequential Patterns):

"X şirketinin hisse fiyatları ile Y şirketinin hisse fiyatları benzer hareket ediyor."

Gözlem sonuçlarının zaman ve mekan özelliklerine göre sıralanmış olarak gösteren sayı dizileridir. Sayısal sıralı verilerdeki trendleri ve döngüleri anlamak için kullanılır. Bu fonksiyonda ilişkili kayıtlar incelenir ve zaman içinde sıkça rastlanan trendler ve benzer trendler bulunur. Bu trendler daha sonra veri içindeki ilişkileri tanımlamak için kullanılır. Bir beyaz eşya perakendecisinin veritabanından buzdolabı alımını takip eden beyaz eşya alımının bulaşık makinesi olduğunun belirlenmesi, doğal afetler veritabanından 6 büyüklüğünde bir deprem olduktan 3 gün sonra Klimanjaro dağının püskürmesi, banka veritabanından ilk üç taksitinden iki veya daha fazlasını geç ödemiş olan müşterilerin %60 olasılıkla kanuni takibe gidiyor olduklarının belirlenmesi gibi örnekleri vardır. Kredi kartı örneğinde belirlenen davranış skoru (behavioral score), başvuru skorundan farklı olarak kredi almış ve taksitleri ödeyen bir kişinin sonraki taksitlerini ödeme/geciktirme davranışını notlamayı amaçlar. Seriler özelliklerine göre "zaman serileri", "mekan serileri", "bölünme serileri" ve "bileşik seriler" olmak üzere dört başlık altında incelenebilirler.

- **Zaman Serisi Analizi / Benzer Zaman Sıraları/ Zaman İçinde Sıralı Örüntüler (Similar Time Sequences / Time Series):** Gözlem sonuçlarının zamana göre sıralanmış şeklidir. Borsada yer alan hisselerin davranışları sık rastlanan bir örneğidir. Günlere göre hisse değeri, yıllara göre faiz oranları, aylara göre üretim fire oranı, vb. gibi örnekleri vardır. Tek bir seri dışında, birden fazla hareket serisi arasında da bağıntı kurmak mümkündür. Bunlar örneğin iki malın zaman içindeki satış miktarları olabilir. Örneğin dondurma satışları ile kola satışları arasında pozitif, dondurma satışları ile salep satışları arasında negatif bir bağıntı beklenebilir.

Zaman serisinde yer alan verilerin davranışları trend ve döngüler (cycle) ile tanımlanır. "Trend" serideki verilerin ortalama değerinde yaşanan değişimi tanımlamak için kullanılır. "Döngü" veride tekrar eden herhangi bir davranışı tanımlamak için kullanılır. Sezonsal veya dönemsel olabilir. Sezonsal olanlar tahmin edilebilir zamanlarda gerçekleşir, (her pazartesi, her yılbaşı, vb.) dönemsel olanlar "n" zaman aralıkları ile kendini tekrarlar.

Zaman serisi analizlerinde veri serisindeki davranışları belirlemek kadar gelecek değerleri tahmin etme çalışmaları da gerçekleştirilir. Hisse değerlerini, ekonomik değerleri, ürün talebini hava durumunu tahmin etmek, vb.)

- **Mekan Serisi:** Gözlem sonuçlarının mekana göre sıralanmış şeklidir. Bölgelere göre satış rakamları, ülkelere göre yaşam süresi, vb.
- **Bölünme Serisi (Frekans):** Gözlem sonuçlarının belirlenen kriterlere göre sıralanmış şeklidir.
- **Bileşik Seri:** Gözlem sonuçlarının iki ya da daha fazla özelliğe göre bir arada gösterilmiş şeklidir.

Veri Madenciliğinin Algoritmaları (Metotları / Teknikleri)

Veri madenciliği, sahip olunan verilerden yola çıkarak daha önce keşfedilmemiş bilgileri ortaya çıkarma ve bunları karar alma sürecinde kullanma yöntemidir. Veri madenciliği, verilerin içerisindeki desenlerin, ilişkilerin, değişimlerin, düzensizliklerin, kuralların ve istatistiksel olarak önemli olan yapıların analiz ve yazılım tekniklerinin kullanılarak ortaya çıkarılmasıdır. Bu açıdan bakıldığında veri madenciliği istatistiksel bir yöntemler serisi olarak görülebilir. Benzer şekilde veri madenciliğiyle ilgili yazılım ürünleri ve uygulamalara bakıldığında da veri madenciliğinin esasen istatistiğin kullanıldığı bir teknik olduğu görülmektedir.

Ancak önemli olan kolaylıkla mantıksal kurallara ya da görsel sunumlara çevrilebilecek nitel modellerin çıkarılmasıdır. Bu bağlamda, veri madenciliği sadece istatistik değildir, insan merkezli bir uygulamadır.

Veri madenciliği literatürü incelendiğinde istatistik ve algoritmalar ağırlıklı sayısız makale ve kitabın olduğu görülmektedir. Bu çalışmada bu yöntemlere alternatif sunmak, yeni yöntemler geliştirmek veya anlatılan konuları tekrarlamak gibi bir hedef ile yola çıkmadık. Asıl olarak ortaya koymak istediğimiz kritik konunun teknikleri en detayına kadar bilmekten çok hangi soruların hangi yaklaşımlar ile çözülebileceği konusunda fikir vermek ve yönetim kalitesinin artırılması için veri kullanımını teşvik etmektir.

Bununla birlikte veri madenciliği fonksiyonlarının kullandığı bazı kritik teknikler ve tanımlamaları şu şekildedir;

- Karar Ağaçları
- Regresyon
- Lojistik Regresyon
- Bayes
- Apriori
- Kümeleme Teknikleri
- Yapay Sinir Ağları

Karar Ağaçları (Decision Trees)

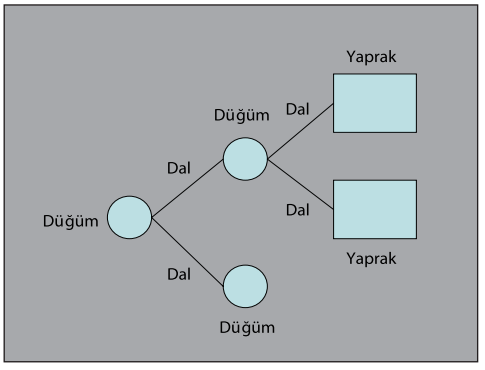
Karar ağaçları, kurgulanmasının, yorumlanmasının ve veri tabanları ile entegrasyonun kolaylığı nedeniyle en yaygın kullanılan öngörü yöntemlerinden / sınıflandırma tekniklerinden biridir. Güvenilirliklerinin iyi olması da bir başka tercih edilme nedenidir. Karar ağaçlarının hedefi bağımlı değişkendeki farklılıkları maksimize edecek şekilde veriyi sıralı bir biçimde parçalarına (farklı gruplara) ayırmaktır. Sınıflandırma ağacı olarak da adlandırılabilir.

İstatistiksel yöntemlerde veya yapay sinir ağlarında veriden bir fonksiyon öğrenildikten sonra bu fonksiyonun insanlar tarafından anlaşılabilecek bir kural olarak yorumlanması zordur. Karar ağaçları ise ağaç oluşturulduktan sonra, kökten yaprağa doğru inilerek kurallar yazılabilir. Bu kurallar uygulama konusunda uzman bir karar vericiye gösterilerek sonucun anlamlı olup olmadığı denetlenebilir. Sonradan başka bir teknik kullanılacak bile olsa karar ağacı ile önce bir kısa çalışma yapmak, önemli değişkenler ve yaklaşık kurallar konusunda karar vericiye bilgi verir.

Yapısı ve Kuruluşu

Karar ağacı, adında belirtildiği şekilde ağaç görünümünde bir tekniktir. Karar düğümleri, dallar ve yapraklardan oluşur.

