VERİ MADENCİLİĞİ

Bölüm I

Öğr. Gör. Merve KESİM ÖNAL

Veri Madenciliğine Giriş

İçinde yaşadığımız bilişim çağında elektronik ortamda mevcut verinin hızlı artışı ve bilginin fazlalaşması sebebiyle öncelikle, genelde Veri Tabanlarında Bilgi Keşfi olarak adlandırılan yeni bir paradigma ortaya çıkmıştır. Daha yaygın bir kullanımla bu alana Veri Madenciliği denilmektedir.

Veri Madenciliği Tanımları

- Veri Madenciliği(Data Mining): Büyük miktarda veri içinden, gelecekle ilgili tahmin yapmamızı sağlayacak bağıntı ve kuralların aranmasıdır. (Knowledge Discovery in Databases)
- Daha önceden bilinmeyen, geçerli ve uygulanabilir bilgilerin geniş veritabanlarından elde edilmesi ve bu bilgilerin işletme kararları verilirken kullanılmasıdır.
- Büyük ölçekli veriler arasından değeri olan bir bilgiyi elde etme işidir.
- Yapısal veritabanlarında depolanmış verilerden geçerli, yeni, potansiyel olarak yararlı ve nihayetinde anlaşılabilir örüntülerin tanımlanması işlemidir.

Bu tanımlamalardan da anlaşıldığı üzere veriler arasındaki ilişkileri ortaya koymak ve gerektiğinde ileriye yönelik tahminlerde bulunmak veri madenciliği çalışmaları sayesinde mümkün olmaktadır. Bunun anlamı, veri madenciliği bir kurumda üretilen tüm verilerin belirli yöntemler kullanarak var olan ya da gelecekte ortaya çıkabilecek gizli bilgiyi ortaya çıkarma süreci olarak değerlendirilmesidir. Bu açıdan bakıldığında veri madenciliği işinin kurumların Karar Destek Sistemleri için önemli bir yere sahip olduğu söylenebilir.

Veri Madenciliği Tanımları

Veri madenciliği çalışmaları, sınıflandırma, ilişki kurma, kümeleme, regresyon, veri özetleme, değişikliklerin analizi, sapmaların tespiti gibi belirli sayıda teknik yaklaşımları içerir.
Veri madenciliği aslında klasik istatistiksel uygulamalara çok benzer. Ancak klasik istatistiksel uygulamalar yeterince düzenlenmiş ve çoğunlukla özet veriler üzerinde çalıştırılır. Ayrıntılı bilgi olsa bile kayıtlar binlerce olabilir. Veri madenciliğinde ise milyonlarca ve hatta milyarlarca veri ve çok daha fazla değişken ile ilgilenilir. Veri sayısı çok olunca da bazı özel algoritmalar gerekli olmuştur.

Veri Madenciliği Tarihçesi

Data Fishing Data Dredging: 1960

istatistikçiler

Data Mining: 1990

veritabanı kullanıcıları, ticari

Knowledge Discovery in Databases (KDD): 1989

Yapay zeka, makine öğrenmesi toplulukları

Data Archaeology, Information Harvesting, Information Discovery, Knowledge Extraction,...

Veri Madenciliği Tarihçesi

Veri madenciliği, kavramsal olarak 1960'lı yıllarda, bilgisayarların veri analiz problemlerini çözmek için kullanılmaya başlamasıyla ortaya çıkmıştır. O dönemlerde, bilgisayar yardımıyla, yeterince uzun bir tarama yapıldığında, istenilen verilere ulaşmanın mümkün olacağı gerçeği kabullenilmiştir. Bu işleme veri madenciliği yerine önceleri veri taraması (data dredging), veri yakalanması (data fishing) gibi isimler verilmiştir.

1990'lı yıllara gelindiğinde Veri Madenciliği ismi, bilgisayar mühendisleri tarafından ortaya atıldı. Bu camianın amacı, geleneksel istatistiksel yöntemler yerine, veri analizinin algoritmik bilgisayar modülleri tarafından değerlendirmesini vurgulamaktı. Bu noktadan sonra bilim adamları veri madenciliğine çeşitli yaklaşımlar getirmeye başladılar. Bu yaklaşımların kökeninde istatistik, makine öğrenmesi (machine learning), veritabanları, otomasyon, pazarlama, araştırma gibi disiplinler ve kavramlar yatmaktaydı.

Veri Madenciliği Tarihçesi

İstatistik, süre gelen zaman içerisinde verilerin değerlendirilmesi ve analizleri konusunda hizmet veren bir yöntemler topluluğuydu. Bilgisayarların veri analizi için kullanılmaya başlamasıyla istatistiksel çalışmalar hız kazandı. Hatta bilgisayarın varlığı daha önce yapılması mümkün olmayan istatistiksel araştırmaları mümkün kıldı. 1990lardan sonra istatistik, veri madenciliği ile ortak bir platforma taşındı. Verinin, yığınlar içerisinden çekip çıkarılması ve analizinin yapılarak kullanıma hazırlanması sürecinde veri madenciliği ve istatistik sıkı bir çalışma birlikteliği içine girmiş bulundular.

Bunun yanısıra veri madenciliği, veritabanları ve makine öğrenimi disipliniyle birlikte yol aldı. Günümüzdeki Yapay Zeka çalışmalarının temelini oluşturan makine öğrenimi kavramı, bilgisayarların bazı işlemlerden çıkarsamalar yaparak yeni işlemler üretmesidir. Önceleri makineler, insan öğrenimine benzer bir yapıda inşa edilmeye çalışıldı. Ancak 1980lerden sonra bu konuda yaklaşım değişti ve makineler daha spesifik konularda kestirim algoritmaları üretmeye yönelik inşa edildi. Bu durum ister istemez uygulamalı istatistik ile makine öğrenim kavramlarını, veri madenciliği altında bir araya getirdi.

Bilgi Keşfi

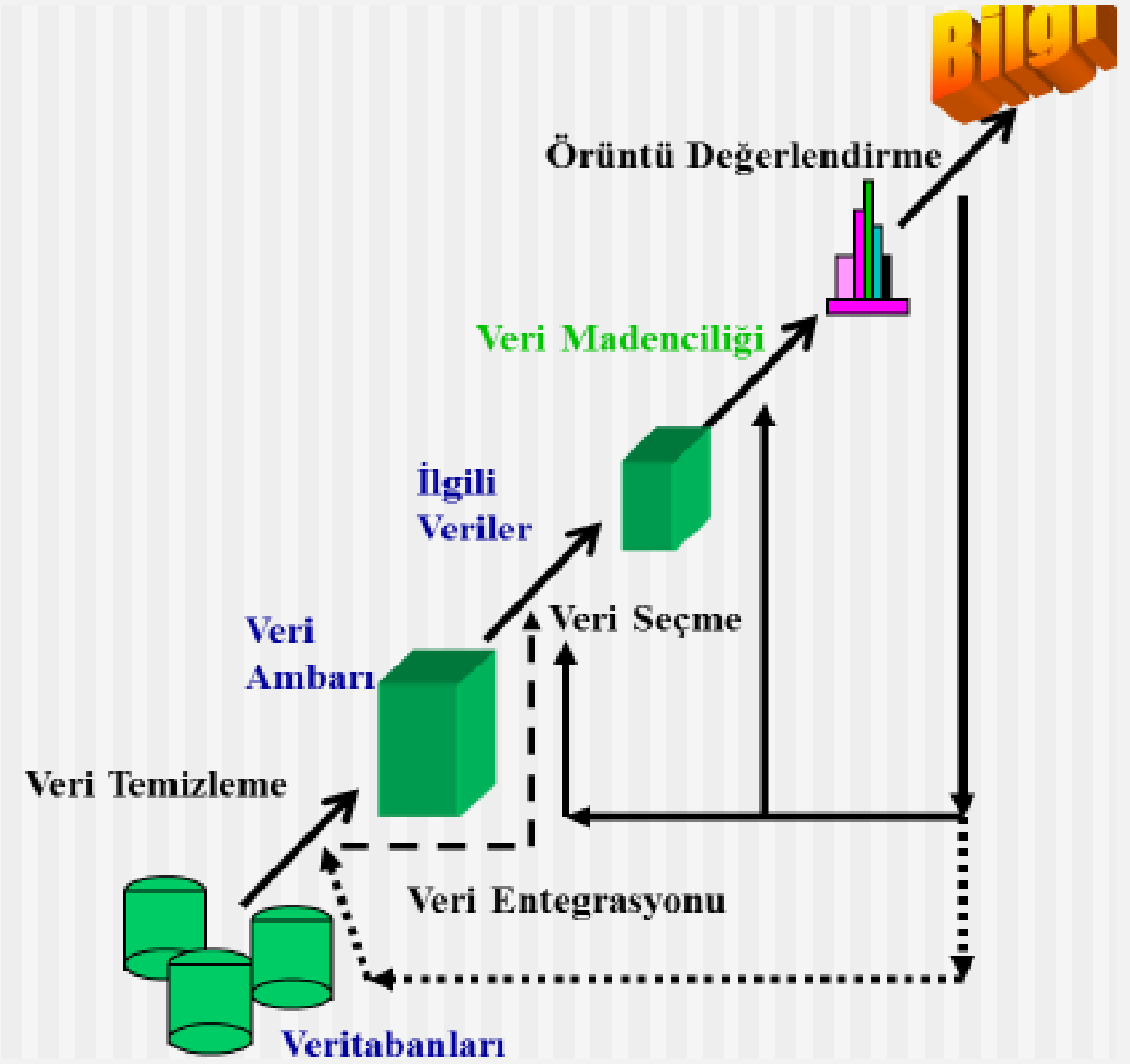
Teoride veri madenciliği bilgi keşfi işleminin aşamalarından biridir.

Pratikte veri madenciliği ve bilgi keşfi eş anlamlı olarak kullanılır.

