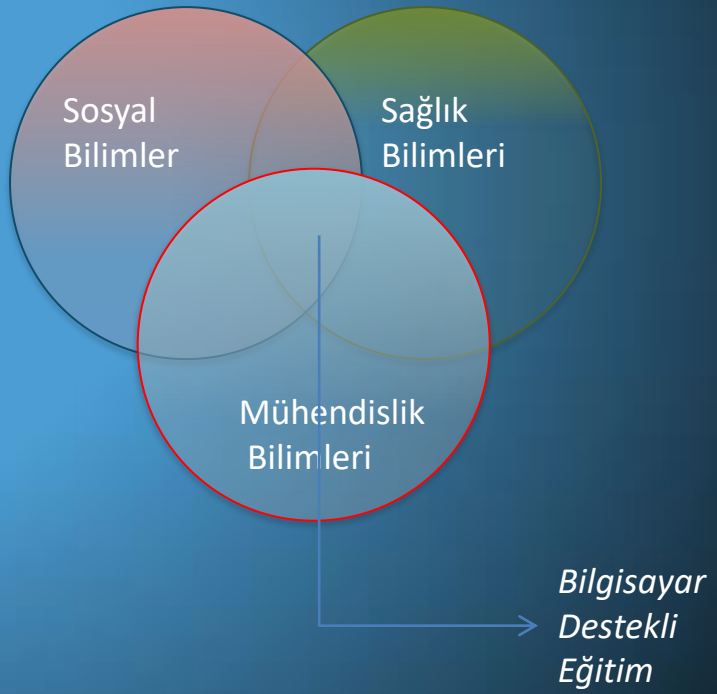
Veri Madenciliği
HAFTA 1

Dr. Öğretim Üyesi Deniz TANIR

İÇİNDEKİLER

İçerik

- VERİ MADENCİLİĞİNİN TARİHSEL GELİŞİMİ
- VERİ MADENCİLİĞİNE ETKİ EDEN DİSİPLİNLER
- VERİ MADENCİLİĞİ KAVRAMI
- VERİTABANLARINDA BİLGİ KEŞFİ SÜRECİ
- VERİ MADENCİLİĞİNDE KULLANILAN MODELLER
- VERİ MADENCİLİĞİNİN DİĞER VERİ ANALİZİ YAKLAŞIMLARI İLE KARŞILAŞTIRILMASI
- VERİ MADENCİLİĞİNİN UYGULANDIĞI ALANLAR



ÖĞRENME HEDEFLERİ

Bu üniteyi çalıştıktan sonra;

- Veri madenciliğinin tarihsel gelişimini özetleyebilecek,
- Veri madenciliğine etki eden disiplinleri betimleyebilecek,
- Veri madenciliği kavramını tanımlayabilecek,
- Veritabanlarında bilgi keşfi sürecini açıklayabilecek,
- Veri madenciliğinde kullanılan modellere ilişkin özellikleri özetleyebilecek,
- Veri madenciliğini diğer veri analizi yaklaşımları ile karşılaştırabilecek,
- Veri madenciliğinin uygulandığı alanları örnekleyebilecek bilgi ve becerilere sahip olabileceksiniz.

1.1. Kitaplar

- Altunkaynak, B. (2019). Veri Madenciliği Yöntemleri ve R Uygulamaları: Seçkin Yayıncılık.
- Veri Madenciliği, AÖF, 2019.
- Veri Madenciliği, İstanbul Üniversitesi Açık ve Uzaktan Eğitim Fakültesi.