- **Karar düğümü:** Veriye uygulanacak test tanımlanır. Her düğüm bir özellikteki testi gösterir. Test sonucunda ağacın dalları oluşur. Dalları oluştururken veri kaybı yaşanmaması için verilerin tümünü kapsayacak sayıda farklı dal oluşturulmalıdır.
- **Dal:** testin sonucunu gösterir. Elde edilen her dal ile tanımlanacak sınıfın belirlenmesi amaçlanır. Ancak dalın sonucunda sınıflandırma tamamlanamıyorsa tekrar bir karar düğümü oluşur. Karar düğümünden elde edilen dalların sonucunda sınıflandırmanın tamamlanıp tamamlanmadığı tekrar kontrol edilerek devam edilir.
- **Yaprak:** Dalın sonucunda bir sınıflandırma elde edilebiliyorsa yaprak elde edilmiş olur. Yaprak, verileri kullanarak elde edilmek istenen sınıflandırmanın sınıflarından birini tanımlar.



Başlangıçta bütün öğrenme örnekleri kök düğümdedir, örnekler seçilmiş özelliklere tekrarlamalı olarak göre bölündükten sonra ağacı temizlemek için (Tree pruning) gürültü ve istisna kararları içeren dallar belirlenir ve kaldırılır. Karar ağacı tekniğini kullanarak verinin sınıflanması üç aşamadan oluşur.

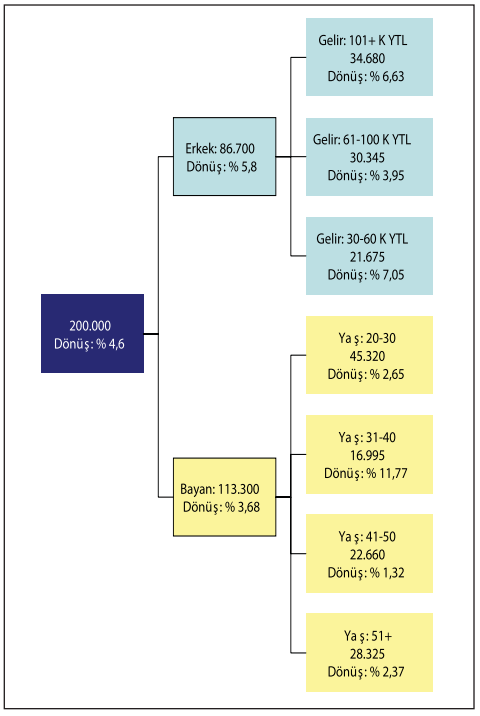
- **Öğrenme:** Önceden sonuçları bilinen verilerden (eğitim verisi) model oluşturulur.
- **Sınıflama:** Yeni bir veri seti (test verisi) modele uygulanır, bu şekilde karar ağacının doğruluğu belirlenir. Test verisine uygulanan bir modelin doğruluğu, yaptığı doğru sınıflamanın test verisindeki tüm sınıflara oranıdır. Her test örneğinde bilinen sınıf, model tarafından tahmin edilen sınıf ile karsılaştırılır.
- **Uygulama:** Eğer doğruluk kabul edilebilir oranda ise, karar ağacı yeni verilerin sınıflanması amacıyla kullanılır.

Uygulama Alanları

Risk grupları kategorileri oluşturmak, gelecekte olabilecek olaylar için tahmin kuralları oluşturmak, kategorilerin birleştirilmesi, yeni bilinmeyen bir örneğin sınıflandırılması gibi durumlarda karar ağaçları kullanılır. Örneğin kredi sınıfını tahmin edecek bir model için aşağıdaki şekilde bir sınıflandırma kuralı oluşturulabilir:

EĞER "yaş" = "41...50" VE "gelir" = yüksek BU DURUMDA "kredi durumu" = mükemmel.

Bu kural gereğince yaşı "41...50" kategorisinde olan (yaşı 41 ile 50 arasında olan) ve gelir düzeyi yüksek bir kişinin kredi durumunun mükemmel olduğu görülür. Oluşturulan bu modelin doğruluğu, bir test verisi aracılığı ile onaylandıktan sonra model, sınıfı belli olmayan yeni bir veriye uygulanabilir ve sınıflama kuralı gereği yeni verinin sınıfı "mükemmel" olarak belirlenebilir.



"Yanıt verme" veya "teklifi kabul etme" gibi beklenen davranış biçimlerini gösterecek sınıfları tanımlamak için (pazarın hareket biçimini anlamak için) uygun ve basit bir yöntemdir. Regresyona göre avantajı lineer olmayan ilişkileri de ortaya çıkarabilmesidir. Bu çalışma ile elde edilen veriler daha farklı modellerde de kullanılabilir. Yandaki örnekte en üst düğümün tüm kampanyanın genel sonucunu gösterdiği bir kredi kartı kampanyası kurgulanmıştır. Kredi kartı satışı için 200.000 kişiye ulaşılmıştır. Sonuç olarak %4,6'sı olan 9.200 kişiden geri dönüş alınmıştır. İlk aşamada cinsiyete göre bir ayrım yapılmıştır. Buradan erkeklerin daha yüksek geri dönüş yaptıkları görülmüştür. Eğer model bu aşamada tamamlansaydı erkekler daha iyi bir hedef kitledir sonucu elde edilecekti ancak cinsiyet ayrımı çok geniş bir ayrım olduğundan her cinsiyet grubunun içindeki alt grupları da bulunmuştur. İkinci aşamada iki grup kendi içlerinde farklı şekillerde alt gruplara ayrılmışlardır. Erkekler için gelir düzeyi, bayanlar için yaş ikinci seviyedeki ayrım olarak düşünülmüş ve sonuç olarak en yüksek yanıt verme ihtimali olan hedef kitleler belirlenmiştir. Bundan sonra yapılacak bir kredi kartı kampanyasında 101.000 YTL ve üzerinde geliri olan erkek ve 31-40 yaşları arasındaki bayanlara ulaşılması durumunda daha az maliyet ile daha yüksek oranlarda geri dönüşün elde edildiği bir kampanya gerçekleştirilmiş olacaktır.

Regresyon Analizi (Regression Analysis)

Bir ya da daha çok değişkenin başka değişkenler cinsinden tahmin edilmesini sağlayacak ilişkiler bulmak ve bunları tanımlamaktır. Regresyon analizinin temelinde gözlenen bir olayın değerlendirilirken, hangi olaylardan etkilendiğini belirlemek yatmaktadır. Bu olaylar bir veya birden çok olacağı gibi etki düzeyleri farklı seviyelerde de olabilir.

Yapısı ve Kuruluşu

Regresyonda, verilerin matematiksel gösterimle, bir fonksiyon olarak tanımlanması gerekmektedir. Regresyon analizi yapılırken kurulan matematiksel modelde yer alan değişkenler bir bağımlı değişken ve bir veya birden çok bağımsız değişkenden oluşmaktadır. Değişkenler sayılabilir veya ölçülebilir niteliktedir. Örneğin bir hissenin fiyatını ile ona dolaylı veya direkt etkili olan faiz oranları, enflasyon, vb. gibi bir veya birden çok değişken ile ilişkilendirmek mümkündür. Sadece faiz oranlarının etkisi ile ilgileniyorsak, tek değişkenli bir matematiksel model, faiz oranları ile birlikte enflasyon oranı ile de ilgileniyorsak, iki değişkenli bir matematiksel model kurulmalıdır. Tek değişkenli modeller basit doğrusal regresyon (doğrusal ilişkiyi temsil eden bir doğrunun denklemi formüle edilir), birden fazla bağımsız değişkenli modeller çoklu regresyon modeli konusunu oluşturmaktadır.

