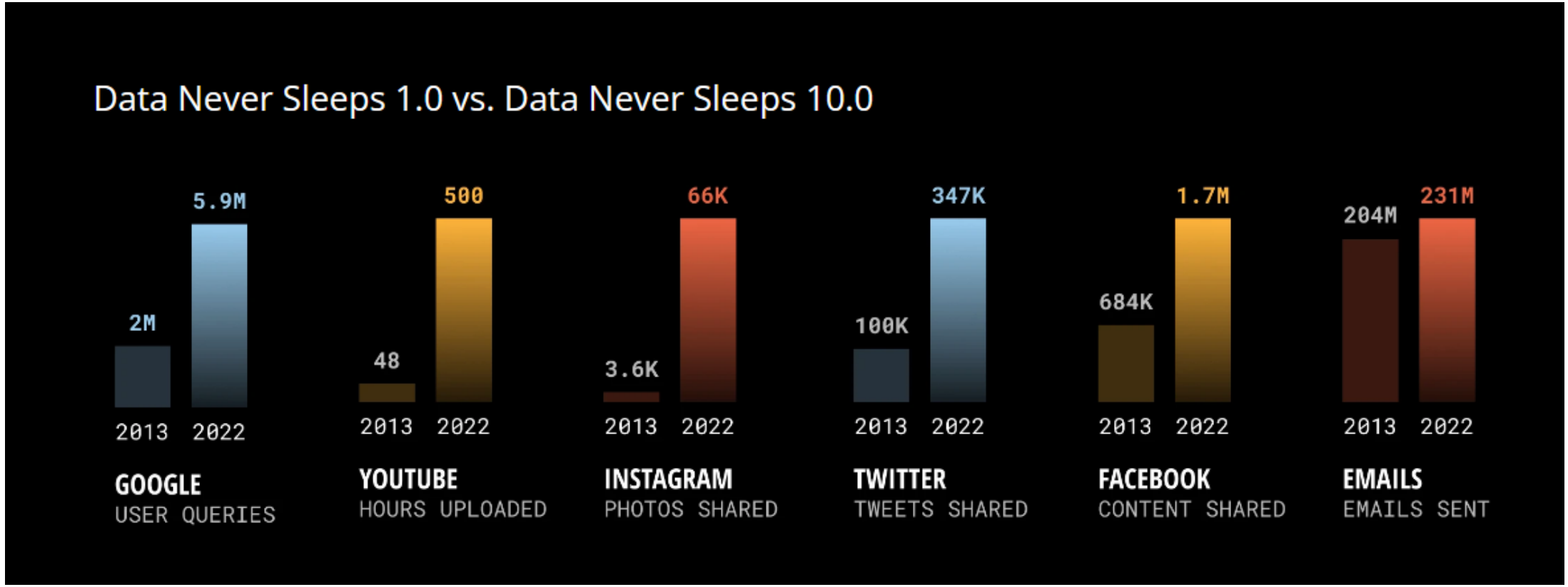
ISE 302 –Veri Madenciliği

DR. ÖĞR. ÜYESİ ESİN AYŞE ZAİMOĞLU

esinzaimoglu@sakarya.edu.tr

Veri Madenciliğine Giriş



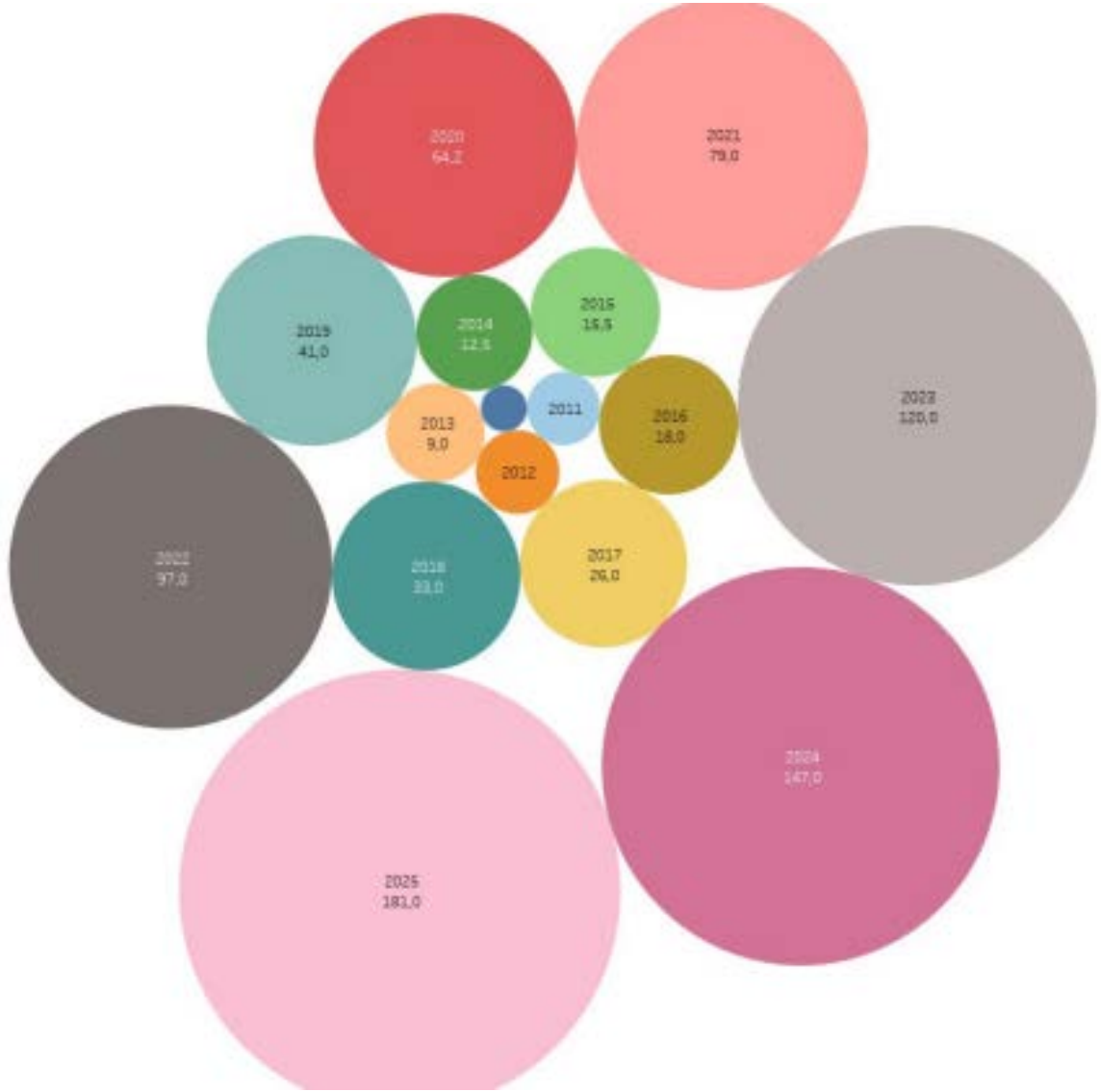
Neden Bu Kadar Önemli?

