Veri Madenciliği

Ders Notları - 1

İçerik

- **Veritabanlarında Bilgi Keşfi Süreci**
- Veri Madenciliği

Veri Tabanlarında Bilgi Keşfi Süreci

- Problemin Tanımlanması,
- Verilerin Hazırlanması,
- Modelin Kurulması ve Değerlendirilmesi,
- Modelin Kullanılması,
- Modelin İzlenmesi

Veritabanlarında Bilgi Keşfi Süreci

Decision
- A ürününün Z bölgesinde reklamını yap.
- P profil ailelerine mail ilavesi
- C türü müşterilere yeni telefon hediye et

"nasıl"

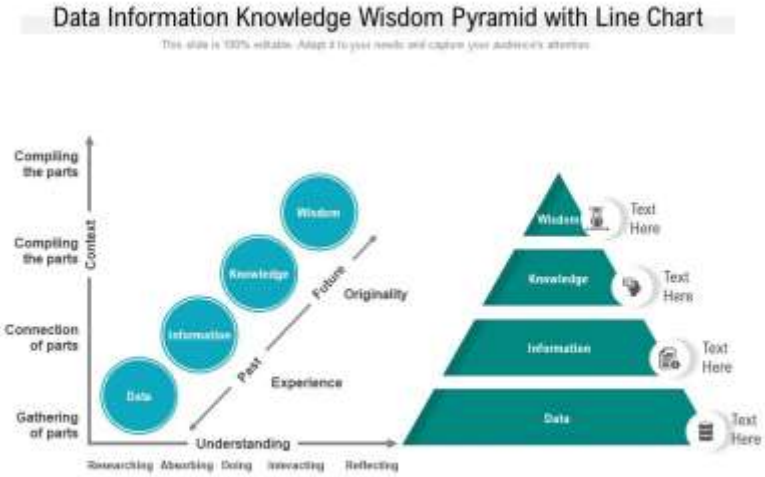
Knowledge
- Belli Y miktarında A ürünü Z bölgesinde kullanılır ,
- Y sınıfı müşteriler D periyodu boyunca x% C kullanırlar

"Niçin bir şeyler yapmalıyım? En iyisi ne? "
- Bilgileri aktif olarak kullanmaya başladığımızda

"kim, ne, ne zaman, nerede"

Information
- X, Z' de yaşar
- S, Y yaşındadır
- X ve S hareket etti
- W' nin Z' de parası var

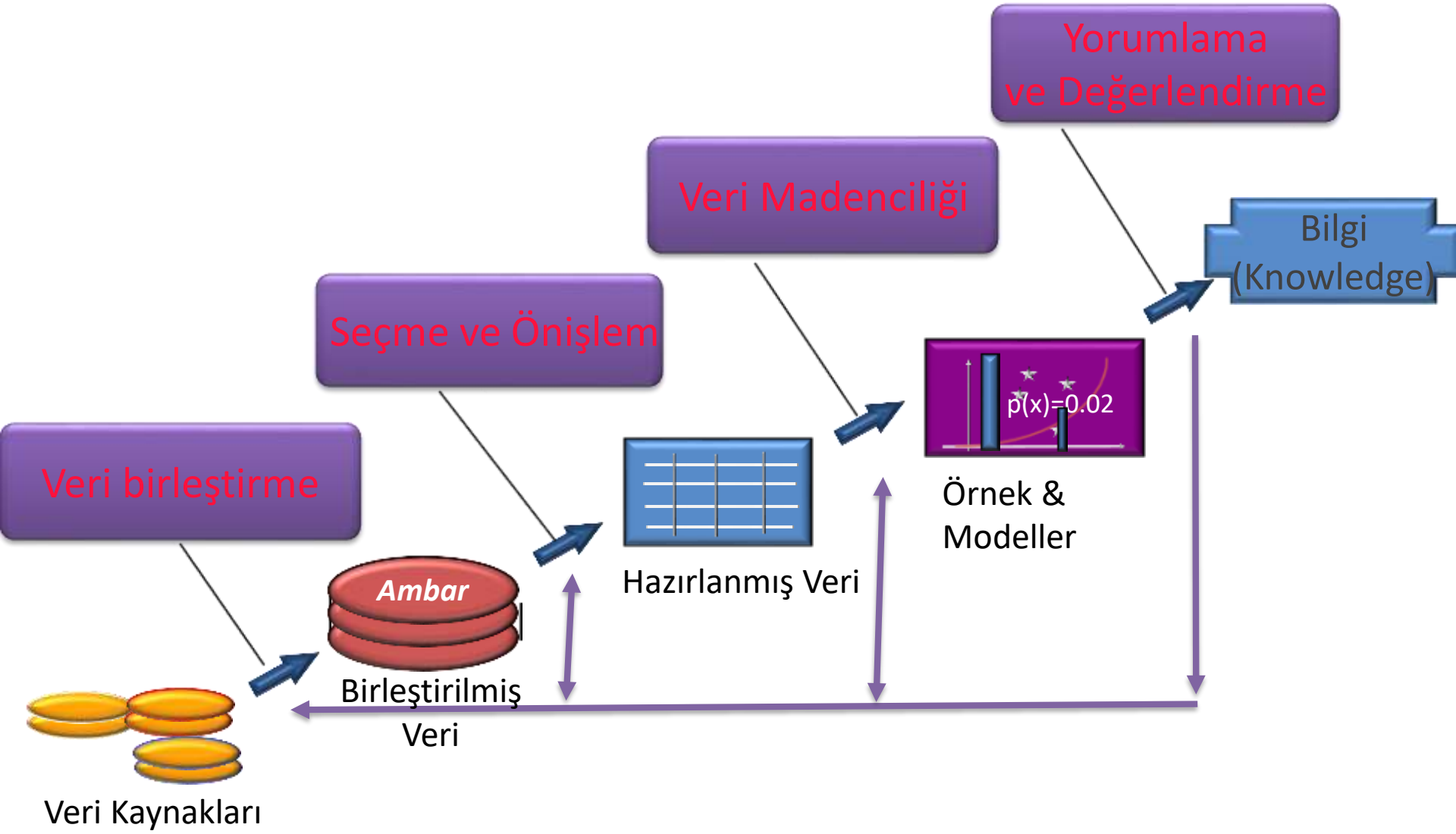
Data
- Müşteri data
- Depo data
- Nüfus data
- Coğrafik data