- **Tek Değişkenli Regresyon - Lineer Regresyon:** Basit lineer regresyon iki sürekli değişken (tahmin edilmeye çalışılan bağımlı değişken ve bağımsız değişken) arasındaki ilişkiyi tanımlamayı amaçlayan bir tekniktir. Teknik verileri kullanarak bir doğru denklemi oluşturmayı hedefler. Bu doğru oluşturulurken tüm veri noktalarından tahmin edilen eğriye olan uzaklığın karelerinin minimize edilmesi ile doğrunun optimize edilmesi sağlanır. Doğru elde edildikten sonra iki değişken arasındaki ilişkinin gücü R-kare (R-Square) değeri ile tanımlanır. R-kare verinin değişiminin ne ölçüde oluşturulan model (çizilen doğru) ile açıklanabildiğini gösterir.
- **Tek Değişkenli Regresyon - Lineer Olmayan Regresyon:** Bazı durumlarda bağımlı ve bağımsız değişkenler arasındaki ilişki doğrusal olmayabilir. Bu gibi durumlarda daha iyi bir uyum için bağımsız değişkeni modifiye etmek gerekebilir.
- **Çoklu Regresyon:** Pazarlama, risk yönetimi, müşteri ilişkileri yönetimi konularında model oluşturulurken birden fazla değişkenin bağımlı değişken üzerinde etki ediyor olması çok doğal ve genellikle rastlanan bir durumdur. Bazı durumlarda değişkenler yüzler ile ifade edilecek seviyelere çıkabilir.

Uygulama Alanları

İki değişken arasındaki ilişkiyi bulmak, ilişki varsa bu ilişkinin gücünü belirlemek, değişkenler arasındaki ilişkinin türünü belirlemek, ileriye dönük değerleri tahmin etmek gibi konularda kullanılır. Regresyon analizi, araştırma, matematik, finans, ekonomi, tıp gibi bilim alanlarında yoğun olarak kullanılmaktadır. "Ev sahibi olan, evli, aynı iş yerinde beş yıldan fazladır çalışan, geçmiş kredilerinde geç ödemesi bir ayı geçmemiş bir erkeğin kredi skoru 825'dir." sonucu bir regresyon ilişkisidir.

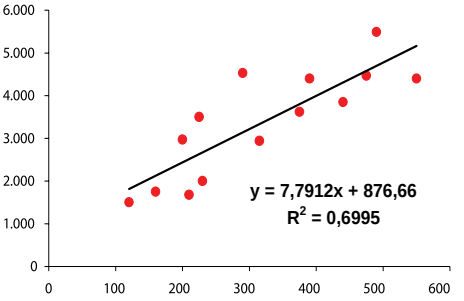
Aşağıdaki örneklerde reklam harcamaları ile satış rakamları arasındaki ilişkiler farklı regresyon yöntemlerine göre belirlenmiş ve R-kare değerleri bulunmuştur. (y: satış değeri, x:reklam değeri, z: enflasyon)

İlk örnekte satış değerlerindeki değişimi reklam harcamaları cinsinden tanımlamak için tek değişkenli lineer regresyon uygulanmıştır. R-karenin aldığı 0,7 değeri, çizilen doğrunun verileri yüksek bir düzeyde açıkladığını göstermektedir.

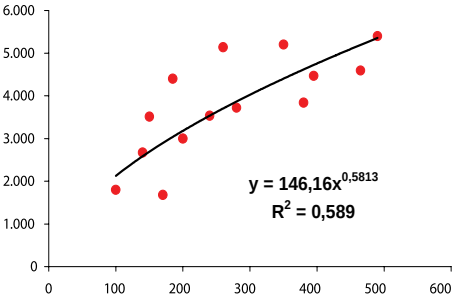
İkinci örnekte satış değerlerindeki değişimi reklam harcamaları cinsinden tanımlamak için tek değişkenli lineer olmayan regresyon uygulanmıştır.

Üçüncü örnekte bir bağımsız değişken daha değerlendirmeye katılmıştır ve satışlar iki boyutlu olarak tanımlanmıştır.

Reklam Harcaması	Satış Değeri
120	1.500
160	1.750
200	2.970
210	1.680
225	3.500
230	2.000
290	4.530
315	2.940
375	3.620
390	4.400
440	3.850
475	4.470
490	5.490
550	4.400



Reklam Harcaması	Satış Değeri
100	1.800
140	2.670
150	3.510
170	1.680
185	4.400
200	3.000
240	3.530
260	5.140
280	3.720
350	5.200
380	3.840
395	4.470
465	4.590
490	5.400



Reklam Harcaması	Enflasyon	Satış Değeri
120	3,4%	1.500
160	3,3%	1.755
205	3,6%	2.970
210	3,5%	1.680
225	3,4%	3.500
230	3,3%	2.000
290	3,2%	4.530
315	3,3%	2.940
375	3,3%	3.620
390	3,4%	4.400
440	3,2%	3.840
475	3,1%	4.470
490	3,2%	5.490
550	3,2%	4.400

$y = 415,6 + 7,9x + 12781z$

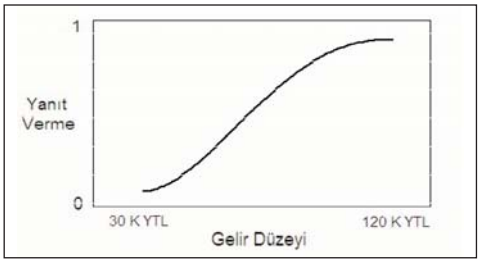
$R^2 = 0,70$

Lojistik Regresyon (Logistic Regression)

Lojistik regresyon lineer regresyona çok benzer olmakla birlikte, lojistik regresyonda bağımlı değişkenin kesikli veya kategorik olması (sürekli olmaması) en önemli farklılıktır. Bu fark özellikle bir teklife yanıt veya bir seçim yapmak gibi kesikli aksiyonları belirlemeye yönelik sınıflandırma modellerinde önem kazanmaktadır. (Sınıflandırma analizlerinde doğrusal regresyonun kullanılması mümkün olmamaktadır.) Lojistik regresyon, çok değişkenli normal dağılım varsayımına ihtiyaç göstermediğinden bu tür uygulamalarda avantaj sağlamaktadır.

Lojistik regresyon ile bağımsız değişkenleri kullanarak ikili çıktısı olan bağımlı değişkenin istenilen durumunun gerçekleşme olasılığını hesaplanır. Regresyon yapabilmek için bağımlı değişken sürekli değere dönüştürülür. Bu değer beklenen olayın olma olasılığıdır.

İşlem şu şekilde yapılır. Her gelir değeri için gelire göre verilen yanıtların ortalamalarından bir olasılık hesaplanır. (p: eldeki verilere göre her gelir düzeyinde teklifin kabul edilme sıklığı), daha sonra her gelir düzeyinde teklifin kabul edilme olasılığı hesaplanır. (p/(1-p) ile). Son aşamada olasılıkların logaritmik değerleri log(p/(1–p)) ile hesaplanır.



Yandaki grafikte gelir ile müşterilere yapılan bir teklife verilen cevaplar arasındaki ilişki kurulmuştur. Yüksek gelir sahiplerinin olumlu yanıt verme ihtimallerinin yüksek olduğu net olarak görülmektedir. Yanıt alma denklemi aşağıdaki şekilde oluşmuştur.

Log (p/(1-p)) = 4,9 + 0,0911 x Gelir

Bayes

İstatistiksel bir sınıflandırıcıdır. Eldeki verilerin belirlenmiş olan sınıflara ait olma olasılıklarını öngörür. İstatistikteki Bayes teoremine dayanır. Bu teorem; belirsizlik taşıyan herhangi bir durumun modelinin oluşturularak, bu durumla ilgili evrensel doğrular ve gerçekçi gözlemler doğrultusunda belli sonuçlar elde edilmesine olanak sağlar. Belirsizlik taşıyan durumlarda karar verme konusunda çok kullanışlıdır. En önemli zafiyeti değişkenler arası ilişkinin modellenmiyor olması ve değişkenlerin birbirinden tamamen bağımsız olduğu varsayımıdır.

Yapısı ve Kuruluşu

Bayes yöntemi koşullu olasılık durumları ile ilgilidir. Her hangi bir koşullu olasılık durumu P(X=x | Y=y) = R şeklinde tanımlanır. Bu ifade; "Eğer Y = y doğru ise, X = x olma olasılığı R'dir" anlamına gelmektedir. X ve Y 'nin alabileceği değerlerin her kombinasyonu için koşullu olasılıkları belirleyen tabloya koşullu olasılık dağılımı adı verilir ve P(X|Y) ile ifade edilir.

Bayes Kuralı şu şekilde tanımlanır.