1995 yılında birincisi düzenlenen Veri Tabanlarında Bilgi Keşfi Konferansı bildiri kitabı aşağıdaki ifadelerle başlamaktadır : “Dünyadaki enformasyon miktarının her 20 ayda bir ikiye katlandığı tahmin edilmektedir. Bu ham veri seli ile ne yapmamız gerekmektedir. İnsan gözleri bunun ancak çok küçük bir kısmını görebilecektir.”

Veri madenciliği son yıllarda çok yoğun bir şekilde bilgi endüstrisinin ilgisini çekmektedir. Bu ilginin temel nedeni ise çok büyük miktarda verilerin elde edilebilmesi ve elde edilen bu ham verilerin hızlı bir şekilde faydalı bilgilere dönüştürülmesi gerekliliğidir.

Veri & Veri Tipleri

Veri : Ölçümlerden ya da istatistiklerden elde edilen gerekli, gereksiz ya da tekrarlı olabilen, anlamlı ifade edebilmesi için işlenmesi gereken sayı ya da karakter dizisi şeklindeki varlık tanımlayıcı ham bilgi.



Kaynak:Statista https://www.statista.com/statistics/871513/worldwide-data-created

Veri & Veri Tipleri

Sayısal(Numerical) Veri :

- Saatte çalan telefon sayısı (Discrete)
- Çocuk boyu boyu 153 cm. (Continuous)

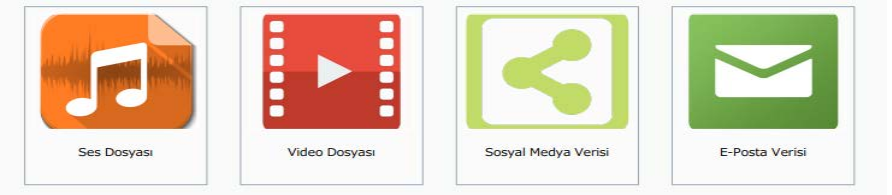
Kategorik(Categorical) Veri:

- Evet/Hayır
- Düşük-OrtaYüksek
- Kabul/Red
- Hasta/Sağlam

Yapısal Veri : Sayı, Metin, Tarih, Para Birimi Tipindeki Veriler ile Diğer uygulamaya özel veri tipleri

Yarı Yapısal Veri : Analiz yapmaya imkan veren belli yapıya sahip veriler (Excel Çalışma Sayfası, XML Dosyası)

Yapısal Olmayan Veri :Tablolarda ya da ilişkisel veri tabanlarında tutulamayan veri tipler



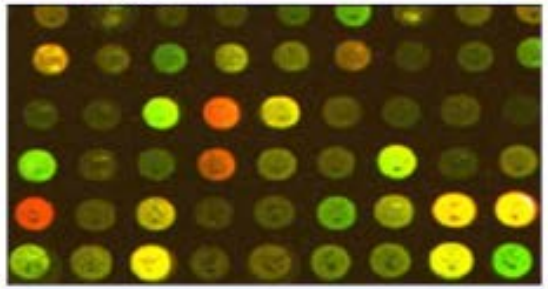
Veri & Veri Tipleri

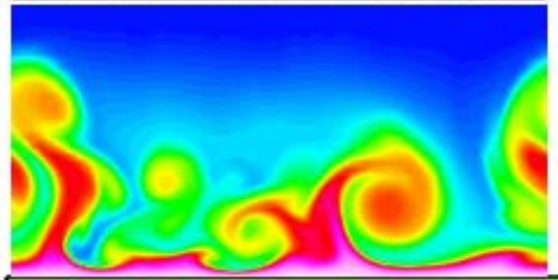
Tip	Gerçek Sıfır	Eşit Aralık	Sıralama	Kategori
Nominal	✗	✗	✗	✓
Ordinal	✗	✗	✓	✓
Interval	✗	✓	✓	✓
Ratio	✓	✓	✓	✓

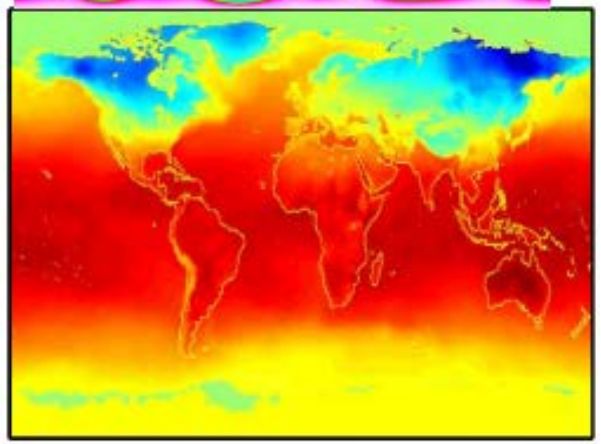
Nominal (Kategorik Ayrık Veri)	Ordinal (Sıralama İçeren Veri)	Interval (Aralık Verisi)	Ratio (Oransal Veri)
• Ülkeler Listesi {«Türkiye», «Gürcistan», «Rusya», «Ukrayna», «Bulgaristan»}	• Hesabınızı ne sıklıkta kontrol ediyorsunuz? Hiç | Her gün | Her hafta | Her ay	• Ölçülen ortam sıcaklığı: C° ya da F°	• Ölçülen ortam sıcaklığı: Kelvin • Yaş, Yükseklik, Genişlik