Veri madenciliği teknikleri veriyi belli bir modele uydurur.

veri içindeki örüntüleri bulur

örüntü: veri içindeki herhangi bir yapı



Bilgi Keşfinin Aşamaları

- Veri Temizleme : Gürültülü ve tutarsız verileri çıkarmak
- Veri Bütünleştirme : Birçok data kaynağını birleştirebilmek
- Veri Seçme : Yapılacak olan analiz ile ilgili olan verileri belirlemek
- Veri Dönüşümü : Verinin veri madenciliği yöntemine göre işlenebilir hale dönüşümünü gerçekleştirmek
- Veri Madenciliği : Verilerdeki örüntülerin belirlenmesi için veri madenciliği yöntemlerinin uygulanması
- Örüntü Değerlendirme: Bazı ölçütlere göre elde edilmiş ilginç örüntüleri bulmak ve değerlendirmek
- Bilgi Sunumu : Elde edilen bilgilerin kullanıcılara sunumunu

Veri Madenciliği için Veri Kaynakları

Veritabanı Sistemleri

Veritabanı sistemleri yazılımları bilgisayar sistemlerinin önemli bir bileşeni olarak değerlendirilir. Veritabanı yönetim sistemleri, birbirleriyle ilişkili veri ve programlar topluluğundan oluşmaktadır. Veri topluluğu bir «Veritabanı» olarak değerlendirilir. Veritabanı bir kuruluşa ilişkin bilgilerin yer aldığı ortamdır.

Veri Ambarı

Veritabanı sistemlerinin avantajlarına rağmen, özellikle karar destek uygulamalarında gereksinimleri karşılamakta zorlandığı görülmüştür. Bu gereksinime paralel olarak verinin farklı biçimlerde saklanması, veriye hızlı erişimin sağlanması, ayrıca kurumun ürettiği tüm bilgilere erişimini sağlayacak kadar büyük boyutlu veriyi işleme ihtiyaçları ile veri ambarı adı verilen ortam oluşturulmuştur. Veri ambarı karar destek sistemlerinin teknik altyapısını oluşturmaktadır. Veri ambarı, kuruluşun sahip olduğu verinin (eskiler de dahil olmak üzere), karar destek amacıyla kullanılmasına olanak sağlamaktadır.

Veri Madenciliği için Veri Kaynakları

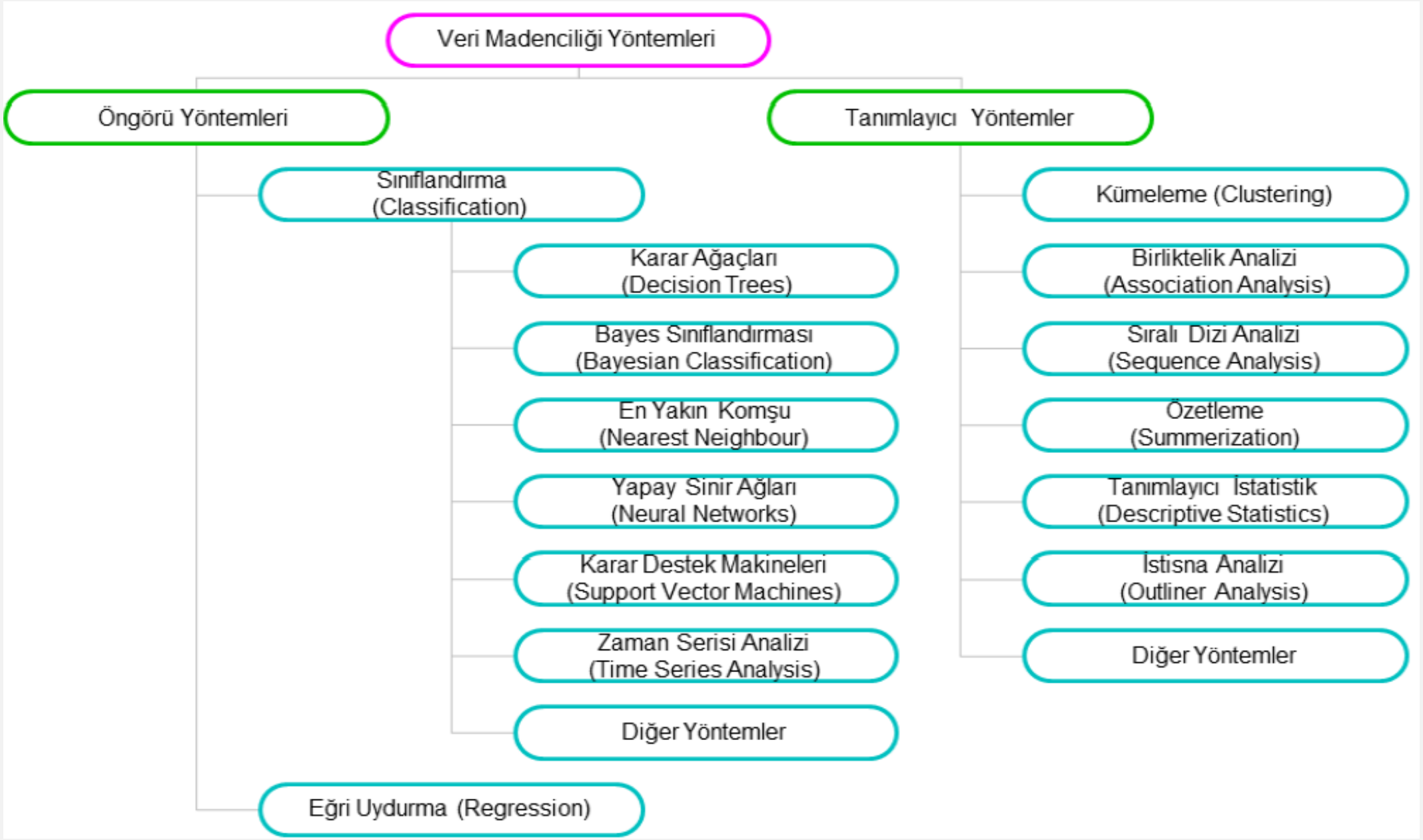
Veri Kümeleri

Veritabanları ve veri ambarı bir yönetim sistemi yardımıyla veriyi düzenler ve yönetilmesini sağlar. Bu yapılar gerektiğinde veri madenciliği sistemi tarafından kullanılır. ancak özellikle bireysel araştırmacılar veri madenciliği alanında veri kaynağı olarak çoğunlukla veri kümesi adını verdiğimiz yapıları kullanırlar.Veri kümesi bir veritabanı tablosu olarak değil, bir Excel, metin dosyası ve benzeri biçimindedir.

Veri Madenciliği Algoritmaları

- amaç : veriyi belli bir modele uydurmak
  - tanımlayıcı
    - En iyi müşterilerim kimler?
    - Hangi ürünler birlikte satılıyor?
    - Hangi müşteri gruplarının alışveriş alışkanlıkları benzer?
  - kestirime dayalı
    - Kredi başvurularını risk gruplarına ayırma
    - Şirketle çalışmayı bırakacak müşterileri öngörme
    - Borsa tahmini
- seçim: veriye uyan en iyi modeli seçmek için kullanılan kriter
- arama: veri üzerinde arama yapmak için kullanılan teknik

Veri Madenciliği Yöntemleri



Veri Madenciliği Yöntemleri

- Sınıflandırma (Classification): Veriyi önceden belirlenmiş sınıflardan birine dahil eder.
  - Danışmanlı (Gözetimli) öğrenme
  - Örüntü tanıma
  - Kestirim
- Eğri uydurma (Regression): Veriyi gerçel değerli bir fonksiyona dönüştürür.
- Zaman serileri inceleme (Time Series Analysis): Zaman içinde değişen verinin değerini öngörür.
- İstisna Analizi (Outlier Analysis): Verinin geneline uymayan nesneleri belirleme
- Kümeleme (Clustering): Benzer verileri aynı grupta toplama
  - Danışmansız (Gözetimsiz) öğrenme
- Özetleme (Summarization): Veriyi alt gruplara ayırır. Her alt grubu temsil edecek özellikler bulur.
  - Genelleştirme (Generalization)
  - Nitelendirme (Characterization)
- İlişkilendirme kuralları (Association Rules)
  - Veriler arasındaki ilişkiyi belirler
- Sıralı dizileri bulma (Sequence Discovery): Veri içinde sıralı örüntüler bulmak için kullanılır.

KAYNAKÇA

- Dr. Öğr. Üyesi Kadriye ERGÜN Ders Notları
- Veri Madenciliği Yöntemleri, Dr. Yalçın ÖZKAN

VERİ MADENCİLİĞİ

Bölüm 2

Öğr. Gör. Merve KESİM ÖNAL

Veri Madenciliğinde Temel Kavramlar

- Veri (Data)
- Enformasyon(Information)
- Bilgi (Knowledge)
- Bilgelik (Wisdom)

Veri

Veri kelimesi Latince'de "gerçek, reel" anlamına gelen "datum" kelimesine denk gelmektedir. "Data" olarak kullanılan kelime ise çoğul "datum" manasına gelmektedir. Her ne kadar kelime anlamı olarak gerçeklik temel alınsa da her veri her daim somut gerçeklik göstermez. Kavramsal anlamda veri, kayıt altına alınmış her türlü olay, durum, fikirdir. Bu anlamıyla değerlendirildiğinde çevremizdeki her nesne bir veri olarak algılanabilir.

Veri, oldukça esnek bir yapıdadır. Temel olarak varlığı bilinen, işlenmemiş, ham haldeki kayıtlar olarak adlandırılırlar. Bu kayıtlar ilişkilendirilmemiş, düzenlenmemiş yani anlamlandırılmamışlardır. Ancak bu durum her zaman geçerli değildir. İşlenerek farklı bir boyut kazanan bir veri, daha sonra bu haliyle kullanılmak üzere kayıt altına alındığında, farklı bir amaç için veri halini koruyacaktır.

Enformasyon (Information)

Yaygın anlamda enformasyon terimi, «haber» veya «mesaj» terimiyle eşanlamlıdır.