P(XIY) = P(YIX) x P(X) / P(Y)

Bu ifade; Y'nin gerçekleşmesi halinde X'in gerçekleşme ihtimalinin ne olduğunu belirtmektedir. Bu değeri bulabilmek için "X'in gerçekleştiği durumlarda Y'nin gerçekleşme ihtimali" ile X'in gerçekleşme ihtimalini çarpmak ve bunu Y'nin gerçekleşme ihtimaline bölmek gereklidir.

Örneğin; bir cep telefonu operatörü müşterileri arasında yaptığı araştırma ile cep telefonu kullanımı arka arkaya 3 ay sürekli düşüş gösteren müşterilerinin %20'sinin hattını kapatarak başka operatöre geçtiğini tespit etmiştir. Ayrıca araştırmalardan her 100 müşterinin 6'sının (çeşitli nedenlerle) hattını kapattığı ve her 100 müşterinin 14'inde arka arkaya 3 ay sürekli düşüş yaşandığı tespit edilmiştir.

Bu bilgiler doğrultusunda hattını kapatan bir müşterinin, kullanımında son 3 ayda sürekli azalma olan bir müşteri olma ihtimali nedir?

P(Düşüş I Kapatmış) = P(Kapatmış I Düşüş) x P(Düşüş) / P(Kapatmış)

P(Düşüş I Kapatmış) = (0,2 x 0,14) / 0,06 = % 47

Bu değer hattını kapatan müşterilerin yaklaşık yarısının kullanımında son 3 ayda sürekli azalış olan müşterilerden geldiğini göstermektedir. Bu oldukça yüksek bir orandır. Şirket bu müşterilerin kimler olduğunu kullanım trendlerinden önceden tahmin edebilmektedir. Eğer bu tür müşteriler yeni alternatifler sunularak ayrılmaktan vazgeçirebilirlerse, toplam kaybedilen müşterinin yarısı elde tutulacaktır.

Örneğin; üç farklı kurye şirketinin faaliyet gösterdiği bir şehirde, gece yaşanan bir trafik kazasının tek görgü tanığı mavi bir kurye aracının kazayı gerçekleştirdiğini söylemektedir.

Şehirde faaliyet gösteren üç kurye şirketinin kırmızı, mavi ve sarı olmak üzere farklı renklerde araçları vardır. Kazanın olduğu gece Kırmızı kuryenin 7, mavi kuryenin 4 ve sarı kuryenin 9 aracı hizmet vermektedir. Gece karanlığında rengi doğru olarak görme olasılığı %70'dir. Bu durumda görgü tanığının ifadesi ne ölçüde doğrudur.

Burada belirlenmek istenen P(Mavi I İddia-Mavi), yani iddianın mavi olduğu durumda aracın gerçekten mavi olması ihtimalidir.

P(Mavi I İddia-Mavi) = P(İddia-Mavi I Mavi) x P(Mavi) / P(İddia-Mavi)

P(İddia-Mavi I Mavi): Görgü tanığının araç mavi ise iddiasının mavi olması olasılığı %70'dir. Bu olasılık doğru görme olasılığıdır.

P(Mavi): Gece toplam çalışan araç sayısına göre aracın mavi olma ihtimali 4/20 = %20'dir.

P(İddia-Mavi): Görgü tanığının mavi iddiasında bulunması iki şekilde olur. Gerçekten mavidir ve doğru görmüştür = 0,2 x 0,7 = 0,14, Mavi değildir ve yanlış görmüştür = 0,8 x 0,3 = 0,24, ikisinin toplamında P(İddia-Mavi) = 0,24+0,14 = 0,38 olur.

P(Mavi I İddia-Mavi) = 0,7 x 0,2 / 0,38 = %37'dir.

Aracın mavi olma ihtimali %37'dir. Bu değerlendirmeye göre görgü tanığının ifadesine rağmen kazayı diğer iki şirketten birinin yapmış olması ihtimali daha yüksektir.

Uygulama Alanları

Belirsiz durumlarda tahmin yapmak, sınıflandırma yapmak için kullanılır.

Aşağıda ABD'de 2008'in ilk üç ayında vizyona giren filmlerin bazı özellikleri ve sonuçta elde edilen hasılatları özetlenmiştir.

60 milyon USD üzerinde hasılat yapılan filmlerin başarılı olarak kabul edildiği varsayılırsa, yukarıdaki tablodaki verilerden yola çıkarak yeni vizyona girecek Küçük bir şirket tarafından yapılan, drama tarzında, erkek starı olan ancak kadın starı olmayan, 20-30 mio bütçeli, bir filmin hasılatının 60 mio USD'yi aşma ihtimali var mıdır?

	Film	Tür	Kadın Star	Erkek Star	Şirket	Bütçe	Hasılat
1	I Am Legend	Bilim Kurgu	Yok	Var	Warner Bros	150	256
2	National Treasure: Book of Secrets	Macera	Yok	Var	Walt Disney Pictures	130	220
3	Alvin and the Chipmunks	Aile Filmi	Yok	Yok	Fox 2000	70	217
4	Juno	Komedi	Yok	Yok	Fox Searchlight	8	144
5	The Bucket List	Komedi	Yok	Var	Storyline	45	94
6	Jumper	Bilim Kurgu	Yok	Yok	20th Century Fox	85	80
7	Cloverfield	Bilim Kurgu	Yok	Yok	Bad Robot	25	80
8	27 Dresses	Komedi	Yok	Yok	Fox 2000	25	77
9	No Country for Old Men	Drama	Yok	Yok	Paramount Vantage	25	75
10	The Spiderwick Chronicles	Aile Filmi	Yok	Yok	Kennedy	90	71
11	Vantage Point	Macera	Var	Var	Columbia	40	72
12	Fool's Gold	Macera	Yok	Yok	Warner Bros	70	70
13	Hannah Montana/Miley Cyrus	Diğer	Var	Yok	Pace	7	65
14	Step Up 2: The Streets	Drama	Yok	Yok	Offspring	25	58
15	Charlie Wilson's War	Drama	Var	Var	Good Time Charlie	75	67
16	P.S. I Love You	Drama	Yok	Yok	Alcon	30	54
17	Sweeney Todd	Diğer	Yok	Var	Dreamworks	50	53
18	Rambo	Macera	Yok	Var	Lionsgate	50	43
19	Welcome Home, Roscoe Jenkins	Komedi	Yok	Var	Universal	25	42
20	There Will Be Blood	Drama	Yok	Var	Ghoulardi	25	40
21	Meet the Spartans	Komedi	Yok	Yok	New Regency	30	38
22	Atonement	Drama	Yok	Yok	Working Title	30	51
23	The Water Horse	Aile Filmi	Yok	Yok	Beacon	45	40
24	First Sunday	Komedi	Yok	Var	Cubevision	25	38
25	Semi-Pro	Komedi	Yok	Yok	Donnes Co.	25	34
26	Definitely, Maybe	Drama	Yok	Yok	Universal	25	32
27	The Eye	Korku	Var	Yok	Lionsgate	25	31
28	One Missed Call	Korku	Yok	Yok	Alcon	25	27
29	The Other Boleyn Girl	Drama	Var	Yok	BBC	35	27
30	Untraceable	Macera	Yok	Yok	Cohen / Pearl	35	29
31	Mad Money	Komedi	Yok	Yok	Big City	22	21
32	The Pirates Who Don't Do Anything	Aile Filmi	Yok	Yok	Big Idea	25	13
33	Be Kind Rewind	Komedi	Yok	Var	Partizan	20	11
34	Penelope	Komedi	Yok	Yok	Stone Village	15	10

Bu hesaplama için öncelikle olasılık değerleri belirlenmelidir. P(60+) = 15/34 P(60-) = 19/34'dür. Diğer detaylı olasılıklar aşağıdaki tabloda hesaplanmıştır.