- 50 sayısının 50 metre olduğunu söylemek “data”
- 50 metrenin uzun veya kısa olduğunu sorgulamak “information”
- Tecrübeleri de katarak uygun uzunluğun bulunması “knowledge”
- Uygun uzunluğu çizen veya uygulayan kişi ya da kişiler de “wisdom”

«Data» ve «information» kavramlarını geçmişe bakmaya benzetirsek, «wisdom» ve «knowledge» kavramlarını da şimdi yaptıklarımız ve gelecekte yapmak istediklerimizle ilişkilendirebiliriz.

VTBK Süreci


Yorumlama
ve Değerlendirme
Veri Madenciliği
Bilgi
(Knowledge)
Seçme ve Önişlem
p(x)=0.02
Veri birleştirme
Örnek &
Modeller
Hazırlanmış Veri
Ambar
Birleştirilmiş
Veri
Veri Kaynakları

Veri Madenciliği ve İş Zekası

1. Problemin Tanımlanması

- Veri Madenciliği (VM) çalışmalarında başarılı olmanın ilk şartı, uygulamanın hangi işletme amacı için yapılacağının açık bir şekilde tanımlanmasıdır. İlgili işletme amacı işletme problemi üzerine odaklanmış ve açık bir dille ifade edilmiş olmalı, elde edilecek sonuçların başarı düzeylerinin nasıl ölçüleceği tanımlanmalıdır. Ayrıca yanlış tahminlerde katlanılacak olan maliyetlere ve doğru tahminlerde kazanılacak faydalara ilişkin tahminlere de bu aşamada yer verilmelidir

2. Verilerin Hazırlanması

- Modelin kurulması aşamasında ortaya çıkacak sorunlar, bu aşamaya sık sık geri dönülmesine ve verilerin yeniden düzenlenmesine neden olacaktır. Bu durum verilerin hazırlanması ve modelin kurulması aşamaları için, bir analizcinin veri keşfi sürecinin toplamı içerisinde enerji ve zamanının % 50 - % 85'ini harcamasına neden olmaktadır
- Verilerin hazırlanması aşaması kendi içerisinde toplama, değer biçme, birleştirme ve temizleme, seçme ve dönüştürme adımlarından meydana gelmektedir

3. Modelin Kurulması ve Değerlendirilmesi

- Tanımlanan problem için en uygun modelin bulunabilmesi, olabildiğince çok sayıda modelin kurularak denenmesi ile mümkündür. Bu nedenle veri hazırlama ve model kurma aşamaları, en iyi olduğu düşünülen modele varılıncaya kadar yinelenen bir süreçtir.
- Model kuruluş süreci denetimli (Supervised) ve denetimsiz (Unsupervised) öğrenimin kullanıldığı modellere göre farklılık göstermektedir

4. Modelin Kullanılması

- Kurulan ve geçerliliği kabul edilen model doğrudan bir uygulama olabileceği gibi, bir başka uygulamanın alt parçası olarak kullanılabilir. Kurulan modeller risk analizi, kredi değerlendirme, dolandırıcılık tespiti gibi işletme uygulamalarında doğrudan kullanılabileceği gibi, promosyon planlaması simülasyonuna entegre edilebilir veya tahmin edilen envanter düzeyleri yeniden sipariş noktasının altına düştüğünde, otomatik olarak sipariş verilmesini sağlayacak bir uygulamanın içine gömülebilir