Neden Veri Madenciliği?- Ticari Bakış

- Çok sayıda veri toplanıyor ve depolanmış
- Web verileri, e-ticaret– departmandaki satın alımlar/marketler
- Banka/Kredi Kartı işlemler
- Bilgisayarlar daha ucuz ve daha güçlü hale geldi
- Rekabet Baskısı Güçlü
- Spesifik bir iş için daha iyi, özelleştirilmiş hizmetler sağlama isteği (örn.Müşteri ilişkileri yönetimi)





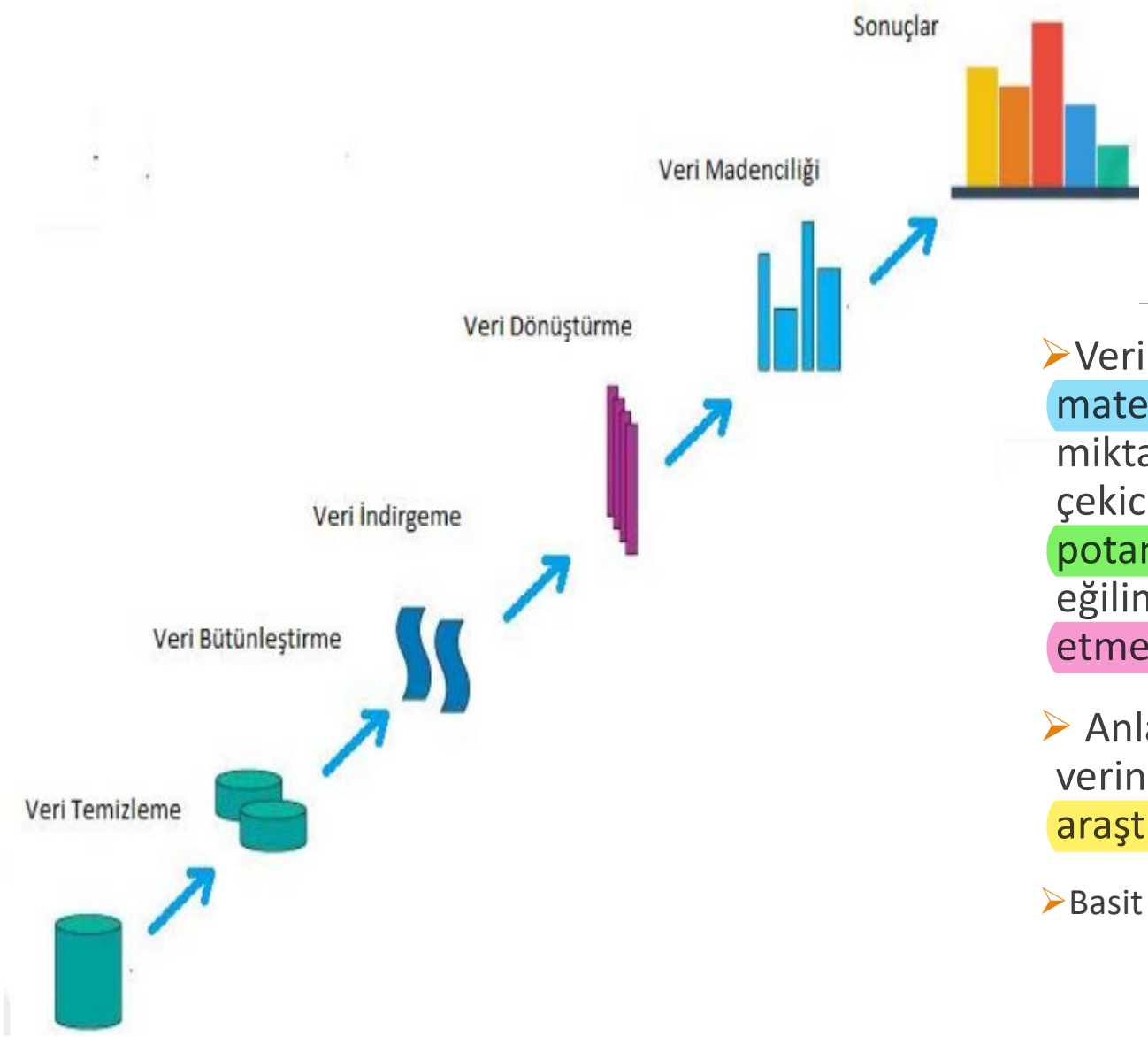


Neden Veri Madenciliği ?- Bilimsel Bakış

- ✓Muazzam hızlı veriler (GB/saat)
  - ✓Toplanan ve saklanan veriler
  - ✓Bir uydudaki uzak sensörler
  - ✓Gökyüzünü tarayan teleskoplar
- ✓ Mikrodiziler üreten gen ifade verileri
- ✓Bilimsel simülasyonlar terabaytlarca veri üretiliyor
- ✓Ham veriler için uygun olmayan geleneksel teknikler
- ✓ Veri madenciliği
  - ✓verileri sınıflandırma ve bölümlere ayırmada
  - ✓Hipotez Oluşturmada bilim adamlarına yardımcı olabilir

Veri Madenciliği Ne Değildir?

- Bir telefon defterinden telefon numarası aramak
- Arama motorlarından anahtar kelime aramak
- Maaşların farklı yaş gruplarına göre dağılım grafiğini çıkarmak
- Bir SQL sorgusuyla veri tabanından sonuç döndürmek
- İlişkisel bir veritabanından çok boyutlu veri küpleri oluşturmak



Veri Madenciliği Nedir?