1.2. İnternet Kaynakları

- https://archive.ics.uci.edu/ml/datasets.php
- https://www.r-project.org/
- https://www.cs.waikato.ac.nz/ml/weka/

Giriş

- İletişim ve bilişim teknolojilerinde yaşanan gelişmeler dünyada her şeyin hızla değişmesine neden olmaktadır.
- İster kâr amaçlı işletmeler, ister diğer kurum ve kuruluşlar açısından olsun, değişimlere ayak uydurabilmek başarı için önemli bir gerekliliktir.
- İşletmeler açısından ele alındığında bu değişimler; ekonomik koşullarda, iş yapma biçimlerinde, müşteri beklentilerinde, müşteri eğilimlerinde, rakiplerin stratejilerinde vb. ortaya çıkmaktadır. İşletmelerin bu değişimlere ayak uydurabilmesi, rakipleriyle yarışabilmesi ve varlıklarını başarılı bir biçimde sürdürebilmesi için, işletmelerde karar verici konumunda olan yöneticilerin, doğru kararlar vererek doğru stratejiler belirlemeleri gerekmektedir.
- Bu da ancak zamanında elde edilebilen doğru bilgilerin kullanımıyla mümkün olacaktır. Bu nedenle işletmelerin iş süreçlerinden ve işletme dışından elde ettikleri verileri karar verme sürecinde anlamlı bilgilere dönüştürebilmeleri önemlidir.
- Günümüzde bilişim teknolojisinde gelinen noktada çok büyük miktarda verinin kolaylıkla elde edilmesi ve kaydedilerek saklanması olanaklı hâle gelmiştir. Bununla birlikte veriler tek başlarına bir anlam ifade etmeyip belirli bir amaca yönelik olarak işlendiklerinde anlamlı bilgilere dönüşürler.
- Verilerin kolaylıkla elde edilip saklanabilmelerine karşın, bu verilerden anlamlı bilgilere ulaşabilmek aynı derecede kolay değildir. Anlamlı bilgilere ulaşabilmek amacıyla geçmişten beri kullanılan farklı yöntemler bulunmaktadır.
- Bununla birlikte verilerin analiz edilmesinde kullanılan geleneksel yöntemler veri miktarında meydana gelen büyük artış karşısında yetersiz kalmaya başlamıştır.
- Veri madenciliğinin ortaya çıkışı da büyük miktarda veriyi analiz edebilme ve işleyebilme ihtiyacından kaynaklanmıştır. Veri madenciliğinin amacı, çok büyük miktarda ve karmaşık durumdaki veriler içinden geleneksel yöntemlerle elde edilemeyecek bilgilere ulaşma ve bu bilgileri rakiplere fark yaratacak kararlarda kullanabilmeye olanak sağlamaktır.
- Buradan anlaşılabileceği üzere veri madenciliği tek başına çözümün kendisi olmayıp çözüme ulaştıracak kararın verilmesine destek sağlayacak bilgilerin ortaya çıkarılmasında kullanılan bir araçtır.

Neden veri madenciliği?

- Bilgisayarların ucuzlayıp aynı zamanda çok güçlü hale gelmeleri
- Teknolojinin gelişimiyle bilgisayar ortamında ve veritabanlarında tutulan veri miktarının da artması (terabyte to petabyte)
- Yeni veri toplama yolları
  - Otomatik veri toplama aletleri, veritabanı sistemleri, bilgisayar kullanımının artması
- Büyük veri kaynakları
  - İş dünyası: Web, e-ticaret, alışveriş, hisse senetleri, …
  - Bilim dünyası: Uzaktan algılama ve izleme, bioinformatik, simülasyonlar..
  - Toplum: haberler, digital kameralar, YouTube, Facebook…
- Ticari rekabet baskısının artması
  - Kişiselleştirilmiş ürünler, CSR yönetimi
- **Veri içinde boğuluyoruz, ancak bilgi elde edemiyoruz!!!**

KGS

VERİ MADENCİLİĞİNİN TARİHSEL GELİŞİMİ

- Veri madenciliğinin tarihi bilgisayarların hayatımıza girmesiyle başlamıştır.
- 1950'li yıllardaki ilk bilgisayarların geliştirilme ve kullanım amacı sayım ve karmaşık hesaplamaları kolaylıkla yapabilmekti.
- Daha sonra kullanıcıların ihtiyaçları doğrultusunda, bilgisayarlar veri depolama işlemleri için de kullanılmaya başlanmıştır.
- Verilerin depolanması ihtiyacı ile birlikte, 1960'lı yıllardan itibaren teknoloji dünyası veri tabanı kavramı ile tanışmıştır. 1960'ların sonunda ise basit öğrenmeli bilgisayarlar geliştirilmiştir.
- Buna karşın, günümüzdeki sinir ağlarının temeli olarak bilinen perseptron'ların yalnızca çok basit olan kuralları öğrenebileceği, bazı basit mantıksal işlemlerde ise yetersiz kaldığı 1969'da Minsky ve Papert tarafından ortaya konulmuştur.