5. Modelin İzlenmesi

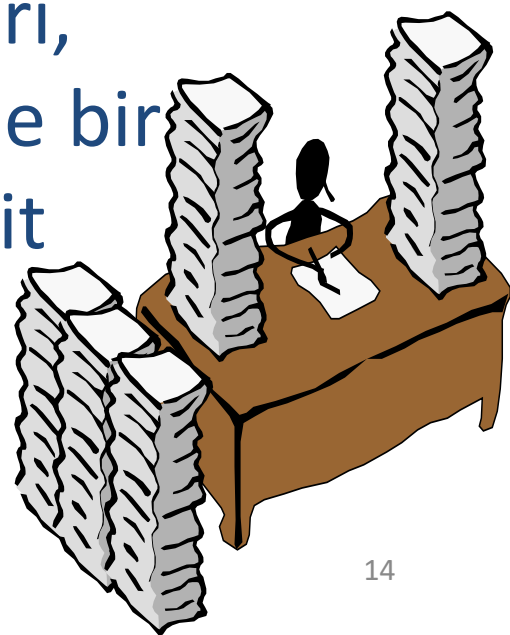
- Zaman içerisinde bütün sistemlerin özelliklerinde ve dolayısıyla ürettikleri verilerde ortaya çıkan değişiklikler, kurulan modellerin sürekli olarak izlenmesini ve gerekiyorsa yeniden düzenlenmesini gerektirecektir. Tahmin edilen ve gözlenen değişkenler arasındaki farklılığı gösteren grafikler model sonuçlarının izlenmesinde kullanılan yararlı bir yöntemdir.

İçerik

- Veritabanlarında Bilgi Keşfi Süreci
- **Veri Madenciliği**

- Her 20 ayda bir dünyadaki bilgi miktarının 10 katına çıktığı tahmin edilmektedir.
- Dünya gözlem uyduları bir günde yüzlerce petabayt veri üretmektedir.

- Bilgi toplama ve toplanan bilgileri saklama olanaklarında büyük bir artış.
- Kredi kartı kullanımı, tıbbi test sonuçları, telefon konuşmaları, süper marketlerde bir kerede satın alınan ürünler gibi en basit hareketler bile bilgisayar ortamına kaydedilmektedir



Veri Madenciliği Nedir?

Veri madenciliği;

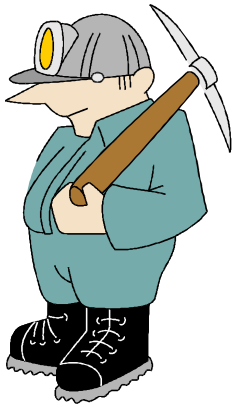
- *veri ambarlarındaki tutulan*
- *çok çeşitli ve çok miktarda veriye dayanarak*
- *daha önce keşfedilmemiş bilgileri*
- *ortaya çıkarmak,*
- *bunları karar verme ve*
- *eylem planını gerçekleştirmek için kullanma sürecidir.*

Büyük miktarda veri içinden, gelecekle ilgili tahmin yapmamızı sağlayacak bağıntı ve kuralların aranmasıdır.

Sayısal verinin miktarı, son 10 yılda bir patlama yaşayarak tahminlerin dışında bir artış göstermiştir. Buna karşılık, bilim adamlarının, mühendislerin ve analistlerin sayısı değişmemektedir.

- Geniş hacimli ve çok boyutlu VM için yeni algoritma ve sistemlerin geliştirilmesi,
- Yeni veri tiplerinin madenciliği için yeni algoritma, teknik ve sistemlerin geliştirilmesi,
- Dağıtık VM için algoritma, protokol ve altyapıların geliştirilmesi,
- Mevcut VM sistemlerinin kullanımının ilerletilip geliştirilmesi,
- VM için özel gizlilik ve güvenlik modellerinin geliştirilmesi.

VM, verilerin içerisindeki desenlerin, ilişkilerin, değişimlerin, düzensizliklerin, kuralların ve istatistiksel olarak önemli olan yapıların yarı otomatik olarak keşfedilmesidir.



Temel olarak VM, veri setleri arasındaki desenlerin ya da düzenin, verinin analizi ve yazılım tekniklerinin kullanılması ile ilgilidir. Veriler arasındaki ilişkiyi, kuralları ve özellikleri belirlemekten bilgisayar sorumludur. Amaç, daha önceden fark edilmemiş veri desenlerini tespit edebilmektir

Örnek Uygulamalar

- Bağıntı

  “Çocuk bezi alan müşterilerin 30%’u bira da alır.” (*Basket Analysis*)

- Sınıflandırma

  “Genç kadınlar küçük araba satın alır; yaşlı, zengin erkekler ise büyük, lüks araba satın alır.”

- Regresyon

  Kredi skorlama (*Application Scoring*)

Örnek Uygulamalar

- Zaman içinde Sıralı Örüntüler

  "İlk üç taksidinden iki veya daha fazlasını geç ödemiş olan müşteriler %60 olasılıkla krediyi geriye ödeyemiyor." (*Behavioral scoring, Churning*)

- Benzer Zaman Sıraları

  "X şirketinin hisselerinin fiyatları Y şirketinin fiyatlarıyla benzer hareket ediyor."