- Veri Madenciliği, örüntü tanıma, istatistik ve matematiksel yöntemlerin kullanımıyla devasa miktardaki güncel ya da geçmiş veri içerisinden ilgi çekici (önemsiz olmayan, gizli, önceden bilinmeyen, potansiyel olarak kullanışlı) bilginin gelecekteki eğilimleri kestirmek ya da sonraki aşamalarda analiz etmek üzere etkin şekilde çıkarılması sürecidir.
- Anlamlı kalıpları keşfetmek için büyük miktarda verinin otomatik veya yarı otomatik yöntemlerle araştırılması ve analizi
- Basit bir tanım yapmak gerekirse bilgiyi maden-leme işidir.

Problem Tanımı

- teknolojinin gelişimiyle bilgisayar ortamında ve veritabanlarında tutulan veri miktarının artması (terabyte -> petabyte)
  - verinin kolayca toplanabilmesi
  - bu veriyi nasıl kullanacağımızı bilmiyoruz
- saklanan veriden bilgi elde etmek için bu veriyi yorumlamamız gerekiyor
- kullanıcıların beklentilerinin artması(basit veritabanı sorgulama yöntemlerinin yeterli olmaması)
- Veri madenciliği yöntemleri fazla miktardaki veri içinden yararlı bilgiyi bulmak için kullanılır.

Bilgi Keşfinin Aşamaları

- Uygulama alanını inceleme
- Amaca uygun veri kümesi oluşturma
- Veri ayıklama ve önişleme
- Veri azaltma ve veri dönüşümü
  - incelemede gerekli boyutları (özellikleri) seçme, boyutlar arası ilişkiyi belirleme, boyut azaltma,
- Veri madenciliği tekniği seçme
  - Sınıflandırma, eğri uydurma, bağıntı kuralları, demetleme
- Veri madenciliği algoritmasını seçme
- Model değerlendirme ve bilgi sunumu
- Bulunan bilginin yorumlanması

Bilgi Keşfi Örnek: web kayıtları

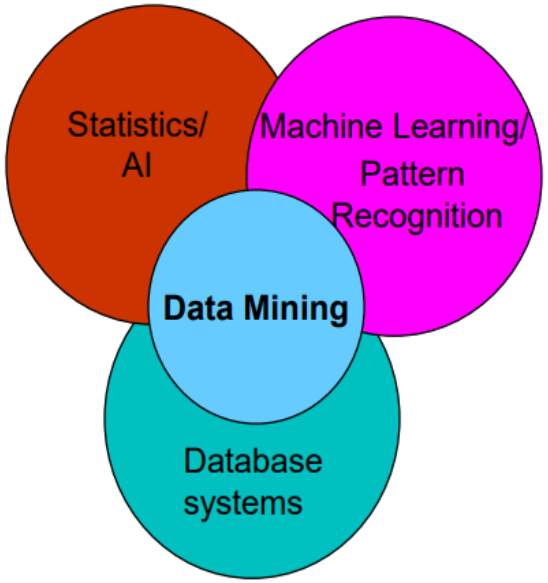
- web sitesinin yapısını inceleme
- verileri seçme: tarih aralığını belirleme
- veri ayıklama, önişleme: gereksiz kayıtları silme
- veri azaltma, veri dönüşümü: kullanıcı oturumları belirleme
- veri madenciliği tekniği seçme: demetleme
- veri madenciliği algoritması seçme: k-ortalama,DBSCAN
- Model değerlendirme/yorumlama: değişik kullanıcı rupları için sıkça izlenen yolu bulma
- Uygulama alanları: öneri modelleri, kişiselleştirme, ön belleğe alma

Veri Madenciliğinin Kökenleri

Makine öğrenimi/yapay zeka, örüntü tanıma, istatistik ve veritabanı sistemlerinden fikirler alır

Geleneksel Teknikler yandaki nedenlerden dolayı uygun olmayabilir:

- Verilerin büyüklüğü
- Verilerin yüksek boyutluluğu
- Verinin heterojen, dağıtılmış doğası



Tan,Steinbach, Kumar Introduction to Data Mining 4/18/2004

Çoğu zaman verilerde kolaylıkla görülemeyen gizli bilgiler vardır.