VERİ MADENCİLİĞİNİN TARİHSEL GELİŞİMİ

- Zaman içinde giderek büyüyen veri tabanlarının organizasyonu, düzenlenmesi ve yönetimi de doğal olarak zorlaşmıştır.
- Bu zorlukların üstesinden gelebilmek amacıyla ise veri modelleme kavramı ortaya atılmıştır.
- İlk veri modelleri; Hiyerarşik Veri Modeli ve Ağ Veri Modeli olarak adlandırılan basit veri modelleridir.
- 1970'lerde İlişkisel Veri Tabanı Yönetim Sistemleri uygulamaları kullanılmaya başlanmış, bu konuyla ilgilenen uzmanlar basit kurallara dayanan uzman sistemler geliştirmişler ve basit anlamda makine öğrenimini sağlamışlardır.
- 1980'lerde veri tabanı yönetim sistemleri yaygınlaşmış ve pek çok farklı alanda uygulanır olmuştur.
- Özellikle işletmeler, müşterileri, rakipleri ve ürünlerine ilişkin verileri düzenli biçimde saklamak amacıyla veri tabanları oluşturmuştur.

VERİ MADENCİLİĞİNİN TARİHSEL GELİŞİMİ

- 1990'lara gelindiğinde ise artık araştırma konusu; veri miktarının sürekli katlanarak arttığı veri tabanları içinden, faydalı bilgilerin nasıl çıkarılabileceği konusudur. Bu amaçla pek çok çalışma ve yayın yapılmıştır.
- Bu çalışmalardan en önemlisi, 1989'da yapılan KDD (Knowledge Discovery in Database) IJCAI-89 Veri Tabanlarında Bilgi Keşfi Çalışma Grubu toplantısıdır.
- 1991 yılında ise KDD (IJCAI)-89'un sonuç bildirgesi sayılabilecek "Knowledge Discovery in Real Databases: A Report on the IJCAI-89 Workshop" makalesi ile Bilgi Keşfi ve Veri Madenciliği ile ilgili temel tanım ve kavramlar ortaya konmuştur.
- Bu makaleden sonra süreç daha da hızlanmış ve 1992 yılında veri madenciliği için ilk yazılım geliştirilmiştir.
- 2000'li yıllarda veri madenciliği sürekli gelişmiş ve hemen hemen tüm alanlara uygulanmaya başlanmıştır.
- Günümüze geldiğimizde veri madenciliğinin pek çok alanda yaygın olarak kullanıldığını görebiliriz. Karar verme sürecinde ihtiyaç duyulan veri analizini gerçekleştirdiği için, operasyonel kararların ötesinde stratejik karar verme süreçlerinde de oldukça önemli bir yere sahiptir. İşletmeler, günümüzde yoğun olarak kullandıkları Müşteri İlişkileri Yönetimi (CRM) ve Kurumsal Kaynak Planlaması (ERP) gibi uygulamalar ve teknikler aracılığıyla veri madenciliği yapmaktadır.