Örnek Uygulamalar

- İstisnalar (Fark Saptanması)

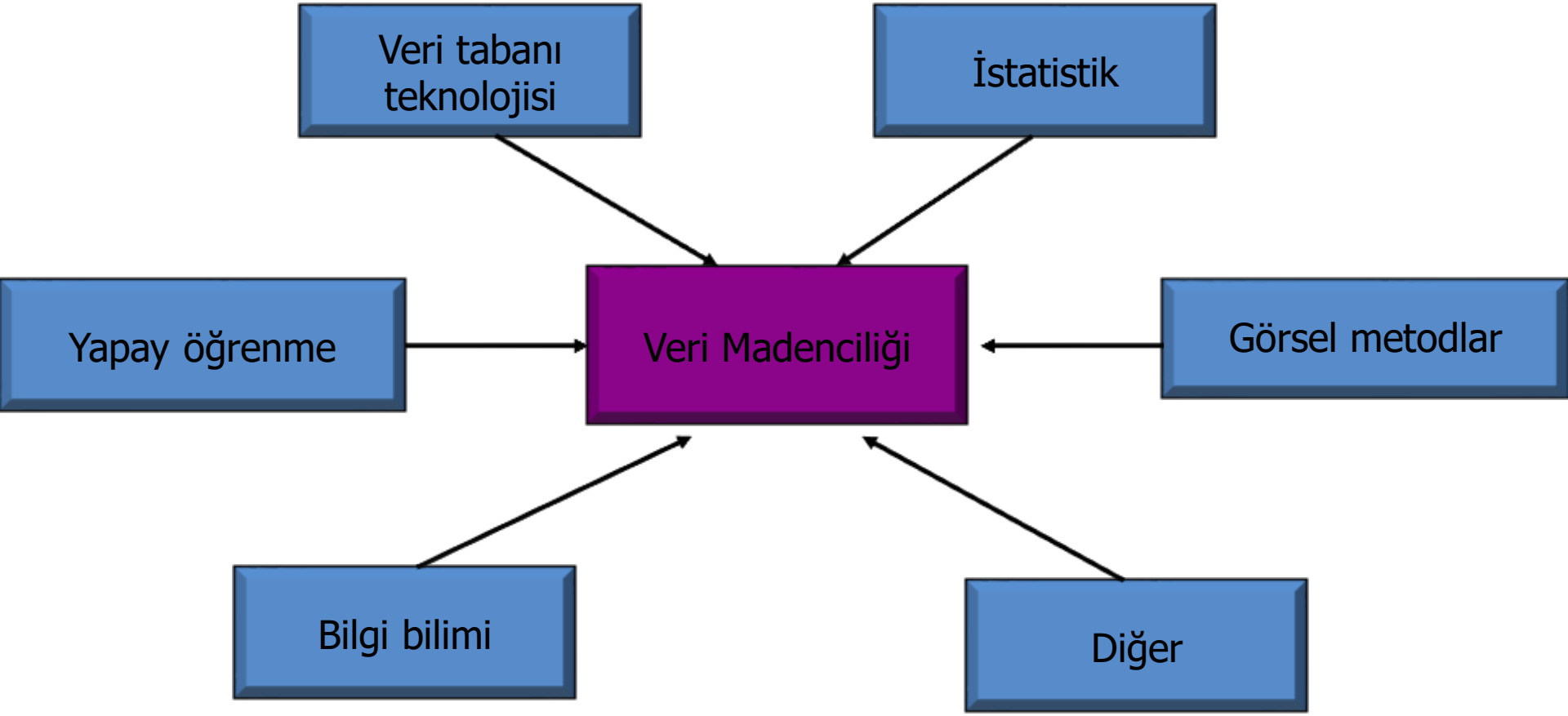
  "Normalden farklı davranış gösteren müşterilerim var mı?"

  Fraud detection

- Döküman Madenciliği (Web Madenciliği)

  "Bu arşivde (veya internet üzerinde) bu dökümana benzer hangi dökümanlar var?"

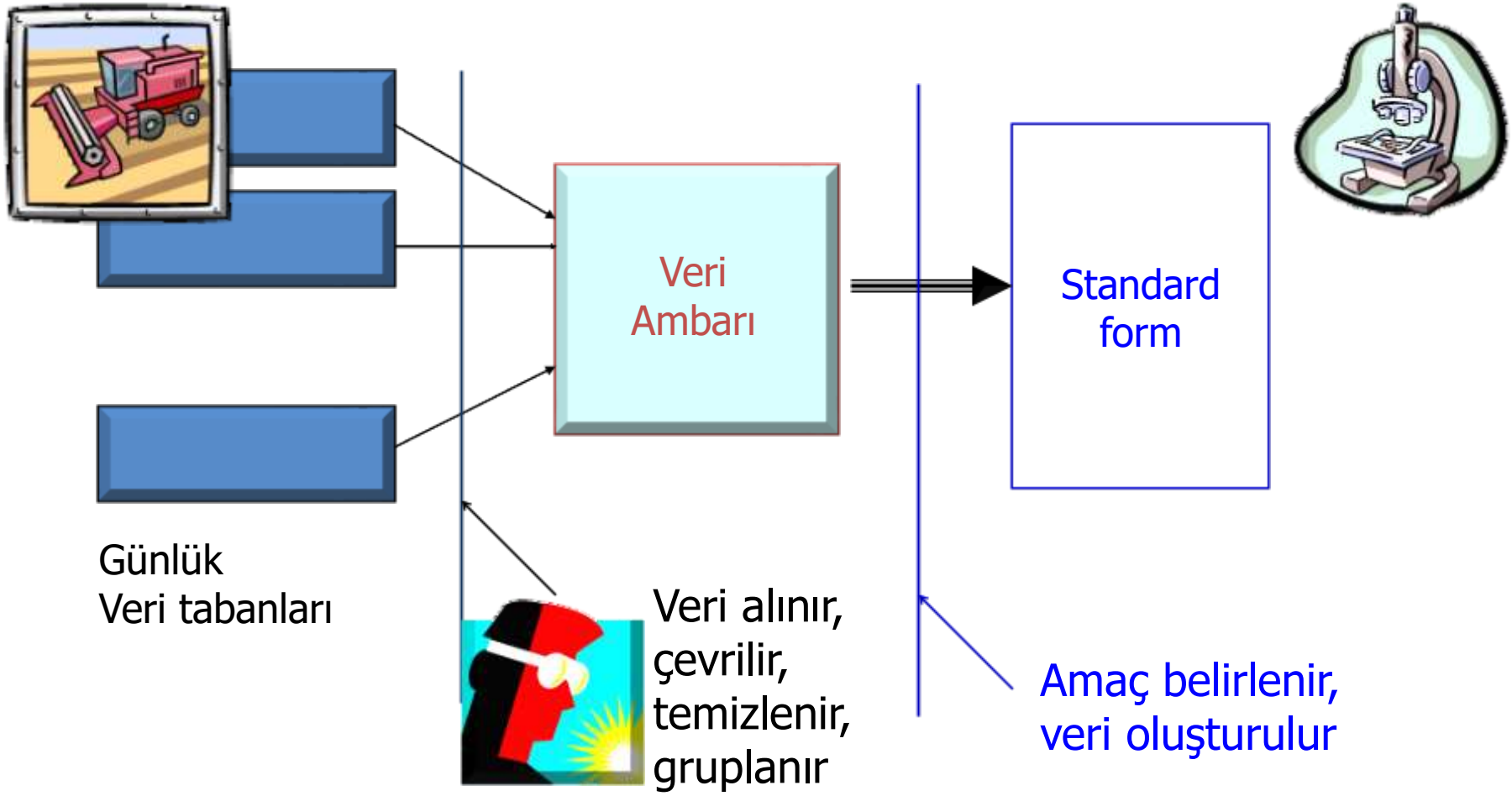
Veri Madenciliği ile diğer disiplinler arasındaki ilişki