Enformasyon, veri kavramının tanımından yola çıkıldığında, verilerin ilişkilendirilmiş, düzenlenmiş, anlamlandırılmış, işlenmiş halidir. Bu haliyle enformasyon, potansiyel olarak içinde bilgi barındıran bir veri halindedir.

Veriler enformasyona dönüştürülerek kullanışlı hale getirilirler. Bu yönüyle enformasyon anlam katılmış verilerdir.

Bilgi (Knowledge)

Bilgi, bu süreçteki üçüncü aşamadır. Enformasyonun, bilgiye dönüşmesi, bireyin onu algılaması, özümsemesi ve sonuç çıkarmasıyla gerçekleşir. Dolayısıyla bireyin algılama yeteneği, yaratıcılık, deneyim gibi kişisel nitelikleri de bu süreci doğrudan etkilemektedir.

«İnsan aklının erebileceği olgu, gerçek ve ilkeler bütünü, malumat» olarak tanımlanan bilgi, bilişim dilinde kurallardan yararlanarak kişinin veriye yönelttiği anlam demektir. Farklı enformasyon parçacıkları arasındaki ilişkilerdir.

Bilgelik (Wisdom)

Bilgelik ulaşılmaya çalışılan noktadır ve bu kavramların zirvesinde yer alır. Bilgilerin kişi tarafından toplanıp bir sentez haline getirilmesiyle ortaya çıkan bir olgudur. Yetenek, tecrübe gibi kişisel nitelikler birer bilgelik elemanıdır.

Neyin bilindiğinin (bilgi) ve en iyinin ne olduğunun (sosyal ve etnik faktörler) dikkate alınarak en uygun davranışın sergilenmesi demektir. Belirli bir alanı veya alanları anlamak için daha geniş ve genelleştirilmiş kuralları ve şemaları temsil etmesiyle bilgiden ayrılır.

Veri Ambarı

Veritabanı: birbirleriyle ilişkili bilgilerin depolandığı alanlardır.

Veri Ambarı: ilişkili verilerin sorgulandığı ve analizlerinin yapılabildiği bir depodur. Veri ambarı veritabanını yormamak için oluşturulmuştur. Bir veri ambarı ilgili veriyi kolay, hızlı, ve doğru biçimde analiz etmek için gerekli işlemleri yerine getirir. Veri ambarı, işlemsel sistemlerdeki veriyi kopyalayıp, karar verme işlemi için uygun formda saklar.

Veri ambarı, bir zaman boyutu içinde analitik işlemlerin yapılması için ihtiyaç duyulan bilgi temelini sağlar.

Veri ambarı, karar verme sürecinde yöneticilere destek vermek üzere hazırlanmış;

- Konuya(Amaca) Yönelik
- Bütünleşik (Birleştirilmiş)
- Zaman Değişkenli
- Sadece okunabilen

veri topluluğudur.

Veri Ambarı - Konuya(Amaca) Yönelik

Bir ambarının ilk özelliği işletmenin belli başlı amaçlarına ya da konularına yönelik olmasıdır. Bunun anlamı veri ambarının işletmedeki yüksek seviyeli varlıklar üzerine odaklanmış olmasıdır.

Klasik işlemsel sistemler, işletmelerdeki işlemler, süreçler ya da fonksiyonlar üzerine yoğunlaşmışlardır. Buna karşılık, veri ambarı işletmedeki konular üzerine yoğunlaşmıştır. Örneğin işlemsel uygulamalar, muhasebe, personel, stok vb. gibi sistemlere ya da fonksiyonlara yönelik olabilir. Veri ambarı ise müşteri, satıcı ürün, eylem gibi konulara yönelik olarak tasarlanabilir. Veri ambarında yer alan verinin tasarımı ve uygulaması söz konusu konulara göre düşünülür.

Kısaca veri ambarı, verinin incelenmesi ve modellenmesi için oluşturulur. Konuyla ilgili karar vermek için gerekli olmayan veriyi kullanmayarak konuya basit, özet bakış sağlar.

Veri Ambarı - Bütünleşik (Birleştirilmiş)

Veri ambarı, veri kaynaklarının veri temizleme ve birleştirme teknikleri uygulanarak birleştirilmesiyle oluşturulurlar. Veri ambarında değişik veri kaynakları arasındaki tutarlılık sağlanmalıdır.

Veri ambarı içindeki veri mutlaka bütünleşik olmalıdır. Bütünleşme farklı biçimlerde olabilir. verinin kodlamasında görüş birliğine varılması, ölçü birimlerinin seçiminde tutarlılık, sayısal değerlerin fiziksel gösterimindeki tutarlılık vb. gibi bütünleştirme kavramlarından söz edilebilir.
Örneğin, cinsiyet ile ilgili alan bulunan bir uygulamada cinsiyet "E" ve "K" olarak ifade edilirken diğer bir uygulamada "1" ve "0" değerleri ile ifade edilmiş olabilir. Eğer böyle bir durum söz konusu ise veri tabanından veri ambarına veri taşınırken bu alanın düzenlenerek tek tip biçime dönüştürmek gerekmektedir.

Veri Ambarı - Zaman Değişkenli

Veri ambarındaki tüm veri zamanın belirli bir anına aittir. Veri ambarındaki verinin bu temel karakteristiği, işlemsel sistemlerdeki veriden oldukça farklıdır. İşlemsel sistemlerde veri o anda var olan değerdir. İşlemsel sistemler de zamana dağılmış olabilir. Ancak bu süre çok geniş bir zaman aralığını kapsamaz. (en çok geçmiş 1 yıl)

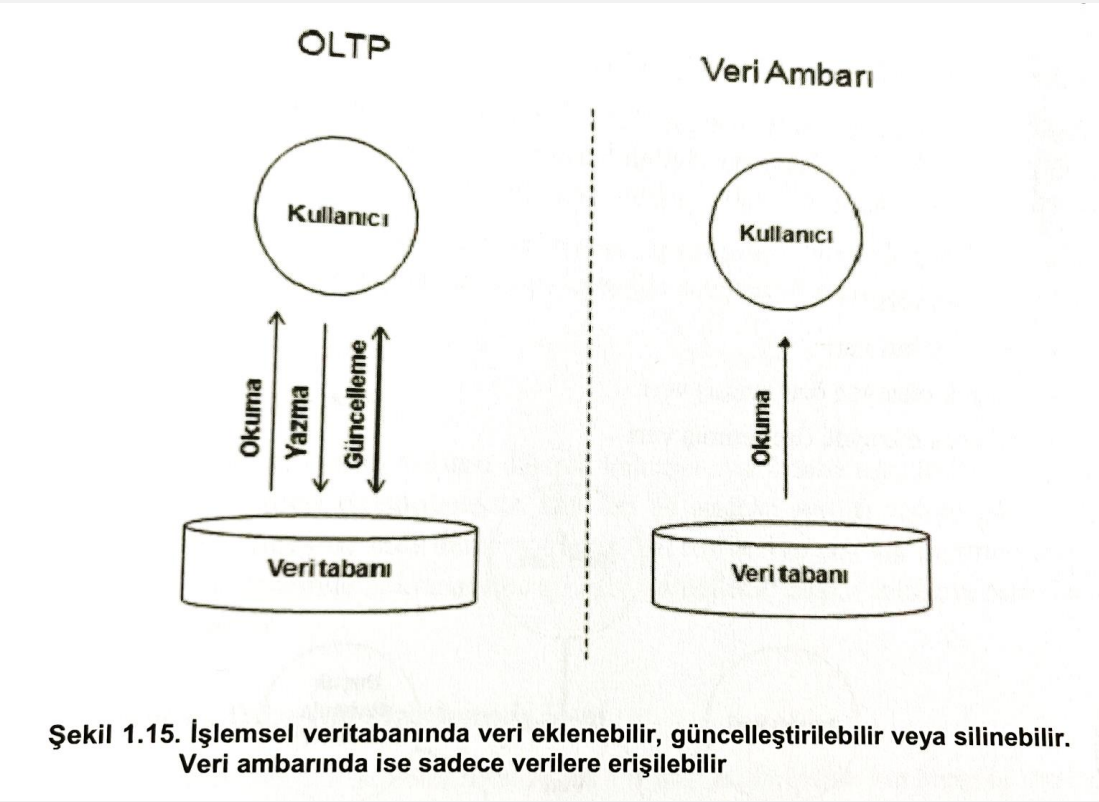
Veri ambarındaki veri ise, sadece o andaki değerlerle değil; geçmişteki değerlerle de ilgilidir. Verinin zaman içinde aldığı değerleri de çözümlemeye katar. (geçmiş 5-10 yıl)

Veri Ambarı - Sadece okunabilen

Veri ambarında veri sadece okunabilir bir yapıdadır. Veri ambarları, yönetimin gereksinimlerine yanıt vermek üzere tasarlandığı için günlük işlemlere tabi tutulamaz; yani silinemez veya güncelleştirilemez. İşlemsel sistemlerde(OLTP) yer alan veri değiştirilebilir, silinebilir ve gerektiğinde yeni veri eklenebilir. Veri ambarında iki tür işlemden söz edilir:

- Verinin yüklenmesi
- Veriye erişilmesi

verinin bildiğimiz anlamda güncelleştirmesi söz konusu değildir.



Şekil 1.15. İşlemsel veritabanında veri eklenebilir, güncelleştirilebilir veya silinebilir. Veri ambarında ise sadece verilere erişilebilir

Veri Ambarının Temel Özellikleri

Önceki söylediklerimizi ve işlemsel sistemlerin özelliklerini de göz önüne alarak veri ambarları ile ilgili aşağıda belirtilen hususları sıralayabiliriz:

a) İşlemsel çevrede yer alan veri bir süzme işlemi sonucunda veri ambarı çevresine aktarılır. İşlemsel ortamdaki verinin tümüyle veri ambarına aktarılması beklenmez. Sadece Karar Destek Sistemlerinin gereksinim duyacağı veri aktarılır.

b) Zaman yelpazesi her iki sistemde farklılık gösterir. İşlemsel ortamdaki veri çok tazedir. Veri ambarındaki veri ise daha eskidir. Ancak zaman açısından bakıldığında, işlemsel ve veri ambarı ortamındaki veri arasında kısa bir örtüşme vardır.

c) Veri ambarı özet bilgileri içerebilir. İşlemsel sistemlerde ise özet veriye yer verilmez.

d) Veri ambarının en önemli özelliklerinden biri olan bütünleştirmeyi sağlamak üzere, verinin önemli bir kısmı belirli bir dönüşümden sonra veri ambarına aktarılır. Böylece işlemsel veri ile veri ambarındaki verinin içeriği açısından farklılıklar oluşabilecektir.