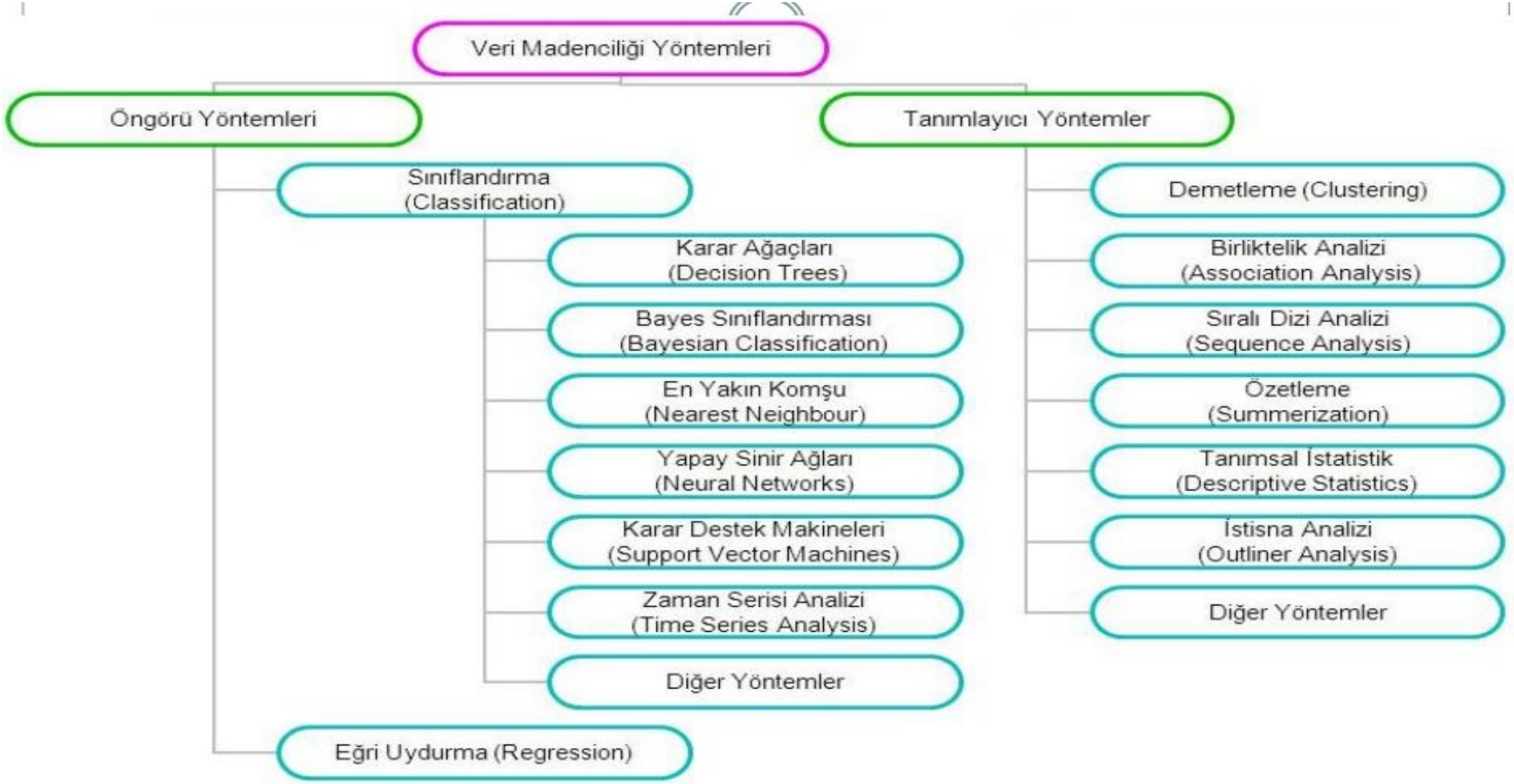
İnsan analistlerin yararlı bilgileri keşfetmesi haftalar alabilir

Verilerin çoğu hiçbir zaman analiz edilmez

Veri Madenciliği

- Mercedes sahibi kişiler ne tür özelliklere sahip?
- Bu müşteri için kredi geri ödemesi ne kadar olmalı?
- Bu işletme için iyi yeni müşteri kimdir?
- Bu makaleye benzeyen başka makaleler var mı?
- Borsa indeksinin değeri önümüzdeki ay ne olabilir?
- Otomobil alan müşterinin hangi öznitelikleri önemlidir?
- Hangi ürün promosyonlarının karlılık üzerindeki etkisi en yüksek?
- En iyi ürün dağıtım kanalı hangisi?
- Semptomlar ve hastalıklar arasındaki ilişkilerin keşfi
- Keşfedilen bu yeni canlı hangi sınıfa ait?
- Market raflarındaki ürünler nasıl dizilmeli?
- ATM'de günlük olarak ne kadar para tutulmalı?

Veri Madenciliği Yöntemleri



İstatistik & Makine Öğrenmesi & Veri Madenciliği

İstatistik

- daha çok teoriye dayalı yaklaşımlar
- bir varsayımın doğruluğunu araştırır

Makine Öğrenmesi

- daha çok sezgisel yaklaşımlar
- öğrenme işleminin başarımını artırmaya çalışır

Veri madenciliği ve bilgi keşfi

- teori ve sezgisel yaklaşımları birleştirir
- bilgi keşfinin tüm aşamalarını gerçekler: veri temizleme, öğrenme, sonucu sunma, yorumlama,...

Aradaki ayrım net değil

Öngörü(Tahmin) Yöntemleri

- Öngörü Yöntemleri sonuçları bilinen verilerden hareket ederek bir model geliştirilmesi ve bu modelden faydalanılarak sonuçları bilinmeyen veriler için tahmin etme amaçlanır.
- Eldeki verilerin benzer özniteliklerine göre bilinmeyen değerlerin tahmin edilmesi
- Yeni bir verinin özniteliklerine göre daha önce belirlenmiş sınıflardan hangisine girebileceğinin belirlenmesi

Örneğin ; Bir bankanın , ilk 3 taksitinden 2 veya daha fazlasını geç ödeyen müşterileri için %75 'inin krediyi geri ödeyememesine dayalı tahmini bu modele örnek olarak verilebilir.

Tanımlayıcı Yöntemler

Tanımlayıcı yöntemler ise mevcut verilerin tanımlanmasını sağlamak olarak ifade edilebilir. Geliri A-B aralığında ve arabası olan çocuklu aileler ile geliri A-B aralığından düşük çocuğu olmayan ailelerin satın alma örüntülerinin birbirine benzerlik göstermesinin saptanması bu modele örnek olarak verilebilir.

Veri madenciliğinin işlevlerine göre;

- Sınıflama ve Regresyon
- Kümeleme
- Birliktelik Kuralları ve Ardışık Zamanlı Örüntüler olarak 3 ana başlıkta incelemek mümkündür. Sınıflama ve Regresyon yöntemleri öngörücü; Kümeleme, Birliktelik Kuralları ve Ardışık Zamanlı Örüntü yöntemleri ise tanımlayıcı yöntemlerdir.

Sınıflama

Veriyi önceden belirlenmiş sınıflardan birine dahil ederek bölümlendirme işlemidir.