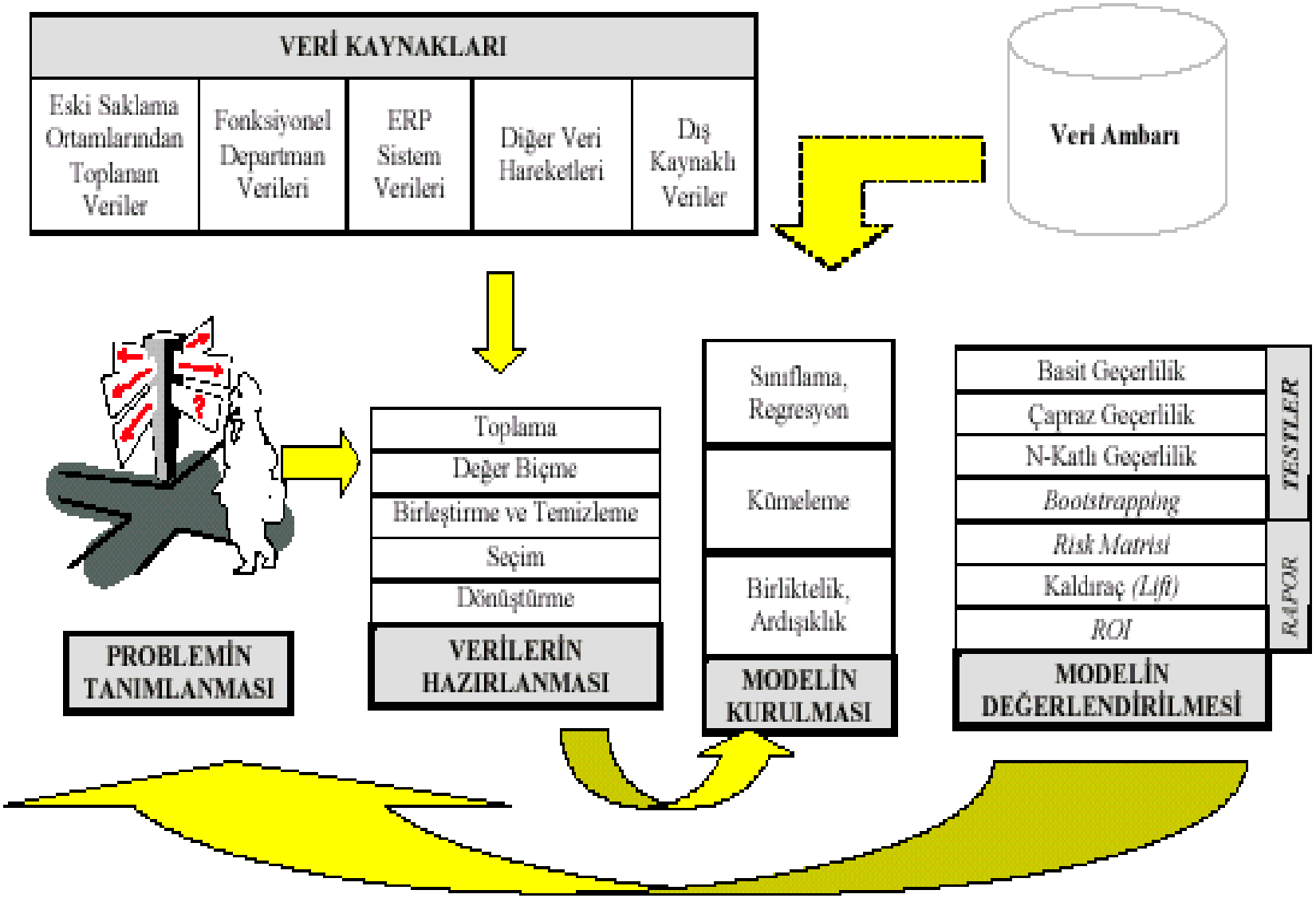


Etkin bir VM uygulayabilmek için

- *Farklı tipteki verileri ele alma*
- *VM algoritmasının etkinliği ve ölçeklenebilirliği*
- *Sonuçların yararlılık, kesinlik ve anlamlılık kıstaslarını sağlaması*
- *Keşfedilen kuralların çeşitli biçimlerde gösterimi*
- *Farklı birkaç soyutlama düzeyi ve etkileşimli VM*
- *Farklı ortamlarda yer alan veri üzerinde işlem yapabilme*
- *Gizlilik ve veri güvenliğinin sağlanması*

Ambardan Madene





MODELİN İZLENMESİ

Adımlar

- Veri Seçimi *(Data Selection)*: Bu adım birkaç veri kümesini birleştirerek, sorguya uygun örneklem kümesini elde etmeyi gerektirir.
- Veri Temizleme ve Ön işleme *(Data Cleaning & Preprocessing)*: Seçilen örneklemde yer alan hatalı tutanakların çıkarıldığı ve eksik nitelik değerlerinin değiştirildiği aşamadır. Bu aşama keşfedilen bilginin kalitesini arttırır.
- Veri İndirgeme *(Data Reduction)*: Seçilen örneklemden ilgisiz niteliklerin atıldığı ve tekrarlı tutanakların ayıklandığı adımdır. Bu aşama seçilen VM sorgusunun çalışma zamanını iyileştirir.
- Veri Madenciliği *(Data Mining)*: Verilen bir VM sorgusunun (sınıflama, kümeleme, birliktelik, vb.) işletilmesidir.