Veri Ambarının İçerdiği Veri

Veri ambarı, içerdiği veri açısından da göz önüne alındığında farklı bir yapıya sahip olduğu anlaşılacaktır. Aşağıda veri ambarının içerdiği veriyi sınıflandırıyoruz [Inmon, 2005; Kimball, 2002; Han, 2006]:

- Metadata
- Ayrıntı veri
- Eski ayrıntı veri
- Düşük düzeyde özetlenmiş veri
- Yüksek düzeyde özetlenmiş veri

Şekil 1.6. Veritabanı çeşitli seviyede veriler içerir

Veri Ambarının İçerdiği Veri

Metadata

Veri ambarlarının en önemli bileşenlerinden biri metadata'dır. Metadata doğrudan işlemsel çevreden gelen veriyi içermez. Metadata veri ambarı içindeki veriden ayrı bir yerdedir. Metadata aşağıda belirtilen özelliklere sahiptir:

a) Karar Destek Sistemleri analistlerine yardım etmek üzere yaratılan bir dizindir ve veri ambarı içeriğinin neler olduğunu belirtir. Kullanılan verinin yapısını ortaya koyan bilgileri içerir.
b) İşlemsel çevreden veri ambarına dönüştürülen verinin konumları hakkında bilgileri içeren bir kılavuzdur.
c) İşlemsel çevreden alınan verinin hangi algoritmaya göre düşük ya da yüksek seviyede özetlendiği hakkındaki bilgileri içeren bir kılavuzdur.

Bu özelliklerinden dolayı metadata aslında "veriye ilişkin veri" olarak da tanımlanabilir.

Ayrıntı Veri

Veri ambarı şu andaki ayrıntı veriyle ilgilendiğini varsayalım. Bu veri en son olayları içermektedir ve henüz işlenmediği için diğerlerine oranla daha büyük hacimlidir. Bu tür veri disk üzerinde saklandığından bunlara erişim ve yönetimi pahalıdır. Ayrıntı veri denilince, sadece şu andaki en son ayrıntı veri kastedilmemektedir. Bu ayrıntılar belirli bir dönemi kapsayabilir. Örneğin satışlara ilişkin veri son beş yılın ayrıntılarını içerebilir.

Veri Ambarının İçerdiği Veri

Eski Ayrıntı Veri

Yukarıda sözü edilen ayrıntı verinin dışında kalanlar, yani daha eski tarihe ait olanlar eski ayrıntı bilgileri oluşturacaktır. Bu veri şu andaki ayrıntı veriye göre daha düşük bir ayrıntı düzeyine indirilerek saklanır. Bu tür veriye çok sık biçimde erişilmediği için, disk üzerinde saklanabileceği gibi, genellikle başka saklama birimlerine yerleştirilebilirler.

Düşük Düzeyde Özetlenmiş Veri

Şu andaki ayrıntı veriden süzülerek elde edilen düşük seviyede özetlenmiş veri de veri ambarının bir parçası olabilir. Bu tür veri disk üzerinde saklanır. Veri ambarının tasarımı esnasında, hangi verinin özetleneceği ve özetleme işleminin ne düzeyde olacağı belirlenir.

Yüksek Düzeyde Özetlenmiş Veri

Şu andaki ayrıntı veri daha yüksek düzeyde özetlenerek, kolayca erişilebilir hale getirilebilir. Bu tür veri de veri ambarının bir bileşeni olarak yer alabilir.

KAYNAKÇA

- Dr. Öğr. Üyesi Kadriye ERGÜN Ders Notları
- Veri Madenciliği Yöntemleri, Dr. Yalçın ÖZKAN
- Prof. Dr. Bülent TUGRUL Ders Notları

VERİ MADENCİLİĞİ

Bölüm 3

Öğr. Gör. Merve KESİM ÖNAL

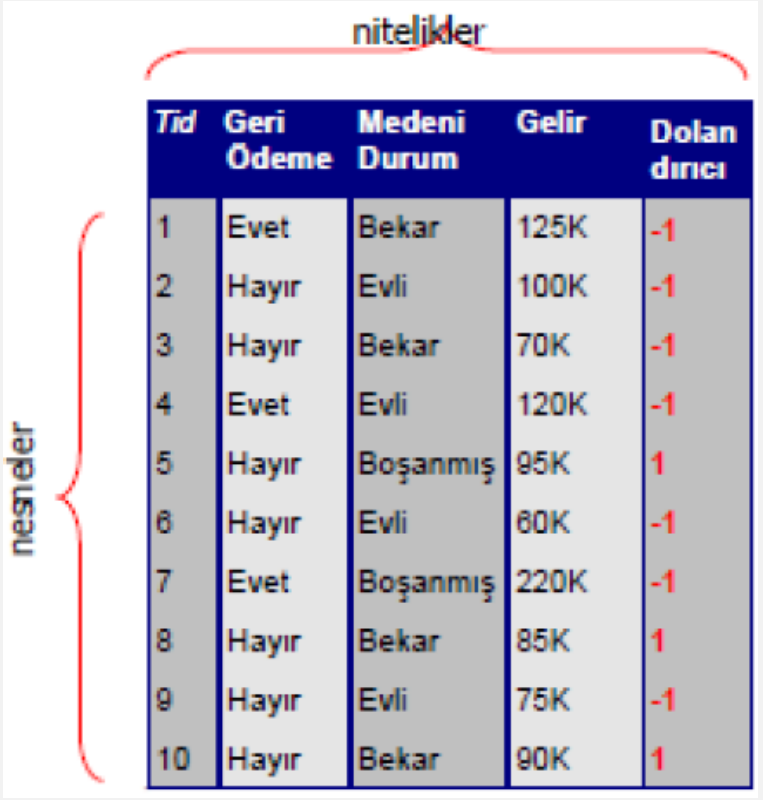
Veri

Veri, nesneler ve nesnelerin niteliklerinden oluşan kümedir.

Nitelik (attribute), nesnenin (object) bir özelliğidir. Nitelik aynı zamanda değişken, alan, karakteristik veya özellik olarak da bilinir

Örnek: bir insanın yaşı, ortamın sıcaklığı…

Nitelikler ve bu niteliklere ait değerler bir nesneyi oluşturur. Nesne, kayıt (record), varlık (entity), örnek (sample, instance) olarak da adlandırılır.



Tid	Geri Ödeme	Medeni Durum	Gelir	Dolandırıcı
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

Veri Nitelik Türleri

Genel olarak nicel ve nitel olarak iki başlık altında incelenirler.

Nicel Veriler : Sayısal, Nümerik, Nicel Veriler de denmektedir. Sayısal ölçekte kaydedilen ölçümlerdir. Boy, yaş, kilo verileri örnek gösterilebilir. Sürekli ve süreksiz olarak iki başlıkta ele alınabilir:

a) Sürekli Nümerik Veriler: Yaş, Sıcaklık
b) Aralıklı Nümerik Veriler (Interval): Çocuk Sayısı, Kaza Sayısı

Nitel Veriler : Sayısal ölçekte kaydedilemeyen ölçümlerdir. Gözlem sonucu elde edilir. Örnek: Renkler, Medeni Durum, Tatlar

Veri Nitelik Türleri

Veriler ölçüm düzeylerine göre de sınıflandırılabilirler.
Nominal Veriler : Kategorik bir veri çeşididir. İkiye ayrılır:

a) Binary Veriler: Var-Yok, Kadın-Erkek, Hasta-Sağlıklı
b) İkiden Çok Kategorili: Medeni Durum, Renk, Kimlik no, Şehir, İsim, Forma Numarası

Ordinal Veriler : Ordinal veriler de kategorik veri türündendir. Fakat değerleri arasında sıralı bir ilişki bulunmaktadır. Nominal veriler, ordinal verilere göre daha az bilgi taşırlar.
Örneğin: Eğitim Düzeyi, Sosyoekonomik ölçek skorları, sıralamalar (örneğin, 1'den 10'a kadar bir ölçekte patates cipsinin tadı), ders harf notları gibi.