Tür	
P(Bilim-Kurgu I 60+)	3/15
P(Aile I 60+)	2/15
P(Macera I 60+)	3/15
P(Drama I 60+)	3/15
P(Komedi I 60+)	3/15
P(Korku I 60+)	0/15
P(Diğer I 60+)	1/15
Kadın Star	
P(Kadın Star Var I 60+)	3/15
P(Kadın Star Yok I 60+)	12/15
Erkek Star	
P(Erkek Star Var I 60+)	5/15
P(Erkek Star Yok I 60+)	10/15
Yapımcı	
P(Büyük Şirket I 60+)	9/15
P(Küçük Şirket I 60+)	6/15
Bütçe	
P(Bütçe 0-20 I 60+)	2/15
P(Bütçe 20-40 I 60+)	4/15
P(Bütçe 40-60 I 60+)	2/15
P(Bütçe 60+ I 60+)	7/15

Tür	
P(Bilim-Kurgu I 60-)	0/19
P(Aile I 60-)	2/19
P(Macera I 60-)	2/19
P(Drama I 60-)	5/19
P(Komedi I 60-)	7/19
P(Korku I 60-)	2/19
P(Diğer I 60-)	1/19
Kadın Star	
P(Kadın Star Var I 60-)	2/19
P(Kadın Star Yok I 60-)	17/19
Erkek Star	
P(Erkek Star Var I 60-)	6/19
P(Erkek Star Yok I 60-)	13/19
Yapımcı	
P(Büyük Şirket I 60-)	5/19
P(Küçük Şirket I 60-)	14/19
Bütçe	
P(Bütçe 0-20 I 60-)	1/19
P(Bütçe 20-40 I 60-)	15/19
P(Bütçe 40-60 I 60-)	3/19
P(Bütçe 60+ I 60-)	0/19

Öğrenilmek istenen durum X= (Drama – Küçük Yapımcı – Erkek Star – 20-30 mio USD bütçe) şeklinde ifade edilebilir.

- P(60+ I X) = P(X I 60+) x P (60+) / P(X) ile
- P(60- I X) = P(X I 60-) x P (60-) / P(X) karşılaştırılacağından iki denklemde de yer alan P(X) değerleri kaldırılabilir.

Soru P(X I 60+) x P (60+) ile P(X I 60-) x P (60-) değerlerinin karşılaştırılması şekline getirilir.

P(X I 60+) x P (60+) = P(Drama I 60+) x P(Küçük Şirket I 60+) x P(Erkek Star Var I 60+) x P(Kadın Star Yok I 60+) x P(Bütçe 20-40 I 60+) x P (60+)

P(X I 60+) x P (60+) = 3/15 x 6/15 x 5/15 x 12/15 x 4/15 x 15/34 = 0,003

P(X I 60-) x P (60-) = P(Drama I 60-) x P(Küçük Şirket I 60-) x P(Erkek Star Var I 60-) x P(Kadın Star Yok I 60-) x P(Bütçe 20-40 I 60-) x P (60-)

P(X I 60-) x P (60-) = 5/19 x 14/19 x 6/19 x 17/19 x 15/19 x 19/34 = 0,024

Değerlendirme sonucunda söz konusu filmin 60 mio USD üzerinde hasılat yapamayacağı görülmektedir.

[4] Warner Bros, Walt Disney Pictures, Fox 2000, Fox Searchlight, 20th Century Fox,, Paramount Vantage, Columbia, Dreamworks, Lionsgate, Universal yapımcı şirketleri büyük şirket, diğerleri küçük şirket kabul edilmiştir.

Apriori Algoritması

Sık tekrarlanan öğeleri bulmak için kullanılan en temel yöntemdir.

Yapısı ve Kuruluşu

Apriori algoritmasında sık geçen öğe kümelerini bulmak için birçok kez veritabanını taramak gerekir. İlk taramada bir elemanlı minimum destek eşik değerini sağlayan sık geçen veriler bulunur. İzleyen taramalarda bir önceki taramada bulunan sık geçen veriler aday veriler adı verilen yeni potansiyel sık geçen verileri üretmek için kullanılır. Aday verilerin destek değerleri tarama sırasında hesaplanır ve aday kümelerinden minimum destek eşik değerin sağlayan veriler o geçişte üretilen sık geçen veriler olur. Sık geçen veriler bir sonraki geçiş için aday veriler olurlar. Bu süreç yeni bir sık geçen veri bulunmayana kadar devam eder.

Uygulama Alanları

Geleneksel kullanım alanı marketlerde ürünler arası ilişkileri tanımlamaktır. Benzer şekilde hızlı tüketim mamulleri üreticisi firmalarda depo sistemlerinin optimizasyonunda da kullanılabilir. Genel olarak birlikte sevk edilen ürünlerin yakın raflara yerleştirilmesi depo içindeki hareketi ve taşıma miktarını azaltıcı sonuçlar sağlayacaktır. Yöntem restoranlarda servis hızının artırılması için de çözümler sağlayabilir. Müşterilerin sipariş etme ihtimali olan ürünleri önceden tahmin ederek hazırlamak veya ilişkili ürünlerden mönüler oluşturmak gibi çözümler üretilebilir.

Aşağıdaki tabloda 15 alışveriş fişinden oluşan küçük bir market veri tabanı tanımlanmıştır. Buradaki alımlardan yola çıkarak hangi ürünlerin birlikte alındıkları belirlenecektir.

Fiş No	Yoğurt (Y)	Meyve (M)	Et (E)	Peynir (P)	Sabun (S)	Deterjan (D)	Çikolata (Ç)
001	●						
002		●	●	●		●	
003	●	●	●	●			
004			●	●			●
005					●		
006				●	●		
007		●	●				
008			●	●	●	●	
009					●	●	●
010		●		●		●	
011	●				●		
012			●	●			
013		●			●		
014					●		
015		●			●		

İlk aşamada her ürünün kaç kez alındığının belirlenmelidir. Destek değeri olarak da adlandırılan bu değerler kümesi C1, minimum alışveriş destek sayısının 2 olduğu varsayımı doğrultusunda tek başlarına sık tekrarlanan ürünler L1 kümesinde görülmektedir. C1 kümesindeki tüm ürünlerin destek sayısı, minimum destek eşik değeri olan 2'den fazla olduğu için C1 tüm ürünler sık tekrarlanan ürün olarak değerlendirilir ve L1 kümesine aktarılır.

Ürün	Destek Değeri (C1)	Sık Tekrarlanan (L1)
Yoğurt (Y)	3	3
Meyve (M)	6	6
Et (E)	7	7
Peynir (P)	8	8
Sabun (S)	9	9
Deterjan (D)	4	4
Çikolata (Ç)	4	4

Sonraki aşamada hangi ürünlerin ikili olarak sık tekrarlandığını belirlemek için L1 kümesindeki ürünlerin ikili kombinasyonları bulunarak C2 kümesi oluşturulur. C2 kümesindeki ürünlerden minimum destek eşik değerini asan ürünler L2 kümesine aktarılır.

Ürün Grubu	Destek Değeri (C2)	Sık Tekrarlanan (L2)
Yoğurt - Meyve	1	-
Yoğurt - Et	1	-
Yoğurt - Peynir	1	-
Yoğurt - Sabun	1	-
Yoğurt - Deterjan	0	-
Yoğurt - Çikolata	0	-
Meyve - Et	3	3
Meyve - Peynir	3	3
Meyve - Sabun	2	2
Meyve - Deterjan	2	2
Meyve - Çikolata	0	-
Et - Peynir	5	5
Et - Sabun	1	-
Et - Deterjan	2	2
Et - Çikolata	1	-
Peynir - Sabun	2	2
Peynir - Deterjan	3	3
Peynir - Çikolata	1	-
Sabun - Deterjan	2	2
Sabun - Çikolata	1	-
Deterjan - Çikolata	1	-

Hangi ürünlerin üçlü olarak sık tekrarlandığını belirlemek için L2 kümesindeki ürünlerin üçlü kombinasyonları bulunarak C3 kümesi oluşturulur. C3 = MEP – MES – MED – MPS – MPD – MSD – EPS – EPD – ESD – PSD olması beklenir ancak Apriori algoritmasına göre, sık tekrarlanan öğelerin alt kümeleri de sık tekrarlanan öğe olması gerekmektedir. Bu nedenle Et-Sabun ikilisi sık tekrarlanan olmadığından bu alt kümeye sahip MES – EPS – ESD elenmiş olur. Geriye kalanların destek değerleri belirlenir.

