SAKARYA ÜNİVERSİTESİ

Veri Madenciliği Uygulamaları

Hafta 2

Yrd. Doç.Dr. Nilüfer YURTAY

Veri Madenciliği Uygulama Alanları

2.1 Giriş

Otomatik veri toplama araçları, yaygın bilgi teknolojileri kullanımı, veri tabanları, veri ambarları ve diğer veri depolama ürünlerinde çok büyük miktarlarda veri ve bilgilerin toplanmasını sağlamakta ve sürekli olarak ta veri miktarı sürekli artmaktadır.

- **Mağazalardaki alış/satış işlemleri**
- **Banka ve Kredi kartı işlemleri**
- **Bilimsel veriler, uydu ve radarlardaki algılayıcılardan gelen veriler**
- **Web teknolojilerinden gelen verileri**
- **Bir çok sektördeki veri ve işlemler**

Veri madenciliğinin karar verici için olası yararları aşağıdaki gibi sıralanabilir:

Mevcut müşterilerin karar verici tarafından daha iyi tanınmasını sağlayabilir.

- **Bankacılık**
  - Kredi taleplerinin değerlendirilmesi.
  - Farklı finansal göstergeler arasında gizli korelasyonların bulunması,
  - Kredi kartı üzerinden dolandırıcılıklarının tespiti,
  - Kredi kartı harcamalarına göre müşteri gruplarının ortaya çıkarılması,
- **Pazarlama**
  - Satış tahminleri
  - Müşterilerin satın alma örüntülerinin ortaya çıkarılması,
  - Pazar sepeti analizi,
  - Müşterilerin demografik özellikleri arasındaki bağlantıların bulunması,
  - Düzenlenen kampanyalarda cevap verme oranının artırılması,
  - Mevcut müşterilerin elde tutulması, yeni müşterilerin kazanılması,
  - Müşteri ilişkileri yönetimi ve değerlendirmesi
- **Sigortacılık**
  - Dolandırıcılıklarının tespiti,
  - Yeni poliçe talep edecek müşterilerin tahmin edilmesi,
  - Riskli müşteri örüntülerinin ortaya çıkarılması.

Web madenciliği, veri madenciliği tekniklerinin kullanılarak web belgelerinden ve servislerinden otomatik olarak bilginin ayıklanması, ortaya çıkarılması ve tahlil edilmesidir.[Etzoni]

- 1 Milyarın Üzerinde Sayfa Sayısı (15 TeraBytes!!)
- Dinamik Yapısı
- Her Gün Yarım Milyon Yeni Sayfa
- Her Gün *Yüz Milyonlarca* Sorgu

Yapılabilecek çalışmalara da aşağıdakiler örnek olarak verilebilir:

- Kaynak Bulma (Resource finding)
- Bilgi seçimi ve ön işleme (Information selection and pre-processing)
- Genelleme (Generalization)