Aralıklı (Interval) Veriler: Sıralayıcı ölçeğe benzer fakat farklar arası anlamlıdır. Toplama, çıkarma işlemleri yapılabilir.
Örneğin: Sıcaklık, IQ değeri gibi

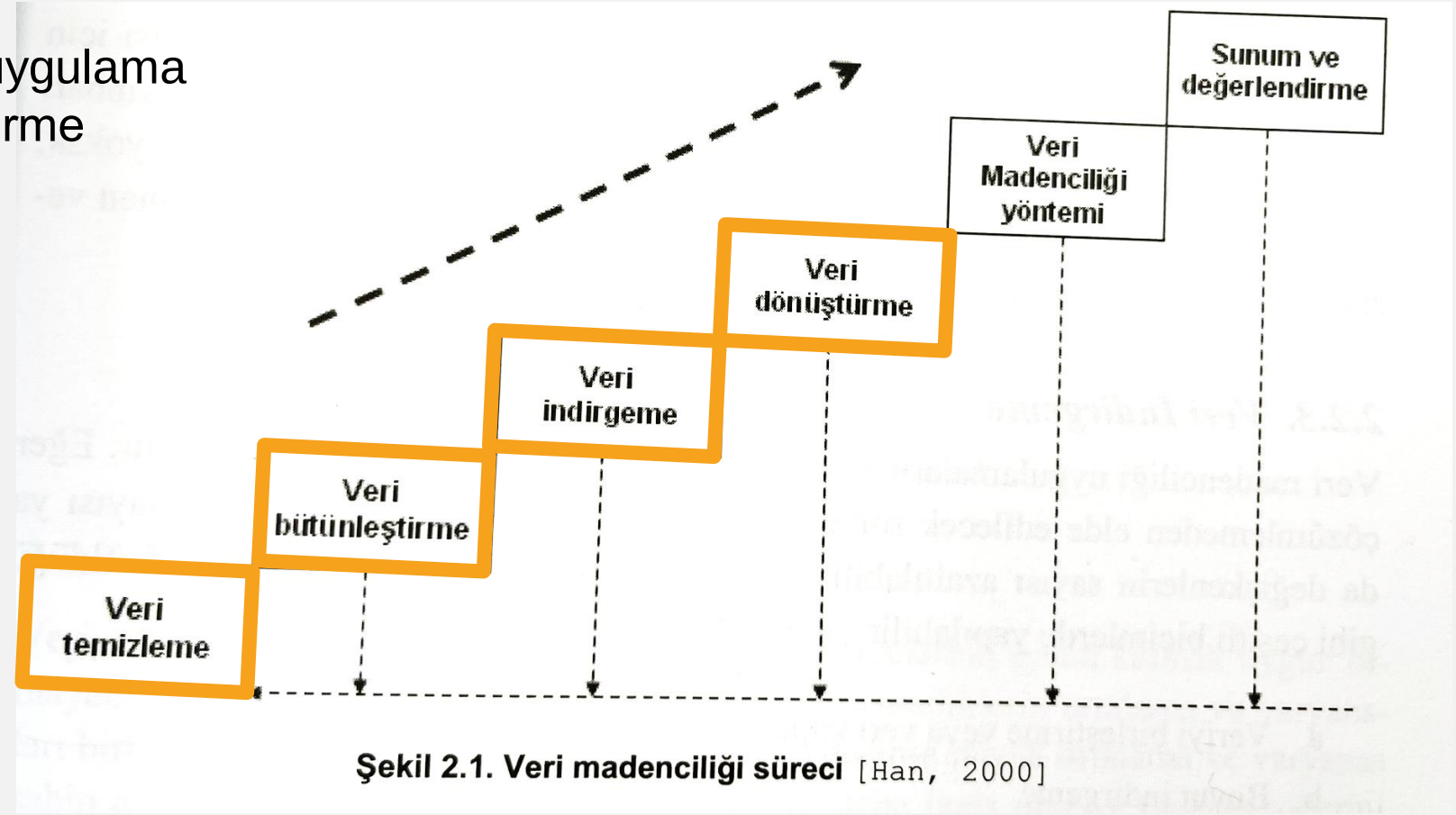
Oransal (Ratio) Veriler : Oran verilebilir veri türlerine Ratio veriler denir. Nümerik verilere benzerler.
Örnek: 100 santigrat derece, 50 santrigat derecenin iki katı denilemez ama derece kelvine çevrilirse 60 kelvin, 30 kelvinin 2 misli sıcak denilebilir. Burada kelvin ratio türünden bir değişken iken, santigrat ise nümerik veri türüne örnek olarak verilebilir. Buna benzer şekilde uzunluk ya da nesne sayımları da ratio türüne örnek verilebilir.

Veri Madenciliği Süreci

Veri madenciliğini belirli bir adım değil, bir süreç olarak değerlendirmek gerekiyor. Söz konusu süreç şu adımları içerir:

- Veri temizleme
- Veri bütünleştirme
- Veri indirgeme
- Veri dönüştürme
- Veri madenciliği algoritmasını uygulama
- Sonuçları sunum ve değerlendirme

Veri Önişleme



Şekil 2.1. Veri madenciliği süreci [Han, 2000]

Veri Önişleme

Veri önişleme, veri setinin analiz ve modelleme için hazır hale getirilmesi işlemidir. Bu işlem, verinin kalitesini ve doğruluğunu artırmayı ve modelleme aşamasında daha iyi sonuçlar elde etmeyi amaçlar.

Veri Önişleme neden gerekli?

Veri önişleme, her veri madenciliği projesinin önemli bir bileşenidir.
Veriler toplanırken veri kalitesini etkileyebilecek bazı problemler;

- Eksik, tutarsız ve hatalı veriler de toplanmış olabilir.
- Veri içerisinde gürültülü ve aykırı değerler bulunabilir.
- Veriler uygun formatta toplanmamış olabilir veya farklı ölçeklerde toplanmış olabilir.
- Yinelenmiş veriler bulunabilir.

Gürültü: Orijinal değerlerin değiştirilmesini ifade eder. Örneğin: telefonla konuşurken kişinin sesinin bozulması, ekranda "kar" görülmesi
Aykırı değer: veri kümesindeki diğer veri nesnelerinin çoğundan oldukça farklı özelliklere sahip veri nesneleridir.

Veriyi Tanıma

Veri tanıma, veri önişleme aşamasının ilk adımıdır ve veri setinin genel bir resmini elde etmeyi amaçlar. Bu aşamada, veri setindeki değişkenlerin türleri, anlamları, dağılımları ve istatistiksel özellikleri hakkında bilgi edinilir. Veri tanıma, veri setinin kalitesini değerlendirmek ve veri önişleme için hangi tekniklerin kullanılacağını belirlemek için önemlidir.

Veriyi tanımlayan özelliklerden bazıları:

- Veri setindeki değişkenlerin sayısı, türleri ve değer aralıkları belirlenir.
- Temel istatistiksel bilgiler (ortalama, medyan, mod, standart sapma, minimum ve maksimum değerler) hesaplanır.
- Histogramlar, dağılım grafikleri, kutu grafikleri ve diğer görselleştirmeler kullanılarak değişkenlerin dağılımları ve eğilimleri incelenir.
- Verilerin normal dağılım gösterip göstermediğini test edilir.
- Farklı değişkenler arasındaki ilişkileri test etmek için korelasyon testleri ve regresyon analizleri yapılır.

Veriyi Tanıma

Merkezi Eğilim;

Ortalama: $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ $\mu = \frac{\sum x}{N}$

- ağırlıklı ortalama $\bar{x} = \frac{\sum_{i=1}^{n} w_i x_i}{\sum_{i=1}^{n} w_i}$
- kırpılmış ortalama: Uç değerleri kullanmadan hesaplama

Ortanca (median): Verinin tümü kullanılarak hesaplanır

- veri sayısı tek ise ortadaki değer, çift sayı ise ortadaki iki değerin ortalaması

$$median = L_1 + (\frac{n/2 - (\sum f)_l}{f_{median}})c$$

Mod

- Veri içinde en sıklıkla görülen değer
- Unimodal, bimodal, trimodal
- deneysel formül: $mean - mode = 3 \times (mean - median)$

Veriyi Tanıma

Merkezi Eğilim;

Ortalama: $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ $\mu = \frac{\sum x}{N}$

- ağırlıklı ortalama $\bar{x} = \frac{\sum_{i=1}^{n} w_i x_i}{\sum_{i=1}^{n} w_i}$
- kırpılmış ortalama: Uç değerleri kullanmadan hesaplama

Ortanca (median): Verinin tümü kullanılarak hesaplanır

- veri sayısı tek ise ortadaki değer, çift sayı ise ortadaki iki değerin ortalaması

$$median = L_1 + (\frac{n/2 - (\sum f)_l}{f_{median}})c$$

Mod

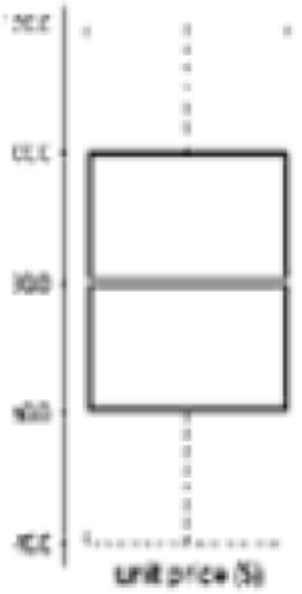
- Veri içinde en sıklıkla görülen değer
- Unimodal, bimodal, trimodal
- deneysel formül: $mean - mode = 3 \times (mean - median)$

Veriyi Tanıma

Veri dağılımı;

Çeyrek, aykırılıklar, kutu grafiği çizimi

- Çeyrek (quartile) : nitelik değerleri küçükten büyüğe doğru sıralanır.
  - Q1: ilk %25, Q3: ilk %75
- Dörtlü aralık (Inter-quartile Range): IQR= Q3-Q1
- Five Number Summary: min, Q1, median, Q3, max
- Kutu Grafiği Çizimi:
  - Q1 ve Q3 aralığında bir kutu
  - kutu içinde ortanca noktayı gösteren bir çizgi
  - kutudan min ve max değerlere birer uzantı
- Aykırılıklar: 1,5xIQR değerinden küçük/büyük olan değerler



Varyans ve standart sapma

$$=\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^2=\frac{1}{n-1}[\sum_{i=1}^{n}x_i^2-\frac{1}{n}(\sum_{i=1}^{n}x_i)^2] \qquad \sigma^2=\frac{1}{N}\sum_{i=1}^{n}(x_i-\mu)^2=\frac{1}{N}\sum_{i=1}^{n}x_i^2-\mu^2$$

Veri Temizleme

Bazı uygulamalarda, üzerinde çözümleme yapılacak verilerin istenen özelliklere sahip olmadığı görülebilir. Örneğin eksik verilerle ve uygun olmayan verilerin oluşturduğu tutarsız verilerle karşılaşılabilir. Veri tabanında yer alan tutarsız ve hatalı veriler gürültü olarak değerlendirilmektedir. Bu gibi durumlarda verinin söz konusu sorunlardan temizlenmesi gerekecektir.