Ürün Grubu	Destek Değeri (C3)	Sık Tekrarlanan (L3)
Meyve - Et - Peynir	2	2
Meyve - Et - Deterjan	1	-
Meyve - Peynir - Sabun	0	-
Meyve - Peynir - Deterjan	1	-
Meyve - Sabun - Deterjan	0	-
Et - Peynir - Deterjan	2	2
Peynir - Sabun - Deterjan	1	-

Hangi ürünlerin dörtlü olarak sık tekrarlandığını belirlemek için L3 kümesindeki ürünlerin dörtlü tek kombinasyonu olan M-E-P-D incelenir. Ancak bu kümenin destek değeri sık tekrarlanan limitinin altında olduğundan Apriori yöntemi tüm sık tekrarlanan öğeleri bularak tamamlanmış olur.

Sık tekrarlanan öğeler bulduktan sonra, veritabanından birliktelik kuralları çıkartılır.

Önerme	Güven	Güven Oranı
"Meyve" ve "Et" alırsa "Peynir" alınır	2/3	%67
"Meyve" ve "Peynir" alırsa "Et" alınır	2/3	%67
"Peynir" ve "Et" alınırsa "Meyve" alınır	2/5	%40
"Meyve" alınırsa "Et" ve "Peynir" alınır	2/6	%33
"Et" alınırsa "Meyve" ve "Peynir" alınır	2/6	%33
"Peynir" alınırsa "Et" ve "Meyve" alınır	2/7	%29

"Deterjan" ve "Et" alırsa "Peynir" alınır	2/2	%100
"Deterjan" ve "Peynir" alırsa "Et" alınır	2/3	%67
"Peynir" ve "Et" alınırsa "Deterjan" alınır	1/4	%25
"Deterjan" alınırsa "Et" ve "Peynir" alınır	2/4	%50
"Et" alınırsa "Deterjan" ve "Peynir" alınır	2/6	%33
"Peynir" alınırsa "Et" ve "Deterjan" alınır	2/7	%29

Minimum güven eşiği değerinin %60 olarak belirlendiği bir durumda; 1., 2,, 7. ve 8. kurallar eşik değerini aştıkları için dikkate alınırlar.

Kümeleme Yöntemleri

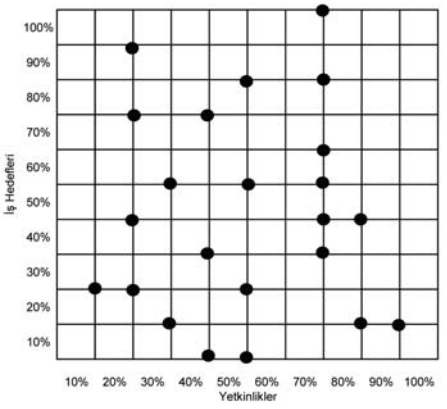
Bölünmeli yöntemler: Veriyi bölerek, her grubu belirlenmiş bir kritere göre değerlendirir. En yaygın olarak kullanılan iki algoritma vardır.

- **K-ortalaması (K-means):** Başlangıç olarak verinin kaç kümeye ayrılacağını belirlemek gereklidir. Küme sayısı "k" değeri olarak adlandırılır. k-means algoritmasının 4 aşaması vardır:
  - Veri kümesinin rastsal olarak k altkümeye ayrılması (her küme bir altküme),
  - Her kümenin ortalaması olan merkez noktanın (kümedeki nesnelerin niteliklerinin ortalaması) hesaplanması
  - Nesnelerin küme merkezine olan uzaklıklarının değerlendirilmesi ve dahil olduğu kümenin merkezinden başka bir küme merkezine daha yakın olan nesnelerin yakın oldukları kümeye dahil edilmesi
  - Yeni nesnelerle artan veya dışarıya nesne vererek azalan kümelerin ortalaması olan merkez noktaları yeniden hesaplanır ve nesnelerin kümelenmesinde değişiklik olmayana kadar aynı şekilde devam edilir.

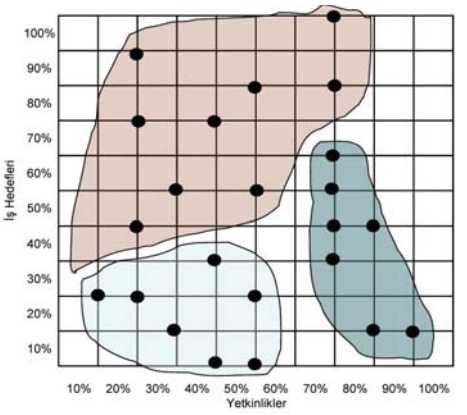
K-means yöntemi kurgulaması kolay ve karmaşıklığı az olan bir tekniktir. Ancak zayıf olduğu bazı önemli noktalar vardır. Sonuçları ilk başta merkez noktaların seçimine bağımlıdır. Merkez noktaların seçimine göre farklı sonuçlar ortaya çıkabilir. Bununla birlikte veri grupları farklı boyutlarda ise, veri gruplarının şekli küresel değilse ve veri içinde ortalamayı önemli ölçüde etkileyecek büyük bileşenler varsa çok iyi sonuçlar alınamayabilir.

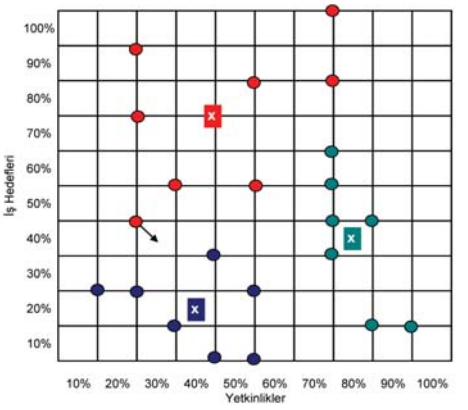
Uygulama

Bir şirkette yapılan performans değerlendirmesi uygulamasında çalışanların yetkinlikleri ve iş hedeflerini gerçekleştirme düzeyleri değerlendirilmiştir. İnsan Kaynakları bölümü çalışanların performans değerlendirme sonuçları doğrultusunda gelişim ve ödüllendirme paketleri oluşturmaktadır. Bu paketler hangi farklı gruplar için oluşturulmalıdır sorusunun yanıtı aranmaktadır.

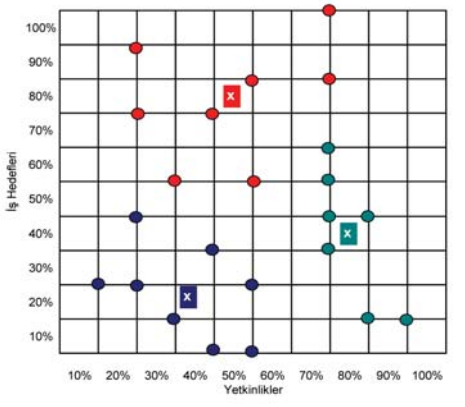


Üç farklı grup olacağı öngörülmüş ve başlangıç olarak yanda görülen gruplama yapılmıştır. Bu gruplamaya göre kümelerin merkezleri belirlenecektir.





Sadece tek bir çalışanın değerleri kendi kümesi dışında başka bir kümeye yakındır. Bu çalışanın da uygun kümeye yerleştirilmesi sonucunda üç grup ve merkezleri şu şekilde oluşmuştur.