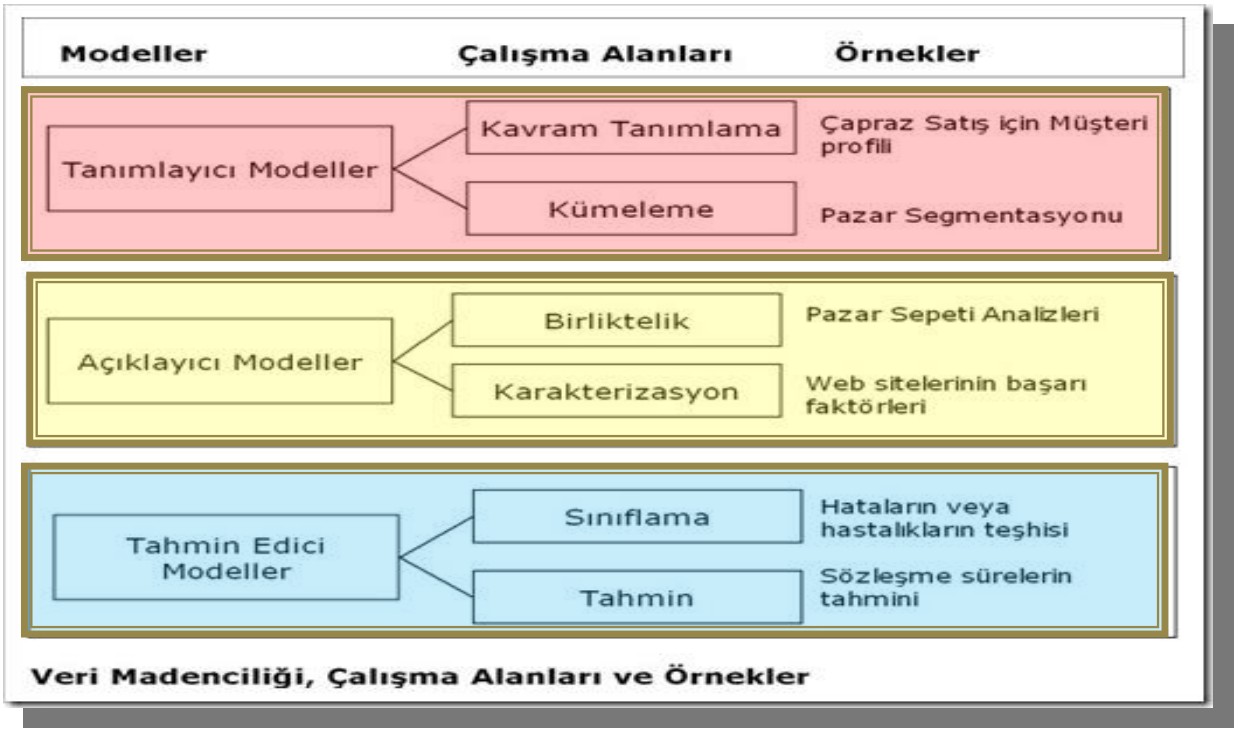
- Analiz (Analysis)

E-Ticaret ve Web Madenciliği ile beklentiler ;

- Müşterilerin yaşam sürelerini öğrenmek
- Yeni stratejiler belirlemek
- Promosyonları değerlendirmek
- Hedef Kitle Seçimek
- Müşterileri tanımak

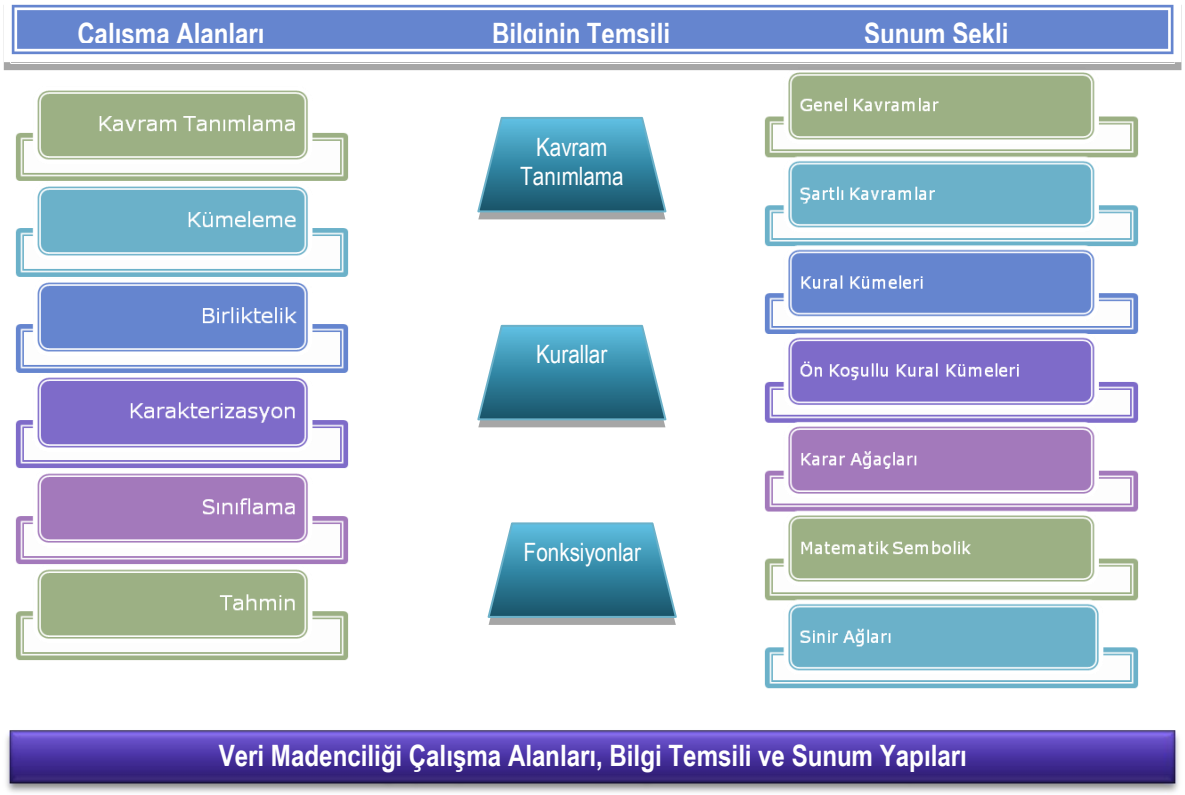
olarak özetlenebilir.