Veri temizleme işlemleri

- Eksik nitelik değerlerini tamamlama
- Aykırılıkların bulunması ve gürültülü verinin düzeltilmesi
- Tutarsızlıkların giderilmesi

Veri Temizleme

Eksik Verinin Temizlenmesi

Eksik verilerin yerine yenileri belirlenerek konulmalıdır. Bunun için aşağıdaki yöntemlerden biri kullanılabilir,

- Eksik değer içeren kayıtlar veri kümesinden atılabilir.
- Kayıp değerlerin yerine bir genel sabit kullanılabilir. Bütün kayıp değerler için aynı sabit kullanılabilir. Örneğin «bilinmiyor» değeri bu eksik veri yerine kullanılabilir. Ancak bütün değişkenlere kayıp değerler yerine aynı sabit değerin kullanımı sorun yaratacaktır.
- Değişkenlerin tüm verileri kullanılarak ortalaması hesaplanır ve eksik değer yerine bu değer konulabilir.
- Değişkenlerin tüm verileri yerine, sadece bir sınıfa ait örneklerin değişken ortalaması hesaplanarak eksik değer yerine kullanılabilir.
- Verilere uygun bir tahmin yapılarak, örneğin regresyon ya da karar ağacı modeli kurularak eksik değer tahmin edilebilir ve eksik değer yerine kullanılabilir.

Örnekler:

- Bir veri setinde, bazı öğrencilerin sınav puanları eksik olabilir. Bu durumda, eksik puan değerlerini, veri setindeki tüm öğrencilerin sınav puanlarının ortalaması ile tamamlayabiliriz.
- Bir veri setinde, hastaların cinsiyet bilgileri eksik olabilir. Bu durumda, eksik cinsiyet değerlerini, benzer özelliklere sahip hastaların cinsiyet bilgileri ile tamamlayabiliriz.

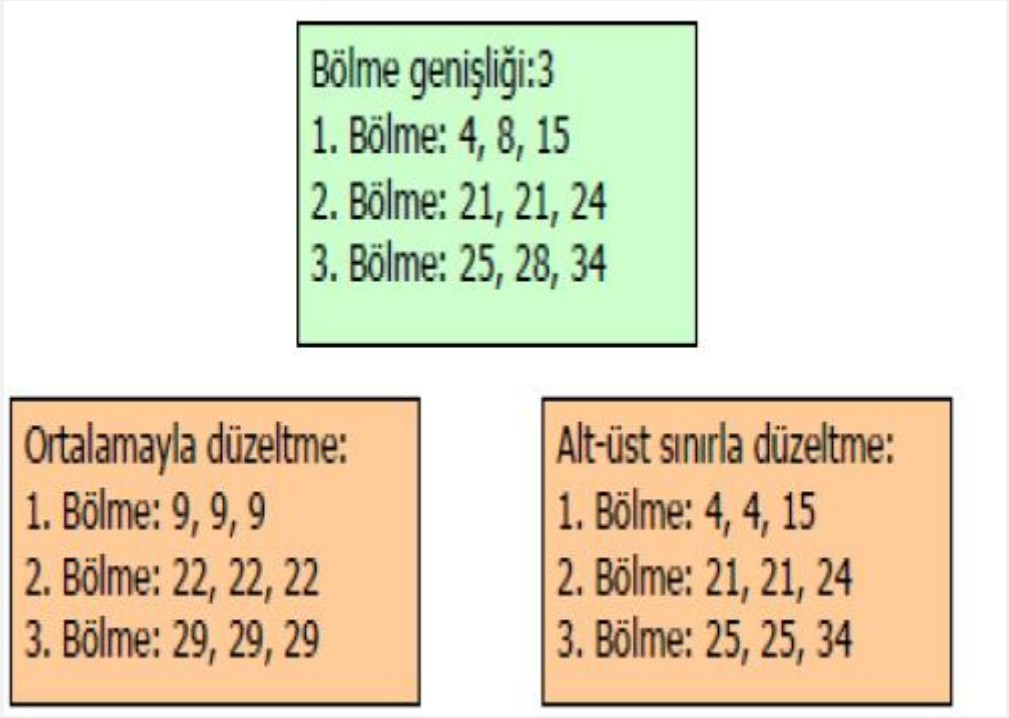
Veri Temizleme

Gürültülü Verinin Temizlenmesi

Bölmeleme

Veri sıralanır, eşit genişlik veya eşit derinlik ile bölünür. Her bölme ortalamayla ya da bölmenin en alt ve üst sınırlarıyla temsil edilir.

Veri sıralanır: 4, 8, 15, 21, 21, 24, 25, 28, 34
Eşit genişlik: Bölme sayısı belirlenir. Eşit aralıklarla bölünür.
Eşit derinlik: Her bölmede eşit sayıda örnek kalacak şekilde bölünür.

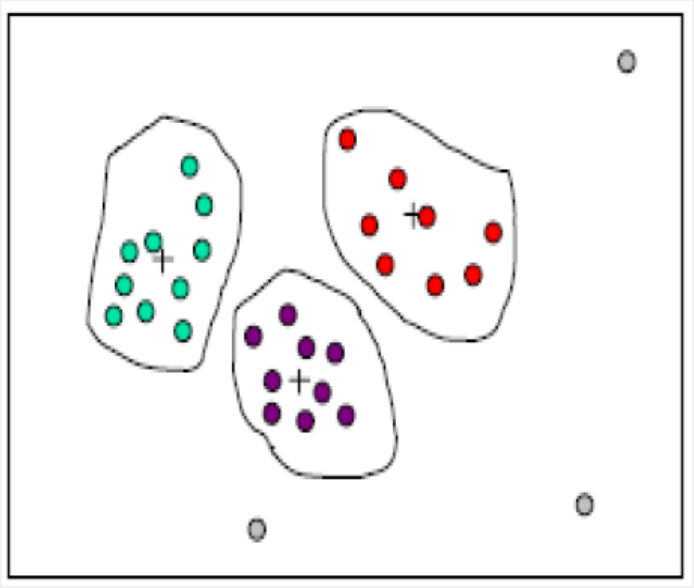


Veri Temizleme

Gürültülü Verinin Temizlenmesi

Kümeleme

Benzer veriler aynı kümede olacak şekilde gruplanır. Bu kümelerin dışında kalan veriler aykırılık olarak belirlenir ve silinir.



Eğri uydurma

Veriyi bir fonksiyona uydurarak gürültüyü düzeltir.

Veri Bütünleştirme

Farklı veri tabanlarından ya da veri kaynaklarından elde edilen verilerin birlikte değerlendirmeye alınabilmesi için farklı türdeki verilerin tek türe dönüştürülmesi yani bütünleştirilmesi söz konusu olacaktır. Veri madenciliği eğer veri ambarı altyapısıyla yapılacaksa, veri bütünleştirme süreci veri ambarı hazırlama aşamasında gerçekleştirilmesi gerekmektedir. Eğer direkt olarak veri kümesi alınacaksa veri bütünleştirme işlemi doğrudan veri kümesi üzerinde uygulanır.

Gereksiz veri

Farklı veri kaynaklarından veriler birleştirilince gereksiz(fazla) veri oluşabilir. Örneğin aynı nitelik farklı kaynaklarda farklı isimle bulunabilir. Böyle olduğu düşünülen nitelikler (nümerik nitelikler) arasında korelasyon hesaplaması yapılarak verilerin aynı olup olmadığına karar verilebilir.

Korelasyon katsayısı:

$$r = \frac{\sum x_i y_i}{\sqrt{\sum x_i^2 \sum y_i^2}} = \frac{\sum (X_i - \overline{X})(Y_i - \overline{Y})}{\sqrt{\sum (X_i - \overline{X})^2 \sum (X_i - \overline{Y})^2}}$$

veya

$$r = \frac{\sum XY - \frac{(\sum X)(\sum Y)}{n}}{\sqrt{\left(\sum X^2 - \frac{(\sum X)^2}{n}\right)\left(\sum Y^2 - \frac{(\sum Y)^2}{n}\right)}}$$

$|r| = 1$ ise tam ilişki,

$0.5 \ll |r| < 1$ ise kuvvetli ilişki,

$0 < |r| < 0.5$ ise zayıf ilişki,

$r = 0$ ise ilişki yok

Veri İndirgeme

Veri madenciliği uygulamalarında bazen çözümleme işlemi uzun süre alabilir.
Eğer çözümlemeden elde edilecek sonucun değişmeyeceğine inanılıyorsa veri sayısı ya da değişkenlerin sayısı azaltılabilir.
Veri indirgeme çeşitli biçimlerde yapılabilir.

- Veri birleştirme veya veri küpü
- Boyut indirgeme
- Veri sıkıştırma
- Örnekleme
- Genelleme

Veri analizlere tüm niteliklerin katılması uygun görülmeyebilir. Bu durumda nitelik seçimi yöntemleri uygulanarak boyut azaltma yoluna gidilebilir.
Örnekleme aşamasındaysa, büyük veri topluluğu yerine onu temsil eden daha küçük veri kümelerinin oluşturulması amaçlanır.
Genelleme verinin tek tek değil genel kavramlarla ifade edilmesini sağlar.

Veri Dönüştürme

Veriyi bazı durumlarda veri madenciliği çözümlemelerine aynen katmak uygun olmayabilir. Değişkenlerin ortalama ve varyansları birbirlerinden önemli ölçüde farklı olduğu taktirde büyük ortalama ve varyansa sahip değişkenlerin diğerleri üzerindeki baskısı daha fazla olur ve onların rolleri önemli ölçüde azaltır. Ayrıca değişkenlerin sahip olduğu çok büyük ve çok küçük değerler de çözümlemelerin sağlıklı biçimde yapılmasını engeller

Bu nedenle bir dönüşüm yöntemi uygulanarak söz konusu değişkenlerin normalleştirilmesi veya standartlaştırılması uygun bir yol olacaktır.

Normalizasyon

- Min-max normalizasyon
- Z-score normalizasyon

Veri Dönüştürme

❖ Min-max normalizasyon

min-max normalleştirmesi ile orijinal veriler yeni veri aralığına doğrusal dönüşüm ile dönüştürülürler. Bu veri aralığı genellikle 0-1 aralığıdır. Bu yöntem, veri içindeki en büyük ve en küçük sayısal değerin belirlenerek diğerlerini buna uygun biçimde dönüştürme esasına dayanmaktadır.