VERİ MADENCİLİĞİNİN TARİHSEL GELİŞİMİ

1950'ler	• İlk bilgisayarlar (sayım ve hesaplama amaçlı)
1960'lar	• Verilerin depolanması ve veritabanları • Perseptronlar
1970'ler	• İlişkisel Veritabanı Yönetim Sistemleri • Basit kurallara dayanan uzman sistemler ve makine öğrenimi
1980'ler	• Büyük miktarda veri içeren veri tabanları • SQL sorgu dili
1990'lar	• Veritabanlarında Bilgi Keşfi Çalışma Grubu ve Sonuç Bildirgesi • Veri madenciliği için ilk yazılım
2000'ler	• Tüm alanlar için veri madenciliği uygulamaları

VERİ MADENCİLİĞİNE ETKİ EDEN DİSİPLİNLER

- Veri madenciliği işlemleri ile amaçlanan; veri analizinde geleneksel istatistiksel yöntemler yerine bu yöntemlerin yetersizliklerini giderecek yeni yaklaşımları, bilgisayar algoritmalarının kullanımı ile uygulamak ve istenilen analizi hızlı ve sağlıklı bir biçimde gerçekleştirmektir.
- Bu yeni yaklaşımların temelinde istatistik, makine öğrenimi, veritabanı sistemleri önemli bir yer tutmaktadır. İstatistik, verilerin analizi ve değerlendirilmesi konusunda geçmişten günümüze yoğun bir biçimde kullanılan bir disiplindir.
- Bilgisayar sistemlerinde hem donanım hem de yazılım alanında sağlanan gelişmeler doğal olarak istatistik alanını da etkilemiştir. İstatistiksel çalışmaların bilgisayar desteğiyle daha güçlü biçimde yapılması, daha önce gerçekleştirilmesi çok mümkün olmayan istatistiksel araştırmaları ve analizleri yapılabilir hâle getirmiştir.
- Bu anlamda 1990'lardan sonra, ilgilenilen verinin yığınlar içinden çekilip çıkarılması ve analizinin yapılarak kullanıma hazır hâle getirilmesi sürecinde istatistik, veri madenciliği ile ortak bir platformda ve sıkı bir çalışma birlikteliği içinde olmuştur.
- Veri madenciliği çalışmalarında etkili olan ve yapay zekâ çalışmalarının da temelini oluşturan makine öğrenimi, kısaca bilgisayarların bazı işlemlerden çıkarsamalar yaparak yeni işlemler üretmesi olarak tanımlanabilir. Makine öğrenimi, insan öğrenmesinde söz konusu olan özelliklerin algoritmalar yardımıyla bilgisayarlara da uygulanabileceği ve bilgisayarların da insanlar gibi öğrenebileceği düşüncesini temel alan bir disiplindir.