Veri Madenciliği uygulamaları, model ve çalışma alanlarına göre düzenlenebilir(Şekil 2.1).



Şekil 2.1 Veri madenciliği, çalışma alanları ve örnekler

Şekil 2.2 de de veri madenciliği çalışma alanları bilgi temsili ve sunum yapıları ile birlikte gösterilemektedir.



Şekil 2.2 veri madenciliği çalışma alanları bilgi temsili ve sunum yapıları

2.2 Veri madenciliği için çapraz endüstri standart süreci (CRISP-DM)

VM çalışmalarının başarılı bir şekilde sonuçlanabilmesi için sistematik bir yaklaşım zorunludur. VM araçları sunan birçok satıcı, danışma firması ve araştırmacı kullanıcılara yönelik bir çok süreç modelleri oluşturmuşlardır.

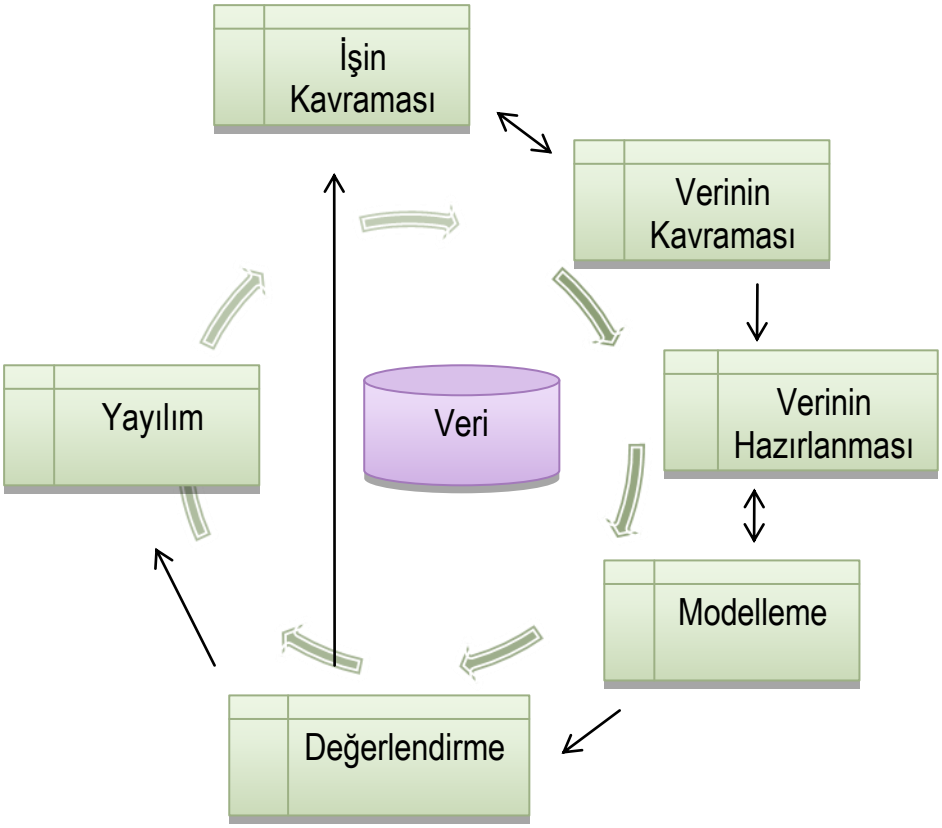
Bunlardan en yaygın olarak kullanılanı, VM araçlarını satan firmalardan bazılarının ve VM uygulamalarını işletme faaliyetlerine uyarlayan ilk kuruluşların oluşturduğu bir konsorsiyum tarafından geliştirilen *Çapraz Endüstri Veri Madenciliği Standart Süreci'*dir.