Söz konusu dönüştürme bağıntısı şu şekilde ifade edilmektedir:

$$X^* = \frac{X - X_{min}}{X_{max} - X_{min}}$$

Burada X^* dönüştürülmüş değerleri,
X gözlem değerlerini,
X_{min} en küçük gözlem değerini
X_{max} en büyük gözlem değerini ifade etmektedir.

Veri Dönüştürme

- Min-max normalizasyon

Aşağıdaki tabloda yer alan X değişkeni değerlerine *min-max* normalleştirme bağıntısını uygulayarak dönüştürmek istiyoruz. Bunun için, veri için önce aşağıdaki değerler belirlenir:

$$X_{min} = 30$$
$$X_{max} = 62$$

Bu değerlere dayanarak X örneğini birinci elemanı için şu şekilde bir hesaplama yapılır:

$$X^* = \frac{X - X_{min}}{X_{max} - X_{min}} = \frac{30 - 30}{62 - 30} = 0$$

Benzer biçimde diğer gözlemler için aynı hesaplamalar yapılır.

Tablo 2.1. *Min-max* normalleştirme dönüşümü sonucu elde edilen değerler

X	X^*
30	0,0000
36	0,1875
45	0,4688
50	0,6250
62	1,0000

Veri Dönüştürme

❖ Z-score normalizasyon

İstatistik çözümlemelerde sıkça kullanılan bir diğer dönüşüm biçimi z-score adıyla anılmaktadır. Bu yöntem, verilerin ortalaması ve standart hatası göz önüne alınarak yeni değerlere dönüştürülmesi esasına dayanmaktadır.

Söz konusu dönüşümlerde şu şekilde bir bağlantıya yer verilir:

$$X^* = \frac{X - \bar{X}}{\sigma_x}$$

Burada X^* dönüştürülmüş değerleri,
X gözlem değerlerini,
$\bar{X}$ verilerin aritmetik ortalamasını,
σx gözlem değerlerinin standart sapmasını ifade etmektedir.

Veri Dönüştürme

- Z-score normalizasyon

Önceki örnekte ele alınan veriye bu kez *Z-score* standartlaştırmasını uygulayarak dönüştüreceğiz. Bu amaçla önce aşağıdaki hesaplamaların yapılması gerekmektedir. Bunların birincisi $\overline{X}$ aritmetik ortalamanın bulunmasıdır.

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i = 44.6$$

Z-score standartlaştırma işlemi için X serisinin standart hatasının bunması gerekmektedir. Söz konusu hata şu şekilde hesaplanır:

$$\sigma_X = \sqrt{\frac{\sum_{i=1}^{n}(X_i - \overline{X})^2}{n-1}} = 12.44$$

Bu durumda birinci satır için Z-score dönüşümü şu şekilde olabilir:

$$X^* = \frac{X - \overline{X}}{\sigma_X} = \frac{30 - 44.6}{12.44} = -1.1735$$

Benzer biçimde diğer gözlemler içinde hesaplamalar yapılarak aşağıdaki tablo elde edilir.

Tablo 2.2. *Z-score* dönüşümü sonucu elde edilen değerler

X	X^*
30	-1,1735
36	-0,6912
45	0,0321
50	0,4340
62	1,3985

Veri Madenciliği Algoritmasını Uygulama

Veri madenciliği yöntemlerini uygulaya-bilmek için önceden bahsedilen işlemlerin uygun görünenleri yapılır. Veri hazır hale getirildikten sonra konuyla ilgili veri madenciliği algoritmaları uygulanır. Bu yöntemlere ilerleyen derslere ayrıntılı şekilde değinilecektir.

Sonuçları Sunum ve Değerlendirme

Veri madenciliği algoritması veriler üzerinde uygulandıktan sonra, sonuçlar düzenlenerek ilgili yerlere sunulur. Sonuçlar çoğu kez grafiklerle desteklenir. Örneğin bir hiyerarşik kümeleme model uygulanmış ise sonuçlar dendrogram adı verilen özel grafiklerle sunulur.

KAYNAKÇA

- Dr. Öğr. Üyesi Kadriye ERGÜN Ders Notları
- Veri Madenciliği Yöntemleri, Dr. Yalçın ÖZKAN
- Prof. Dr. Bülent TUGRUL Ders Notları

VERİ MADENCİLİĞİ

Bölüm 4

Öğr. Gör. Merve KESİM ÖNAL

Veri Madenciliği Yöntemleri

Veri madenciliği konusunda çok sayıda yöntem ve algoritma geliştirilmiştir. Bu yöntemlerin bir çoğu istatistiksel tabanlıdır. Veri madenciliği modellerini temel olarak şu şekilde gruplandırabiliriz.

- Sınıflandırma(Classification)
- Kümeleme(Clustering)
- Birliktelik kuralları (Association rules)

Elimizdeki verinin sınıf sayısı belli, hangi girdilerin hangi sonuçları ürettiği mevcutsa ve bu bilgileri kullanarak bir öğrenme yapılıyorsa bu gözetimli (supervised) öğrenmeye girmektedir. Sınıflandırma işlemi örnek olarak verilebilir.

Fakat elimizdeki verinin kaç sınıfa ayrıldığını, girdilerin hangi sonuçları ürettiğini bilmeden yani ham veriden bir anlam çıkarmaya çalışılıyorsa bu işlem gözetimsiz (unsupervised) öğrenmedir. Kümeleme işlemi örnek olarak verilebilir.

Sınıflandırma (Classification)

Verinin içerdiği ortak özelliklere göre ayrıştırılması işlemine sınıflandırma denir. Sınıflama veri madenciliğinde sıkça kullanılan bir yöntem olup veri tabanlarındaki gizli örüntüleri ortaya çıkarmakta kullanılır. Resim, örüntü tanıma, dolandırıcılık tespiti, kalite kontrol çalışmaları ve pazarlama alanlarında sınıflandırma tekniği sıkça kullanılır. Verinin sınıflandırılması için belirli bir süreç izlenir. Öncelikle var olan verinin bir kısmı eğitim amacıyla kullanılarak sınıflandırma kurallarının (sınıflandırma modeli) oluşturulması sağlanır. Daha sonra eğitim verisinden elde edilen bu kurallar test verisine uygulanarak sınanır.
Böylelikle yeni bir durum ortaya çıktığında nasıl karar verileceği belirlenir.

Sınıflandırma (Classification)

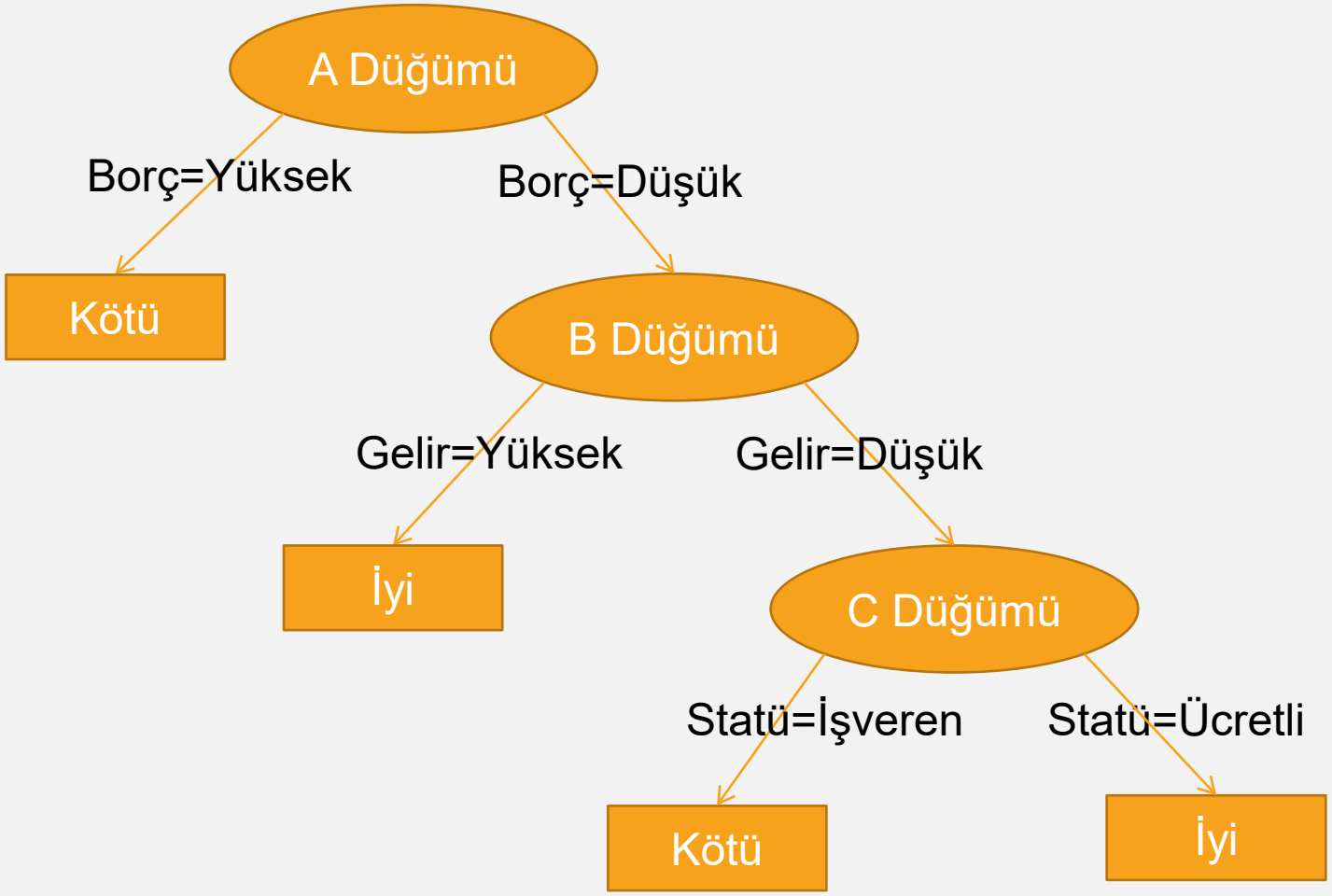
Örnek: Bir bankanın kredi verdiği müşterilerinin risk durumunu karar ağaçları yardımıyla ortaya koymak istediğini varsayalım.
Bu sayede belirli özelliklere sahip müşterilerden kredi talebi geldiğinde karar ağacı bilgilerine dayanarak kredi verip vermeme konusunda karar verilecektir.