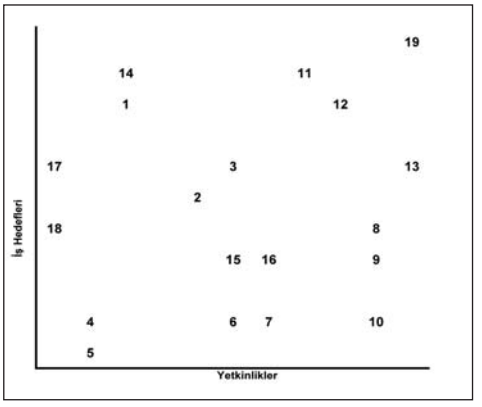
- **K-medoids:** K-means yönteminde; sadece kümenin ortalamasının tanımlanabildiği durumlarda kullanılma ve değeri çok büyük bir nesnenin kümede olması durumunda (kümenin ortalaması ve merkez noktası büyük ölçüde değişebileceğinden) kümenin hassasiyetinin bozulabilmesi gibi iki önemli zafiyet vardır. Bu sorunu gidermek için kümedeki nesnelerin ortalamasını almak yerine, kümede ortaya en yakın noktada konumlanmış olan nesne (medoid) kullanılabilmekte ve bu işlem k-medoids yöntemi ile tanımlanmaktadır. k-medoids yöntemi şu aşamalardan oluşur;
  - Veri kümesi merkezi bir medoid olan k adet kümeye ayrılır.
  - Veri kümesindeki nesneler, kendilerine en yakın olan medoide göre k adet kümeye yerleşirler.
  - Bu bölünmelerin ardından kümenin ortasına en yakın olan nesneyi bulmak için medoid, medoid olmayan her nesne ile yer değiştirir. Bu işlem en verimli medoid bulunana kadar devam eder.

Hiyerarşik yöntemler: Veri kümelerini önceden belirlenmiş bir kritere göre, kümeler ağacı seklinde gruplara ayırma esasına dayanır . Hiyerarşik kümeleme yöntemleri, hiyerarşik ayrışmanın yönüne göre ikiye ayrılır.

- Agglomerative (HAC / AGNES (AGlomerative NESting),) hiyerarşik kümelemede, hiyerarşik ayrışma aşağıdan yukarıya doğru olur .İlk olarak her nesne kendi kümesini oluşturur ve ardından bu atomik kümelerin içinde aralarında en az uzaklık olanlar birleşerek, tüm nesneler bir kümede toplanıncaya dek daha büyük kümeler oluştururlar.
- Divise (DIANA (Dlvise ANAlysis) hiyerarşik kümelemede, hiyerarşik ayrışma yukarıdan aşağıya doğru olur. İlk olarak tüm nesneler bir kümededir ve her nesne tek başına bir küme oluşturana dek, kümeler daha küçük parçalara bölünürler.

Uygulama

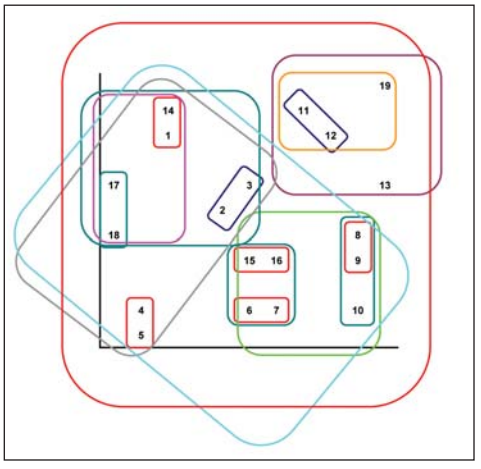
19 çalışanın yukarıdaki örnekteki biçimde iki boyutta değerlendirildiği bir performans çalışması gerçekleştirilmiştir. Bu çalışma sonuçları aşağıdan yukarıya hiyerarşik yöntemle gruplandırılmak istenmektedir.



Çalışanlar aralarındaki minimum uzaklıklara göre aşağıdaki tabloda belirtilen adımlar ile gruplanabilirler.

Adım	Grup	Mesafe
1	1 - 14	1,0
2	4 - 5	1,0
3	6 - 7	1,0
4	8 - 9	1,0
5	15 -16	1,0
6	2 - 3	1,4
7	11 - 12	1,4
8	17 - 18	2,0
9	(15-16) - (6-7)	2,0
10	(8-9) - 10	2,5
11	(11-12) - 19	2,9
12	(11-12-19) - 13	3,4
13	(15-16-6-7) - (8-9-10)	3,5
14	(17-18) - (14-1)	4,0
15	(17-18-14-1) - (2-3)	3,7
16	(17-18-14-1-2-3) - (4-5)	5,9
17	(17-18-14-1-2-3-4-5) - (15-16-6-7-8-9-10)	5,4
18	(17-18-14-1-2-3-4-5-15-16-6-7-8-9-10) – (11-12-19-13)	7,0

Sonuca grafiksel olarak bakıldığın farklı seviyelerde elde edilen kümeler aşağıdaki gibidir.



Bu çalışma sırasında yapılan iterasyonlarda kümelerin merkez noktalarından hareket edilmiştir. Bu nedenle bazı durumlarda kümeler arası mesafe bir önceki iterasyondan daha kısa olmuştur.

İterasyonlar gereksinime göre dört farklı şekilde yapılabilir.

- Kümelerin minimum noktalarının bağlanması
- Kümelerin maksimum noktalarının bağlanması
- Kümelerin elemanlarının ortalama değerlerinin bağlanması
- Kümelerin merkez noktalarının bağlanması

Yukarıdaki grafikle elde edilen sonuçlar dendrogram şekli ile de ifade edilebilir.

Yoğunluk tabanlı yöntemler: Nesnelerin yoğunluğuna göre kümeleri oluşturur. Kümelerin içinde yer alan ortalamaları bozan çok büyük veya çok küçük değerlerden etkilenmeyen yöntemlerdir. Kümeleme iterasyonun sona ermesi önceden belirlenmiş bir yoğunluk parametresi ile olur. En bilinen yöntemleri; Dbscan ve Optics yöntemleridir.

- **Dbscan:** Bu yöntemde iki kriter tanımlanır ve bunun doğrultusunda kümeleme işlemi yapılır. Birinci kriter etki yarıçapı, ikinci kriter minimum eleman sayısıdır. Amaç minimum eleman sayısına ulaşmak ve bunu minimum etki yarıçapı ile gerçekleştirmektir. Etki yarı çapı iterasyonlar ile artırılarak minimum eleman sayısını kapsayıncaya kadar devam edilir. Minimum eleman sayısına ulaşıldığında nokta kümenin çekirdeği olur ve bu işleme diğer noktalar ile devam edilir.

Model tabanlı yöntemler: Her kümeyi oluşturan verilerin bir matematiksel modele uyduğu varsayılır.

Yapay Sinir Ağları (Artificial Neural Networks)

İnsan beyninin işleme mantığını temel alarak, nöronların matematiksel olarak modellenmesidir. Bu yöntem, kurulan modeli kontrol etmekte ve öğrenme faaliyeti ile modeli geliştirmektedir. Süreç davranış biçimlerini anlamak ve hatayı en aza indirmek üzerine kuruludur. Bilgiyi almak ve daha sonra her uygulamadan bir ders çıkarmak gibi düşünülebilir. İstatistiksel yöntemler gibi veri hakkında parametrik bir model öngörmez.

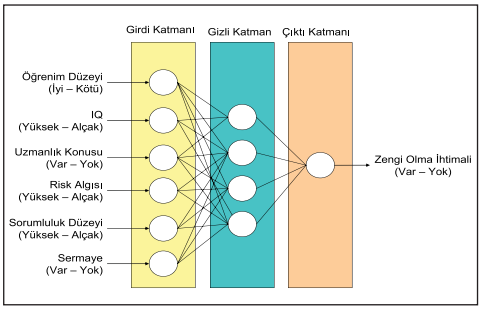
Doğru sınıflandırma sağlayan, doğru sonuçlar veren bir yöntem olmakla birlikte en önemli zafiyetleri öğrenme süresinin uzun olması ve çıkan sonucun ifade edilmesinin / tanımlanmasının güç olmasıdır.