CRISP-DM, bilgi keşfi için yapılacak VM'nin temel adımlarını tanımlayan hiyerarşik bir süreçtir. VM projesi yaşam döngüsünün gözden geçirilmesini sağlamaktadır. CRISP-DM'de VM projesinde neler yapılması gerektiği üzerine vurgu yapılmaktadır.

Projelerde yapılan ticari işlemlere ilişkin bilgiler CRISP-DM'in aşamaları uygulanarak yeni ve ilginç sonuçların elde edilmesi hedeflenmiştir.

VM projesi yaşam döngüsü altı aşamayı içermektedir. Bu aşamalar aşağıda verilmiştir:

1. İşin Kavranması
2. Verinin Kavranması
3. Verinin Hazırlanması
4. Modelleme
5. Plana Göre Yayılma
6. Değerlendirme



Şekil 2.3 VM projesi yaşam döngüsü[1]

2.3 Sağlık Sistemlerinde Veri madenciliği

Sağlık ve tıp çağımızın en önemli araştırma alanlarıdır.Hastane bilgi sistemleri de araştırmalar için önemli veri kaynaklarıdır. Hastane bilgi sistemlerinde

- Bilgi ve iletişim sistemlerindeki gelişmeler
- Bilgi modelleme ve tanı araçlarında gelişmeler
- Elektronik ortamda saklanabilen birçok veri
  - Hasta demografik bilgileri
  - Hastalık ve tedavi durumları
  - Yapılan tetkikler
  - Faturalama ve idari işlere ait veriler
  - ve diğerleri

yer almaktadır.

[1] Kaynak: Clay Helberg, *Data Mining with Confidence*, SPSS, 2002

Tarihçe olarak bakarsak:

- **Veri madenciliği ve Bilgi Keşfi tanımı:**
  - Verilerde daha önceden bilinmeyen, anlamlı ve değerli bilgiler elde etme işlemleridir.
- **Uzman sistemler (1970 ler), yaygınlaşamama nedenleri:**
  - Verilerin hızlı değişmesi
  - Uzmanlar arasındaki görüş ayrılıkları
- **Yapay Sinir Ağları(1990 lar)**
  - Hastaların gelecekteki sağlık durumlarının tahmini
  - Maliyet tahminleri

olarak bir gruplama yapılabilir.

Veri madenciliği teknikleri ile bu alanda yapılabileceklere örnek olarak aşağıdakileri vermek mümkündür:

- Belirli bir hastalığa sahip hastaların ortak özelliklerinin tahmin edilmesi
- Tıbbi tedavilerinin sonuçlarının tahmin edilmesi
- Hastane maliyetlerinin tahmin edilmesi
- Ölüm oranları ve salgın hastalıkların tahmin edilmesi
- Hastalıkların Yönetimi
  - Hastalıkların durumları (Karar ağaçları kullanılmış)
    - Yaş
    - Cinsiyet ve
    - Diğer özellikler kullanılarak hastalıklara ait bilgilerin keşfedilmesi
  - Maliyet modelleme

- Hastaların sürekli doktor değiştirmelerinin nedenleri
- Hastaların hastanelerdeki kalış süreleri
- Hile Tespiti çalışmaları
- İlaçların yan etkileri
- Tıp ve sağlık alanındaki yapısal olmayan dokümanlar için metin madenciliği çalışmaları
- Biyolojik araştırmalar tıbbi gelişmelere önemli katkılar sağlar.
  - BIOMART
    - İlişkisel veriler için geliştirilen veri madenciliği sistemidir.
  - OMIM(Online Mendelian Inheritance in Man)
    - National Center for Biotechnology Information(NCBI) tarafından geliştirilmiştir
    - Genetik bozukluklarla ilgili bilinen hastalıkların saklandığı veritabanıdır.