Örneğin : Kızlar pembe, erkeklere mavi kıyafet giyer gibi…

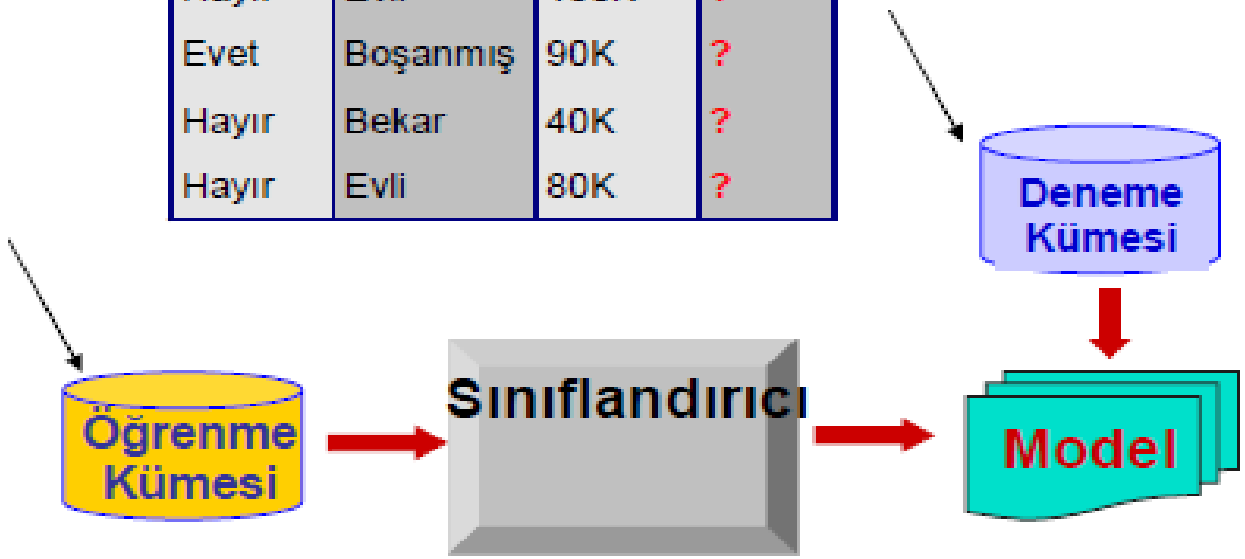
Uygulama Örneği:

- BMW sahibi kişilerin diğerlerine göre bariz özellikleri nelerdir?
- Kredi kartı borcunu ödememe ihtimali olan müşteriler kimlerdir?
- Daha sonra incelenmesi gereken şüpheli işlemler nelerdir?

Sınıflama

Tid	Geri Ödeme	Medeni Durum	Gelir	Dolandırıcı
	kategorik	kategorik	sürekli	sınıf
1	Evet	Bekar	125K	-1
2	Hayır	Evli	100K	-1
3	Hayır	Bekar	70K	-1
4	Evet	Evli	120K	-1
5	Hayır	Boşanmış	95K	1
6	Hayır	Evli	60K	-1
7	Evet	Boşanmış	220K	-1
8	Hayır	Bekar	85K	1
9	Hayır	Evli	75K	-1
10	Hayır	Bekar	90K	1

Geri Ödeme	Medeni Durum	Gelir	Dolandırıcı
Hayır	Bekar	75K	?
Evet	Evli	50K	?
Hayır	Evli	150K	?
Evet	Boşanmış	90K	?
Hayır	Bekar	40K	?
Hayır	Evli	80K	?



Regresyon :

Regresyon analizi, iki ya da daha çok nicel değişken arasındaki ilişkiyi ölçmek için kullanılan analiz metodudur. Kısaca : Veriyi gerçel değerli bir fonksiyona dönüştürmedir.

Örneğin : Vizeden 90 alan dersten geçmeli veya evi ve arabası olan krediyi öder gibi…

Uygulama Örneği:

- Herhangi bir gün için dünya çapında test sürüşü isteklerimiz kaç tane olacak?
- Perakende mağazaları mevsim ve promosyonlara bağlı olarak hangi üründen kaç adet istemeliyiz?
- Yeni bir otomobili satışa çıkarıldığında belli bir perakende satış mağazası bu otomobilden 1 yıl içinde kaç tane satabilir?
- Bu müşteri için kredi geri ödemesi kaç para olacak?
- Telefon, Mobil, TV ürünlerini birlikte satışa çıkarıldığında bunun satış fiyatı ne olacak?

Kümeleme :

Benzer verileri aynı gurupta toplama işlemidir.

Örneğin: Her birine farklı bir ürün grubu kullanılarak kampanya yapılabilecek farklı müşteri grupları oluşturmak.

İçindeki önemli terimlere bakarak birbirine en yakın/benzeyen belgeleri çıkarmak.

Önceden bilinmeyen kritik müşteri özellikleri ve önem derecelerini de ortaya çıkararak gerekli öngörüyü sağlamaktır.

Birliktelik Kuralları :

Eş zamanlı olarak gerçekleşen ilişkilerin tanımlanmasında kullanılır.

Örneğin : Müşteriler sigara satın aldığında %50 ihtimalle çakmakta satın alırlar .

Mercedes sahibi kişilerin başka hangi araç ya da ürünleri var?

Market sepet analizi: HepsiBurada size şu ürünü öneriyor.

Yeterince müşteri tarafından bir arada alınan ürünleri tespit etme gibi

Ardışık Zamanlı Örüntüler :

Birbirleri ile ilişkisi olan ancak birbirini izleyen dönemlerde gerçekleşen ilişkilerin tanımlanmasında kullanılır.

Örneğin : X ameliyatı yapıldığında 15 gün içinde %45 ihtimalle Y enfeksiyonu oluşacaktır.

Müşterinin bir sonraki alışverişinde hangi ürünleri satın alacağının tahmin edilmesi için kullanılan örüntüler

Veri Ambarı

Çok fazla miktarda üzerinde işlem yapılan veri var

- Çoğunlukla farklı veritabanlarında ve farklı ortamlarda
- Veri farklı formatlarda ve yerlerde (heterojen ve dağıtık)

Karar destek birimleri veriye sanal olarak tek bir yerden ulaşabilmeli

- Ulaşım hızlı olmalı

Veri Ambarı Özellikleri

Amaca yönelik

- Müşteri, ürün, satış gibi belli konular için düzenlenebilir
- Verinin incelenmesi ve modellenmesi için oluşturulur
- Konuyla ilgili karar vermek için gerekli olmayan veriyi kullanmayarak konuya basit, özet bakış sağlar

Birleştirilmiş

- Veri kaynaklarının birleştirilmesiyle oluşturulur
- Canlı veri tabanları, dosyalar
- Veri temizleme ve birleştirme teknikleri kullanılır
- Değişik veri kaynakları arasındaki tutarlılık sağlanır