MÜŞTERİ	BORÇ	GELİR	STATÜ	RİSK
1	Yüksek	Yüksek	İşveren	Kötü
2	Yüksek	Yüksek	Ücretli	Kötü
3	Yüksek	Düşük	Ücretli	Kötü
4	Düşük	Düşük	Ücretli	İyi
5	Düşük	Düşük	İşveren	Kötü
6	Düşük	Yüksek	İşveren	İyi
7	Düşük	Yüksek	Ücretli	İyi

Tablodaki veriler karar ağacının oluşturulması amacıyla eğitim verisi olarak kullanılacaktır. Söz konusu verileri kullanarak karar ağaçlarını oluşturmak üzere veri madenciliğinin çok sayıda yöntemi bulunmaktadır.

Sınıflandırma (Classification)

Verilen örnek için C4.5 algoritması yardımıyla karar ağacı aşağıdaki gibi oluşur.



Sınıflandırma (Classification)

Sınıflandırma Kural Tablosu

Eğer borç = yüksek ise **risk=> kötü**;
Eğer borç = düşük ve gelir = yüksek ise **risk => iyi**
Eğer borç = düşük ve gelir = düşük ve statü=işveren ise **risk => kötü**
Eğer borç = düşük ve gelir = düşük ve statü=ücretli ise **risk => iyi**

MÜŞTERİ	BORÇ	GELİR	STATÜ	RİSK
1	Düşük	Yüksek	İşveren	?
2	Düşük	Düşük	İşveren	?
3	Düşük	Düşük	Ücretli	?

Kümeleme(Clustering)

Kümeleme verilerin kendi aralarındaki benzerliklerin göz önüne alınarak gruplandırılması işlemidir. Bu özelliği nedeniyle pek çok alanda uygulanabilmektedir. Örneğin, pazarlama araştırmalarında yaygın biçimde kullanılmaktadır. Bunun dışında desen tanımlama, görüntü işleme, uzaysal harita verilerinin analizinde kullanılmaktadır.

Kümeleme(Clustering)

Örnek: Aşağıdaki gözlem değerlerini göz önüne alalım.
Bu gözlem değerinin X1 ve X2 gibi iki değişkeni bulunmaktadır. Bu gözlem değerlerine dayanarak verilerdeki kümelenmeleri belirlemek istiyoruz.

Gözlem	X1	X2
1	1	1
2	2	1
3	4	5
4	7	7
5	5	7

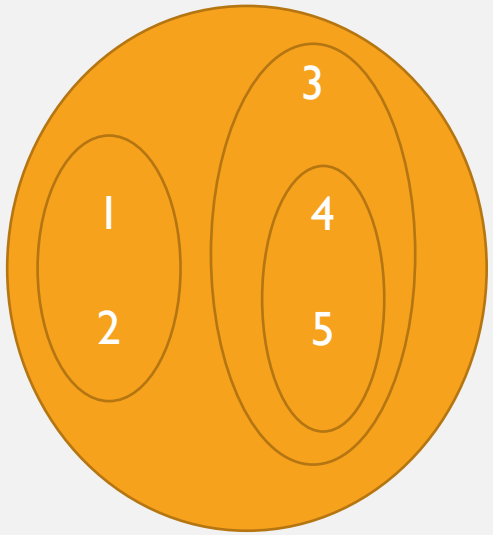
Kümeleri ortaya koymak üzere bir çok veri madenciliği ve istatistiksel yöntem bulunmaktadır.

Kümeleme(Clustering)

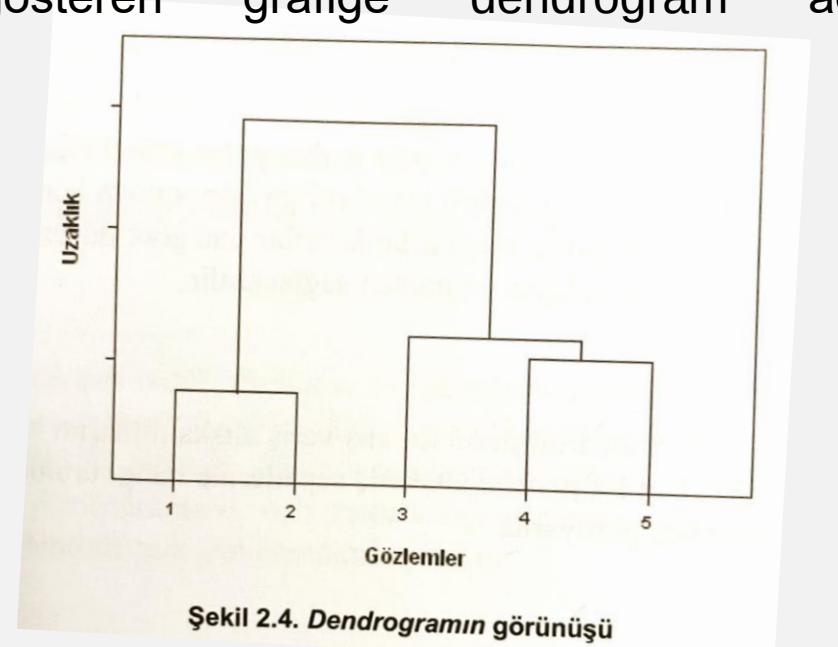
Söz konusu verilere hiyerarşik kümeleme yöntemlerinden «en yakın komşu» algoritmasını uygulandığında bir sonraki tabloda belirtilen kümeler elde edilir.

	Kümeler
Küme 1	(1,2)
Küme 2	(4,5)
Küme 3	(3,4,5)
Küme 4	(1,2,3,4,5)

Kümelere daha açık biçimde aşağıda belirtildiği biçimde de gösterilebilir.



Bu kümelere uygun olarak kümeleme grafiği çizilebilir. Kümeleri gösteren grafiğe dendrogram adı verilmektedir.



Şekil 2.4. *Dendrogramın* görünüşü

Birliktelik Kuralları (Association Rules)

Veri tabanı içinde yer alan kayıtların birbirleriyle olan ilişkilerini inceleyerek hangi olayların eş zamanlı olarak birlikte gerçekleşebileceklerini ortaya koymaya çalışan veri madenciliği yöntemleri bulunmaktadır. Bu ilişkilerin belirlenmesi ile birliktelik kuralları (association rules) elde edilir.

Birliktelik kuralları özellikle pazarlama alanında uygulama alanı bulmuştur. Pazar sepet analizleri adı verilen uygulamalar bu tür veri madenciliği yöntemlerine dayanmaktadır. Bu tür çözümlemelerden hareketle müşterilerin alışveriş alışkanlıkları belirlenmeye çalışılır.

Pazar sepet analizleri yardımıyla bir müşteri herhangi bir ürünü aldığında sepetine başka hangi ürünleri de koyduğu belirli bir olasılığa göre konulur. Birlikte satın alınan ürünler belirlendiğinde mağazalarda raflar ona göre düzenlenerek müşterilerin bu tür ürünlere daha kolayca erişmeleri sağlanabilir.

Birliktelik Kuralları (Association Rules)

Örnek: Bir mağazada alışveriş yapan müşterilerin alışveriş alışkanlıklarını belirlemek istediğimizi varsayalım. 5 müşterinin alışveriş sepetlerine hangi ürünleri koyduğunu bir sonraki slaytta görüyoruz.

Müşteri	Alışveriş sepetindeki ürünler
1	Makarna , yağ , meyve suyu , peynir
2	Makarna ketçap
3	Ketçap , yağ , meyve suyu, bira
4	Makarna, ketçap, yağ, meyve suyu
5	Makarna, ketçap, yağ, bira

Birliktelik Kuralları (Association Rules)

Bu verilerden yararlanarak birliktelik çözümlemeleri yapılır. Apriori algoritması yardımıyla aşağıdaki sonuçlar elde edilir.

{Ketçap, Meyve suyu} → {Yağ} (s=0.4, c=1.0)
{Ketçap, Yağ} → {Meyve suyu} (s=0.4, c=0.67)
{Yağ, Meyve suyu} → {Ketçap} (s=0.4, c=0.67)
{Meyve suyu} → {Ketçap, Yağ} (s=0.4, c=0.67)
{Yağ} → {Ketçap, Meyve suyu} (s=0.4, c=0.5)
{Ketçap} → {Yağ, Meyve suyu} (s=0.4, c=0.5)

Bu sonuçların herbir satırını şu şekilde yorumlayabiliriz:

- Ketçap ve Meyve suyunu birlikte alanlar mutlaka yağ da alıyorlar.
- Ketçap ve yağ satın alan müşteriler %67 olasılıkla meyve suyu da alıyorlar.
- Yağ ve meyve suyunu birlikte satın alanlar %67 olasılıkla ketçap da alıyorlar.
- Meyve suyu alanlar %67 olasılıkla ketçap ve yağ da satın alıyorlar.
- Yağ alanlar %50 olasılıkla ketçap ve meyve suyu da alıyorlar
- Ketçap alanlar %50 olasılıkla yağ ve meyve suyu da alıyorlar.

KAYNAKÇA

- Dr. Öğr. Üyesi Kadriye ERGÜN Ders Notları
- Veri Madenciliği Yöntemleri, Dr. Yalçın ÖZKAN
- Prof. Dr. Bülent TUGRUL Ders Notları