- İnsan Genom projesinin tamamlanması ile birlikte genom bilgileri ile klinik bilgiler arasında bağ kurabilmek için projeler ve veritabanları geliştirildi:
  - HapMap Projesi
  - DECIPHER Projesi

- Radyoloji Bilgi Sistemi

Doğru ve zamanında karar almanın hasta sağlığı üzerindeki etkisi tartışmasız çok önemlidir. Hastane bünyesinde toplanan operasyonel veriler, hasta verileri, uygulanan tedavi yöntemi ve tedavi sürecine dair veriler yöneticiler açısından bakıldığında; hastanedeki servislerin ve programların başarısının görüntülenmesi, kaynakların maliyetlerle göreceli olarak kullanımı, kaynak kullanımı ve hasta sayıları ile ilgili trendlerin tahmini, harcamalarla ilgili normal olmayan durumların anlık tespiti ve yolsuzlukların engellenmesi, hastanede uygulanan tedavi yöntemlerinin başarısının irdelenmesi açısından önemli biligleri içermektedir. Bu veriler başarılı tedavi sonuçları almada etken faktörlerin belirlenmesi, ameliyatlarda yüksek risk faktörlerinin sınanması, hasta verilerinin yaş, cinsiyet, ırk ve tedavi yöntemi gibi faktörlere göre sınıflanması, hasta sağlığı açısından geriye dönük faktörlerin sınanması, tedavi yöntemi geliştirme vb. amaçlarla kullanılmaktadır. Dünya çapında çok sayıda başarılı uygulama örneği mevcuttur. Örneğin, San Francisco Hearth Institute; hasta sonuçlarının iyileştirilmesi, hastanın hastanede kalma süresinin azaltılması, vb amaçlarla bir çalışma başlatmış ve kurum bünyesinde toplanan verilerden hastanın geçmişine ait veriler, laboratuvar verileri, kollestrol verileri, diğer medikal verileri SPSS çözümleri ile bilgiye dönüştürmüştür. Yapılan analizler sonucunda geliştirilen stratejiler ile hastaların hastanede kalma süreleri azaltılmış ve yılda $500.000 tasarruf sağlamıştır. SPSS ile hastaların yaşamını yitirme riskini tahmin etme ve hasta ölümlerini azaltma amacıyla, 17 hasta faktörüne bağlı olarak bir risk modeli geliştirilmiş ve bu yönde başarılı sonuçlar elde edilmiştir.

FLORIDA Hastanesi (US) masrafları azaltma ve hastalarına daha iyi bakım sağlama amaçlı olarak ileri istatistiksel analiz ve veri madenciliği yöntemlerini kullanmış ve ilk yıl masraflarını azaltma yönünde $180.000'lık bir başarı sağlamıştır. Hollanda'da bulunan Ignatius Hospital ise tedavi sürelerinin ve belirli bir sürede tedavi ihtiyacı olan kişi sayısının ve her bir hasta için tedavi süresinin tahmin edilmesi yönünde uygulamalar geliştirmiş ve yaptığı analizler sonucunda elde ettiği bilgi ile hastane kadro ve kaynak ihtiyaçlarının doğru belirlenmesini sağlamış ve geçmiş hasta verilerinden elde ettiği bilgi ile kalp hastalıklarında bypass ameliyatlarının riskini minimuma indirmeyi başarmıştır.

2.4 Bankacılık ve Finans Sistemlerinde Veri Madenciliği

Pek çok banka ve finansal kuruluş hesap açtırma, yatırım yapma gibi banka hizmetleri, otomobil, iş, ev gibi kredi hizmetleri sunmaktadır. Bankacılıkta ve finansal sanayide kullanılan verinin tam, güvenilir ve yüksek kalitede olması sistematik veri analizi uygulanmasını kolaylaştırır.
Bankacılık sektöründe, veri madenciliği kredi sahtekarlığı tahminleme, risk değerlendirme, eylem analiz değerlendirme, kar analizi gibi alanlarda kullanılır.