Veri Ambarı Özellikleri

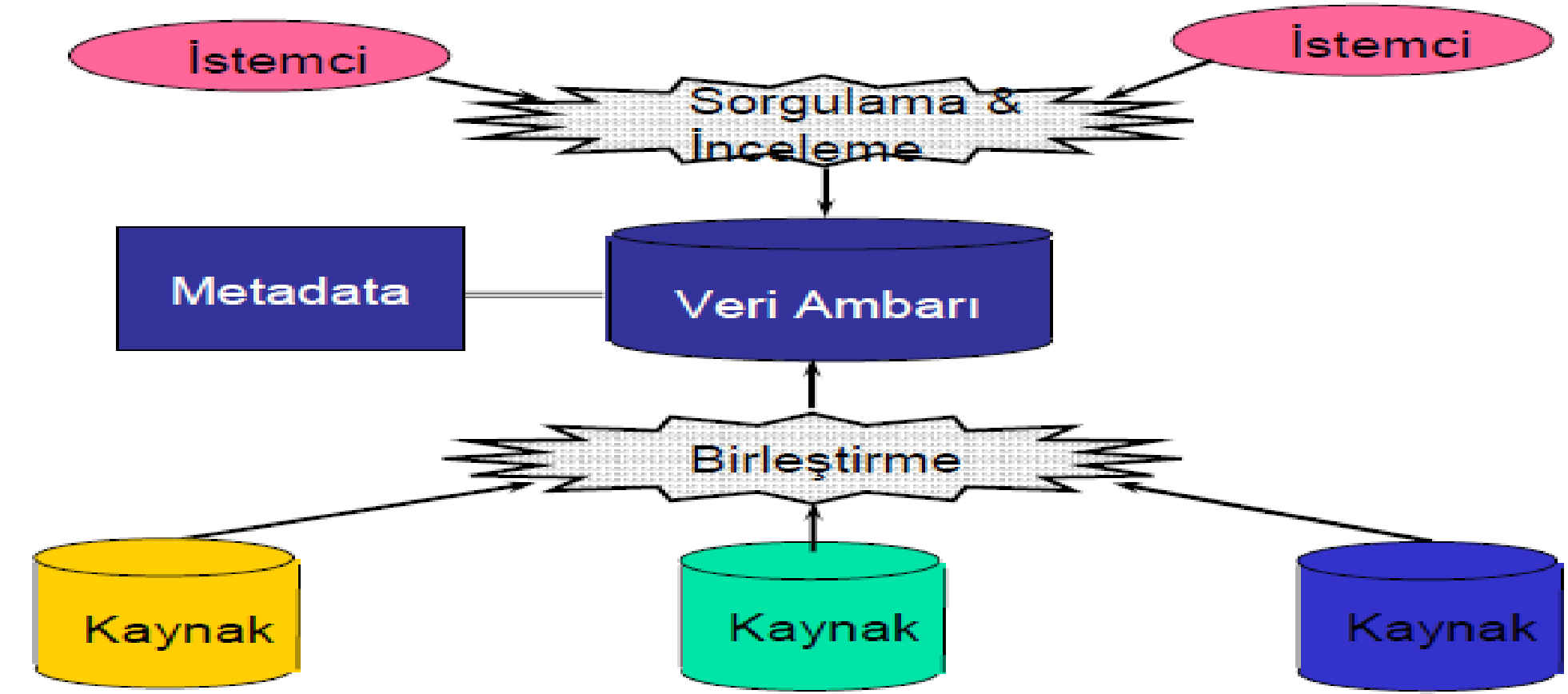
Zaman değişkeni canlı veri tabanlarına göre daha uzundur

- Canlı veri tabanları: Güncel veriler bulunur (en çok geçmiş 1 yıl)
- Veri ambarları: Geçmiş hakkında bilgi verir (geçmiş 5-10 yıl)

Değişken Değil

- Canlı veritabanlarından alınmış verinin fiziksel olarak başka bir ortamda saklanması
- Canlı veritabanlarındaki değişimin veri ambarlarını etkilememesi

Veri Ambarı Mimarisi



Veri Madenciliği & OLAP

OLAP (**O**n-**L**ine **A**nalytical **P**rocessing)

- Veri ambarlarının üzerinde çalışan işlevleri temsil eder
- Veriyi inceleme ve karar verme
- OLTP (**O**n-**L**ine **T**ransaction **P**rocessing) saatler sürebilen işlemler

OLAP avantajları

- Daha geniş kapsamlı sonuçlar
- Daha kısa süreli işlem

OLAP dezavantajları

- Kullanıcı neyi nasıl soracağını bilmesi gerekiyor
- Genelde veriden istatistiksel inceleme yapmak için kullanılır.

OLAP NE sorusuna cevap verir, veri madenciliği NEDEN sorusuna cevap verir.

Veri Madenciliğinde Sorunlar

Güvenlik ve sosyal haklar

- Kişilere ait verilerin toplanarak, kişilerden habersiz ve izinsiz olarak kullanılması
- Gizlilik ve veri madenciliği politikalarının düzenlenmesi

Kullanıcı Arabirimi

- Sonucun anlaşılabilir ve yorumlanabilir hale getirilmesi
- Bilginin sunulması

Etkileşim

- Veri madenciliği ile elde edilen bilginin kullanılması
- Veri madenciliği yöntemine müdahele etmek
- Veri madenciliği yönteminin sonucuna müdahele etmek

Veri Madenciliğinde Sorunlar

Veri madenciliği yöntemi

- Farklı tipte veriler üzerinde çalışabilme
- Farklı seviyelerde kullanıcı ile etkileşim halinde olabilme
- Uygulama ortamı bilgisini kullanabilme
- Veri madenciliği ile elde edilen sonucu anlaşılır şekilde sunabilme
- Gürültülü ve eksik veri ile çalışabilme (ve iyi sonuç verebilme)
- Değişen veya eklenen verileri kolayca kullanabilme
- Örüntü değerlendirme: önemli örüntüleri bulma