- Değerlendirme *(Evaluation)*: Keşfedilen bilginin geçerlilik, yenilik, yararlılık ve basitlik kıstaslarına göre değerlendirilmesi aşamasıdır

Uygulama Alanları

- *Pazarlama*
  - Müşterilerin satın alma örüntülerinin belirlenmesi,
  - Müşterilerin demografik özellikleri arasındaki bağlantıların bulunması,
  - Posta kampanyalarında cevap verme oranının artırılması,
  - Mevcut müşterilerin elde tutulması, yeni müşterilerin kazanılması,
  - Pazar sepeti analizi *(Market Basket Analysis)*
  - Müşteri ilişkileri yönetimi *(Customer Relationship Management)*
  - Müşteri değerlendirme *(Customer Value Analysis)*
  - Satış tahmini *(Sales Forecasting).*

Uygulama Alanları

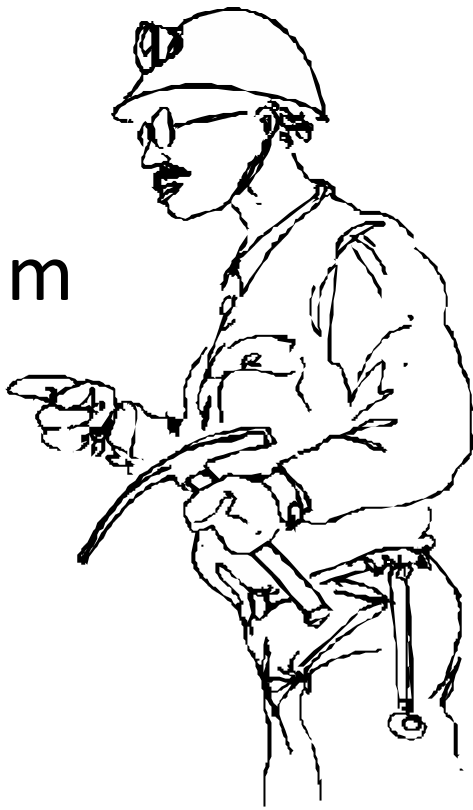
- *Bankacılık*
  - Farklı finansal göstergeler arasında gizli korelasyonların bulunması,
  - Kredi kartı dolandırıcılıklarının tespiti,
  - Kredi kartı harcamalarına göre müşteri gruplarının belirlenmesi,
  - Kredi taleplerinin değerlendirilmesi.
  - Gelecek tahmini (hisse senedi fiyatları...)

Uygulama Alanları

- *Sigortacılık*
    - Yeni poliçe talep edecek müşterilerin tahmin edilmesi,
    - Sigorta dolandırıcılıklarının tespiti,
    - Riskli müşteri örüntülerinin belirlenmesi.