Finansal kuruluşlarda, stok fiyat tahminleme, gümrük ölçümleme, portföy yönetimi, mal fiyat tahminleme gibi alanlarda yapay sinir ağları kullanılır.

2.5 Hizmet Sektöründe Veri Madenciliği

Perakende sanayi, satış bilgileri, müşteri profili, ürün nakil bilgileri, servis kayıtları gibi pek çok veriyi işlediğinden veri madenciliği kullanır.

Perakende veri madenciliğinin kazanımları:

- Müşteri alış davranışı belirlenmesi
- Müşteri alışveriş örneklenmsei ve eylimin belirlenmesi
- Müşteri servis kalitesinin artırılması
- Müşteri hatırlama ve memnuniyetinin sağlanması
- Ürün tüketim oranının artırılması
- Ürün nakil ve dağıtım düzeninin tasarlanması
- İş maliyetinin düşürülmesi

olarak sıralanabilir.

Telekominikasyon sanayide, ses, fax, cep telefonu, resim, e-posta, Internet gibi kapsamlı iletişim servisleri hem uzak hem de yakın mesafelere sunulduğundan büyük miktarda veri transferi söz konusudur.

Bilgisayar ve Internet teknolojilerinin gelişimiyle bu sanayide hızla gelişmekte ve rekabet artmaktadır.

Rekabetin arttığı durumda müşteri memnuniyeti önem kazanır bu nedenle, iş içeriğini daha iyi anlamak, telekominikasyon örnekleri belirlemek, kaynakları daha iyi kullanmak, servisin kalitesini artırmak ve sahtekarlık girişimlerini engellemek gibi hizmetlerde veri madenciliği kullanılır.

2.6 Görsel Veri Madenciliği

Görsel veri madenciliği, saklı ve yararlı bilgileri görsel veri ve bilgi tekniklerini kullanarak geniş veri setlerinden ortaya çıkarır. İnsan görsel sistemi, güçlü, yüksek paralel süreçleme yapabilen ve büyük bir bilgi tabanını içeren sonlandırma mekanizması olan, göz ve beyin tarafından kontrol edilir. Görsel veri madenciliği aslında, verideki veri dağılımları (data distributions), desenler (patterns), kümeler(clusters), ve sınırdışı(outliers) öğelerinin gücünü daha çekici ve etkin bir araç (tool) yapmak için birleştirir.

Görsel veri madenciliği, iki disiplinin kombinasyonu olarak görülebilir: Veri görselleme ve veri madenciliği. Ayrıca, bilgisayar grafikleri, çoklu ortam sistemleri, desen (pattern) tanıma ve yüksek performans hesaplama konularıyla da yakından ilgilidir.

2.7 Bilimsel ve İstatistiksel Veri Madenciliği

Kapsamlı bilimsel veriler için özellikle sayısal veriler üzerinde veri analizi için başvurulan (örn:psikoloji, tıp, elektrik mühendisliği, ekonomi, sosyal bilimler) teknikler vardır.

Ana istatiksel metotlar şunlardır:

- Regresyon
- Genelleştirilmiş doğrusal modeller
- Regrasyon ağaçları (Regression trees)
- Uyuşmazlık analizi (Analysis of variance)
- Karma-etki modeli (Mixed effect model)
- Faktör analizi (Faktor analysis)
- Ayırım analizi (Disciriminant analysis)
- Zaman serileri

- Yaşam analizi (Survival analysis)
- Kalite kontrolü

Kaynaklar

1] Irma Becerra-Fernandez, Stelios H. Zanakis, Steven Walczak, " Knowledge Discovery Techniques for Predicting Country Investment Risk," **Computers& Industrial Engineering,** No: 43, 2002, s.787-800

[2]Pete Chapman, v.d., "CRISP-DM 1.0 Step-by-Step Data Mining Guide," (Çevrimiçi) http://www.crisp-dm.org/CRISPWP-0800.pdf , 12 Aralık 2004, s. 13.

[3] Becker, Ghedini, **a.g.e.**, s. 97.

[4] P.,Yıldıırım,M.,Uludağ,,A.Görür,"Hastane Bilgi Sistemlerinde Veri Madenciliği", Akademik Bilişim 2008, Çanakkale 18 Mart Üniversitesi.