Veri Madenciliğinde Sorunlar

Başarım ve ölçeklenebilirlik

- Zaman karmaşıklığı ve yer karmaşıklığı kabul edilebilir
- Örnekleme yapabilme

- Paralel ve dağıtık yöntemler
  - Artımlı veri madenciliği
  - Parçala ve çöz

Uyarı :

- Veri madenciliği yöntemleri bilinçsiz olarak kullanılmamalı
- Veri madenciliği yöntemleri geçmiş olaylara bakarak örüntüler bulur: Gelecekteki olaylar geçmiştekilerle aynı değildir.
- İlişkiler her zaman nedenleri açıklamaz

Bilişim Sistemleri

İstenmeyen web içerikleri ve mesajların belirlenmesi ve filtrelenmesi

Bilgisayar ağ güvenlik köprülerinin tespiti ve korunması

Sektörel Uygulama Örnekleri

Perakendecilik ve Lojistik

Market-sepet analizi Lojistik optimizasyonu için farklı ürün tiplerine ait tüketim seviyelerinin tahmini

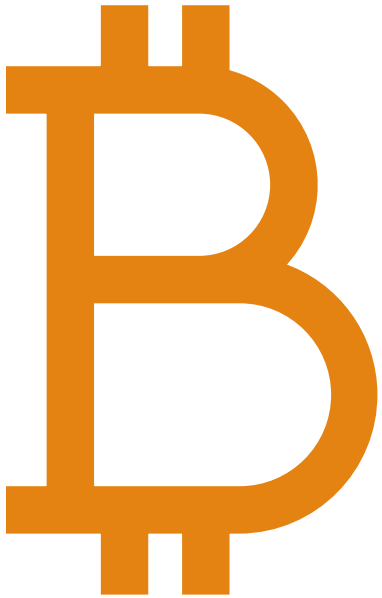
Tedarik zincirindeki ilginç örüntülerin keşfi

Sektörel Uygulama Örnekleri

Sektörel Uygulama Örnekleri

Borsa ve Menkul Kıymetler

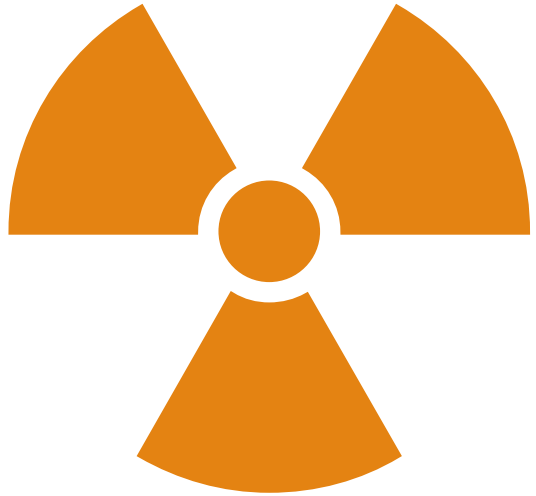
- Belirli hisse fiyatlarının ne kadar ve ne zaman değişeceğinin tahmini
- Sermaye dalgalanmalarının yönü ve oranının tahmini
- Bazı olaylar ve konuların pazardaki hareketliliğe etkisinin değerlendirilmesi
- Menkul kıymetler ticaretindeki şüpheli aktivitelerin tespiti ve önlenmesi



Sektörel Uygulama Örnekleri

Güvenlik ve Hukuk

- Suç ve terörizm ile ilgili örüntülerin tespit edilmesi
- Biyolojik ve kimyasal saldırıların tespiti ve ortadan kaldırılması
- Bilgi altyapısına yönelik kötü niyetli atakların tespiti ve durdurulması

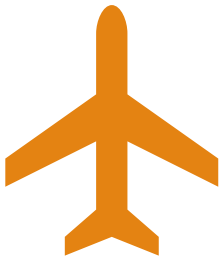


Eğlence

Prime-time’da gösterilecek programlara ve reklamlara nerede yer verilmesi gerektiğine karar verilmesi

Filmlerin finansal başarısının tahmini

Sektörel Uygulama Örnekleri



Seyahat

Farklı hizmetlerin (uçak bilet tipleri, oda tipleri, araç kiralama) satış tahmini

En karlı müşterilerin tespiti ve özelleştirilmiş hizmetlerin sağlanması

Sektörel Uygulama Örnekleri

Sektörel Uygulama Örnekleri

			
Bankacılık ve Sigortacılık	Kredi kartı ve sigorta dolandırıcılıklarının tespiti,	Kredi kartı harcamalarına göre müşteri gruplarının belirlenmesi,	Kredi skoru hesaplama Yeni sigorta poliçesi talep edecek müşterilerin tahmin edilmesi

Sektörel Uygulama Örnekleri

Sektörel Uygulama Örnekleri

Tıp

Klinik testler ile hastalıkların erken teşhisi

Hastalıkların teşhisi için görüntü analizi

Semptomlar ve hastalıklar arasındaki ilişkilerin keşfi

Sektörel Uygulama Örnekleri

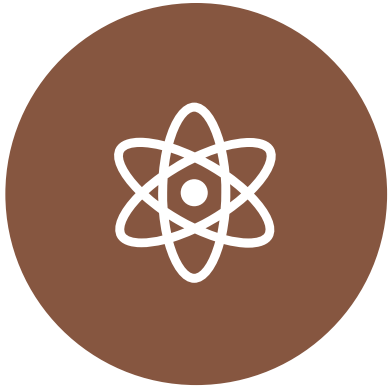
Eczacılık

Yeni ilaçların belirlenmesi

Eczane açılabilecek yerlerin tespit edilmesi

Sektörel Uygulama Örnekleri

BİLİMSEL ANALİZ



ALT GALAKSİ KÜMELERİNİN İNCELENİP YENİ GALAKSİLERİN TESPİT EDİLMESİ

KEŞFEDİLEN YENİ CANLI TÜRLERİNİN SINIFLANDIRILMASI