Yeni Uygulamalar

- İş ve Elektronik Ticaret Verileri
- Bilimsel, Mühendislik ve Sağlık Bakım Verileri
- Web Verileri

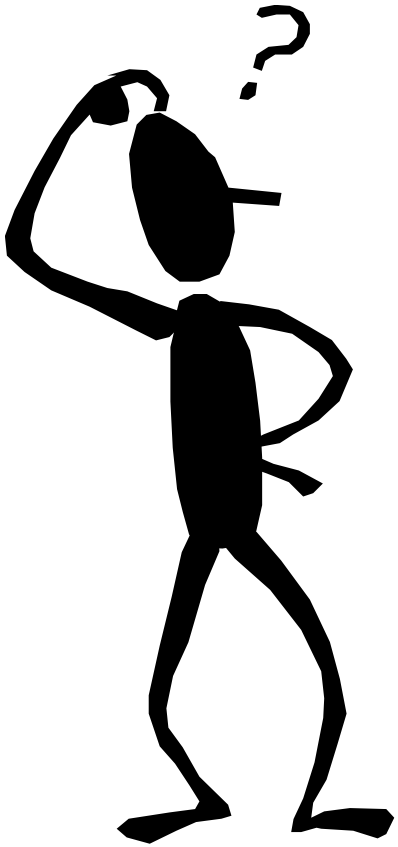


Veri Madenciliğini Etkileyen Eğilimler

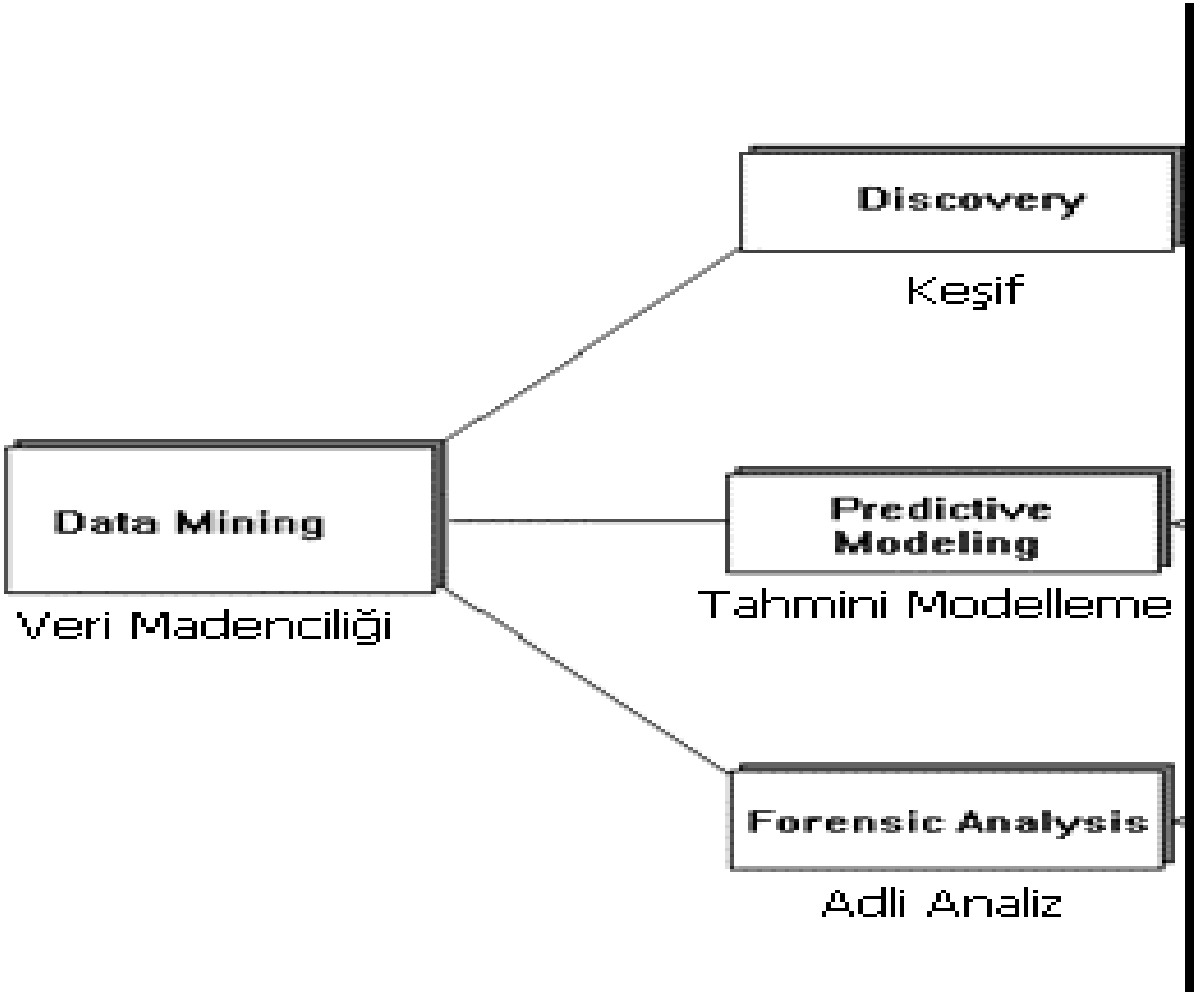
- Veri
- Donanım
- Bilgisayar Ağları
- Bilimsel Hesaplamalar
- Ticari Eğilimler

Veri Madenciliğinde Karşılaşılan Problemler

- **Veritabanı Boyutu**
- **Gürültülü Veri**
- **Null Değerler**
- **Eksik Veri**
- **Artık Veri**
- **Dinamik Veri**