Yapısı ve Kuruluşu

Yapı nöronlar arasındaki bağlantılar ve bağlantıların ağırlıkları (öğrenme mekanizması ile geliştirilen) üzerine kurulur. Modelin karmaşıklığı bu bağlantı yapısına bağlıdır. Nöronların bir araya geldiği alanlara katman denir. (Giriş katmanı, çıkış katmanı ve bu iki katman arasında yer alan gizli katman). Model kurulduktan sonra eğitim verileri sürekli olarak modele girilir ve elde edilen sonuçlar gerçek sonuçlar ile karşılaştırılarak modelde iyileştirmeler (ağırlıklarda değişiklikler) yapılır. Minimum kabul edilebilir hata seviyesine ulaşıldığında model tamamlanmış olur.

Uygulama Alanları

Yapay sinir ağları; halka arzlar, hisse senedi piyasaları tahmini, kredi değerlendirmesi, belirtilere göre hastalık tahmini, vb. alanlarında kullanılmaktadır.



Örneğin dört girdiden oluşan bir sistemde, aşağıdaki tablodan görülmektedir ki, en az 3 girdinin 1 olması halinde çıktı 1 olmaktadır.

Girdi 1	Girdi 2	Girdi 3	Girdi 4	Çıktı
1	1	0	1	1
0	0	0	1	0
1	1	1	1	1
1	0	0	0	0
1	1	1	0	1
0	0	0	1	0
1	0	1	0	0
0	0	1	1	0
0	1	0	0	0
1	1	1	0	1
1	0	0	1	0

Çıktının fonksiyonu şu şekilde tanımlanabilir.

Çıktı = 1 eğer (0,3xG1 + 0,3xG2 + 0,3xG3 + 0,3xG4 – 0,8) > 0

Çıktı = 0 eğer (0,3xG1 + 0,3xG2 + 0,3xG3 + 0,3xG4 – 0,8) < 0

Yukarıdaki örnekte sınırlı sayıda veri olduğundan tüm veriler gözle hesaplanacak kadar basit ağırlık değerleri kullanılarak doğru biçimde sınıflandırılmıştır. Çok sayıda veriden olan sistemlerde Verilerin tümünü doğru sınıflandırmak için "ağırlıkları belirleme işlemi" (eğitme işlemi) şu adımlar ile gerçekleştirilir.

- Başlangıç ağırlık değerleri vermek,
- Bu başlangıç değerlerine göre tüm verilerin sonucuna bakmak,
- Oluşan hayatı belirlemek
- Hataların karelerini minimum edecek şekilde ağırlıkları değiştirmek.

Sonuç

SONUÇ

Veri Madenciliği istatistik biliminin teknolojiyle bütünleşmesi sonucu oluşmuş bir yöntemler serisidir. Bilgi teknolojilerinin gelişmesi ve konu ile ilgili yeni programların üretilmesi çalışmaları kolaylaştırmaktadır. Ancak veri madenciliği sadece program kullanmak değildir.

Veri madenciliği için iş deneyimine, sorunları tanımlama becerisine ve temel istatistik bilgisine ihtiyaç vardır. Veri madenciliği veriden bilgi üreterek ortalama kararlar yerine veriye dayalı özgün kararlar verilmesini destekleyen, satışları, kârlılığı, yenilikçiliği ve kaynak kullanımında etkinliği artıran önemli bir yönetim aracıdır. Veriye dayalı kararların kalitesi ve güvenilirliği artar; bu veriye dayalı kararlarla çalışan kurumların kaynak kullanım etkinliği ve değer yaratma potansiyeli de gelişir.

Teoride, teori ile pratik arasında fark yoktur ama pratikte vardır. - Jan L. A. van de Snepscheut

Okuma Önerileri

Akyouş Selim, Veri Madenciliği Yöntemlerine Genel Bakış (Sunum)

Alpaydın Ethem, Zeki Veri Madenciliği (Sunum)

Argüden, R. Yılmaz (1982). Management of Large Data Sets: A Case Study with California Oil Wells. The RAND Corporation, P-6802

Argüden, R. Yılmaz (1988). Principles for Dealing with Large Programs and Large Data Files in Policy Studies. The RAND Corporation, P-7409

Ayres, Ian, (2008). Super Crunchers, Bantam Books

Bishop, C., (1996). Neural Networks for Pattern Recognition, Oxford Univ Press

Berry Michael J., Linoff, Gordon S., (2000). Mastering data mining. New York: Wiley

Berry Michael J., Linoff, Gordon S., (2004). Data Mining Techniques For Marketing Sales And Customer Support, New York: Wiley

Bilgin, Turgay T., Maltepe Üniversitesi Bilgisayar Mühendisliği (BIL 416) (Ders Notları)

Crisp-DM 1.0 (2000), SPSS

Edelstein, H., A. (1999). Introduction to data mining and knowledge discovery (3rd ed). Potomac, MD: Two Crows Corp

Fayyad, U. M., Piatetsky-Shapiro, G., Smyth, P., & Uthurusamy, R. (1996). Advances in knowledge discovery & data mining. Cambridge, MA: MIT Pres

Han Jiawei, Kamber Micheline (2006), Data Mining Concepts and Techniques, Morgan Kaufmann

Hastie, T., Tibshirani, R., & Friedman, J. H. (2001). The elements of statistical learning : Data mining, inference, and prediction. New York: Springer.

Mitchell, T. (1997). Machine Learning, McGraw-Hill

Horning Mark F., Marcade Erik, Venkayala Sunil (2007). Java Data Mining: Strategy, Standard and Practice, Morgan Kaufmann

Levitt, Steven D., Dubner, Stephen J., (2005) Freakonomics, Harper Collins

Linoff, Gordon S., Survival Data Mining (Sunum)

Olson David L., Delen Dursun (2008). Advanced Data Mining Techniques, Springer

Pregibon, D. (1997). Data Mining: Statistical Computing and Graphics 7-8.

Relles, Dan (1986). Allocating Research Resources: TheRole of a Data Management Core Unit. The RAND Corporation, N-2383-NICHD

Rencher, A.C., (1995). Methods of Multivariate Analysis, Wiley

Rud, Olivia Par, (2001). Data Mining Cookbook - Modeling Data for Marketing, Risk, and Customer Relationship Management, New York: Wiley

Thearling_K._An_Introduction_to_Data_Mining (Sunum)

Toros Hüseyin, Veri Madenciliğine Giriş (Sunum)

Van Tessel, Dennie (1978). Programming Style, Design, Efficiency, Debugging, and Testing. Prentice-Hall, Inc., Englewood Cliffs

Widner G., Fürnkranz J., Clustering (Sunum)

Weiss, S. M., & Indurkhya, N. (1997). Predictive data mining: A practical guide. New York: Morgan-Kaufman

Westphal, C., Blaxton, T. (1998). Data mining solutions. New York: Wiley.

Witten, I. H., & Frank, E. (2000). Data mining. New York: Morgan-Kaufmann

Ye, Nong (2003) Handbook of Data Mining, Lawrence Erlbaum Associates Publishers

Okuma Önerileri –Teknik

Chakrabarti, Soumen (2003). Mining The Web - Discovering Knowledge From Hypertext Data, Morgan Kaufmann

Pal, Nikhil R., Jain Lakhmi (2004). Advanced Techniques in Knowledge Discovery and Data Mining, Springer

Hand, David, Mannila, Heikki and Smyth, Padhraic (2001). Principles of Data Mining, MIT Pres

Chen, Hsinchun, Fuller, Sherrilynne S ., Friedman, Carol, Hersh, William (2005). Medical Informatics: Knowledge Management and Data Mining in Biomedicine, Springer

Witten, Ian H., Frank Eibe, (2005). Data Mining: Practical Machine Leraning Tools and Techniques, Elsevier

Wang, John (2006). Encyclopedia of Data Warehousing and Mining, Idea Group

Sumathi, S., Sivanandam, S. (2006). Introduction to Data Mining and its Applications, Springer

Felici, Giovanni, Vercellis, Carlo (2008). Mathematical Methods for Knowledge Discovery and Data Mining, Information Science

Evangelos, Triantaphyllou, Giovanni, Felici (2006). Data Mining and Knowledge Discovery Approaches Based on Rule Induction Techniques (Massive Computing), Springer

Larose, D.T., (2007). Data Mining Methods and Models, Wiley

Cook, Diane J., Holder, Lawrence B., (2007). Mining raph Data, Wiley