Veri Madenciliği İşlevleri



Veri Madenciliği Algoritmaları

- **Hipotez Testi Sorgusu**
- **Sınıflama Sorgusu**
- **Kümeleme Sorgusu**
- **Ardışık Örüntüler**
- **Birliktelik Kuralları**

Hipotez Testi Sorgusu

- Hipotez testi sorgusu algoritması, doğrulamaya dayalı bir algoritmadır. Bir hipotez öne sürülür ve seçilen veri kümesinde hipotez doğruluğu test edilir. Öne sürülen hipotez genellikle belirli bir örüntünün veritabanındaki varlığıyla ilgili bir tahmindir. Bu tip bir analiz özellikle keşfedilmiş bilginin genişletilmesi veya rötuşlanması işlemleri sırasında yararlıdır.
- Hipotez ya mantıksal bir kural ya da mantıksal bir ifade ile gösterilir. Her iki biçimde de seçilen veritabanındaki nitelik alanları kullanılır. X ve Y birer mantıksal ifade olmak üzere “IF X THEN Y” biçiminde bir hipotez öne sürülebilir.
- Verilen hipotez seçilen veritabanında doğruluk ve destek kıstasları baz alınarak sistem tarafından sınanır.

Sınıflama Sorgusu

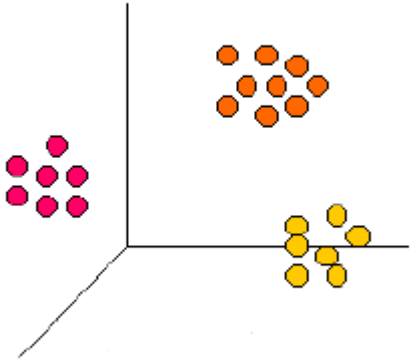
- Sınıflama sorgusu, yeni bir veri elemanını daha önceden belirlenmiş sınıflara atamayı amaçlar.
- Veritabanında yer alan çoklular bir sınıflama fonksiyonu yardımıyla kullanıcı tarafından belirlenmiş ya da karar niteliğinin bazı değerlerine göre anlamlı ayrık alt sınıflara ayırır. Sınıflama algoritması bir sınıfı diğerinden ayıran örüntüleri keşfeder. Sınıflama algoritmaları iki şekilde kullanılır.
  - Karar Değişkeni ile Sınıflama
  - Örnek ile Sınıflama
- Yaygın kullanım alanları, banka kredisi onaylama işlemi, kredi kartı sahteciliği tesbiti ve sigorta risk analizidir.

Kümeleme Sorgusu

- Kümeleme algoritması veritabanını alt kümelere ayırır. Her bir kümede yer alan elemanlar dahil oldukları grubu diğer gruplardan ayıran ortak özelliklere sahiptir

Kümeleme modellerinde amaç, şekilde görüldüğü gibi küme üyelerinin birbirlerine çok benzediği, ancak özellikleri birbirlerinden çok farklı olan kümelerin bulunması ve veri tabanındaki kayıtların bu farklı kümelere bölünmesidir.

Yaygın kullanım alanları nüfusbilimi, astronomi vb.dir.



Ardışık Örüntüler

- Ardışık örüntü keşfi, bir zaman aralığında sıklıkla gerçekleşen olaylar kümelerini bulmayı amaçlar.
    - Bir yıl içinde Orhan Pamuk'un "Benim Adım Kırmızı" romanını satın alan insanların %70'i Buket Uzuner' in "Güneş Yiyen Çingene" adlı kitabını satın almıştır.
    - X ameliyatı yapıldığında, 15 gün içinde % 45 ihtimalle Y enfeksiyonu oluşacaktır,
    - İMKB endeksi düşerken A hisse senedinin değeri % 15'den daha fazla artacak olursa, üç iş günü içerisinde B hisse senedinin değeri % 60 ihtimalle artacaktır,
    - Çekiç satın alan bir müşteri, ilk üç ay içerisinde % 15, bu dönemi izleyen üç ay içerisinde % 10 ihtimalle çivi satın alacaktır.
- Bu tip örüntüler perakende satış, telekomünikasyon ve tıp alanlarında yararlıdır.

Birliktelik Kuralları

- Bir alışveriş sırasında veya birbirini izleyen alışverişlerde müşterinin hangi mal veya hizmetleri satın almaya eğilimli olduğunun belirlenmesi, müşteriye daha fazla ürünün satılmasını sağlama yollarından biridir. Bununla birlikte bu teknikler, tıp, finans ve farklı olayların birbirleri ile ilişkili olduğunun belirlenmesi sonucunda değerli bilgi kazanımının söz konusu olduğu ortamlarda da önem taşımaktadır
- Birliktelik kuralları eş zamanlı olarak gerçekleşen ilişkilerin tanımlanmasında kullanılır.
    - *Müşteriler bira satın aldığında, % 75 ihtimalle patates cipsi de alırlar,*
    - *Düşük yağlı peynir ve yağsız yoğurt alan müşteriler, % 85 ihtimalle diyet süt de satın alırlar*
- Yaygın kullanım alanları katalog tasarımı, mağaza ürün yerleşim planı, müşteri kesimleme, telekomünikasyon vb.dir.