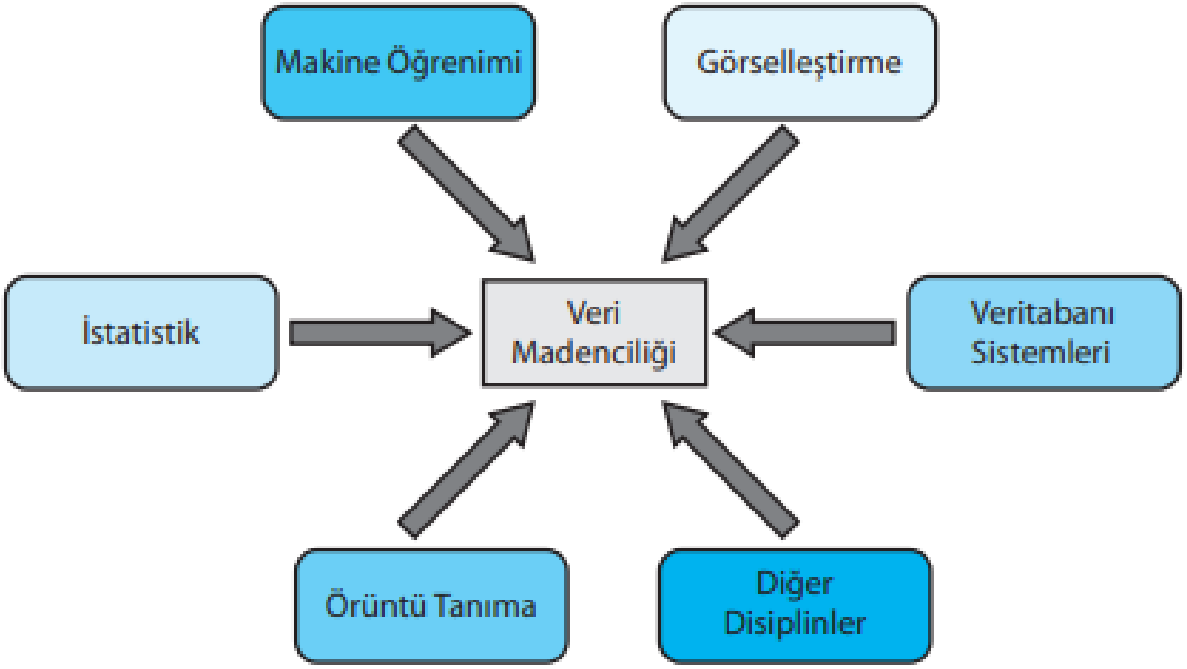
VERİ MADENCİLİĞİNE ETKİ EDEN DİSİPLİNLER

- İnsanlar nasıl öğrenirler? İnsanlar çocukluk dönemlerinden itibaren öğrenmeye başlarlar. Bu, etraflarında gördükleri tüm nesneleri gözlemleme ve bu gözlemler aynı türde nesneler üzerinde tekrarlandıkça nesneleri kavramlara dönüştürme biçiminde gerçekleşir.
- Aynı türde nesnelere ilişkin farklı örnekleri görmeyi, incelemeyi sürdürdükçe nesneye ilişkin kavram netleşir ve benzer örnekleri ilgili nesne sınıfına konumlandırarak bir sınıflama modeli oluşturur.
- Örneğin, ilk kez kedi gören bir çocuğa gördüğü varlığın bir kedi olduğu söylendiğinde bu bilgiyi alan çocuk, başka bir gün başka bir kedi gördüğünde bir önceki deneyimi hatırlayarak o varlığın kedi olduğunu düşünür. Bu deneyim tekrarlandıkça öğrenme de gerçekleşmiş olur.
- Artık herhangi bir kedi görüldüğünde, bu kedi öncekilerden farklı özellikler (daha küçük, daha farklı renkte) taşısa da çocuk o varlığın kedi olduğunu bilerek ortak özellikleri temel alarak kedi tanımlamasını yapabilir.
- Makine öğrenimi de bilgisayarların kendisine algoritmalar yoluyla verilen kuralları uygulaması ve büyük veri kümeleri içinden örnekler çıkararak verileri bu kurallara göre sınıflamaları, tanımlamaları ve dolayısıyla öğrenmeleri olarak ifade edilebilir. Bu öğrenmeler sonucunda çıkarımlarda bulunarak geçmiş veri örnekleri yardımıyla gelecekte daha iyi sonuçlar üretme konusunda veri madenciliği uygulamasına katkıda bulunurlar.

VERİ MADENCİLİĞİNE ETKİ EDEN DİSİPLİNLER

- Veri madenciliğinde söz konusu diğer bir disiplin olan görselleştirme; verilerin, tablolar ve grafikler gibi görseller yardımıyla sunulmasını sağlayan teknolojileri ifade eder.
- Görselleştirme; verilerin daha kolay anlaşılmasına, analiz edilmesine ve geleceğe yönelik tahminlerde bulunulmasına önemli katkı sağlamaktadır.
- Veri madenciliğinde kullanılan görselleştirme teknikleri ilk zamanlarda sadece iki boyutlu serpilme ve serpilme matris çizimleri ya da üç boyutlu grafikler biçimindeydi.
- Ancak zaman içinde, verilerin öznitelik sayılarındaki artış klasik istatistiğin sunduğu iki veya üç boyutlu grafiklerin yetersiz kalması sonucunu da birlikte getirmiştir.
- Bu durum da çok daha fazla boyutun görselleştirilmesine imkân sağlayan yeni grafik araçlarının geliştirilmesine neden olmuştur.
- YerKonum veri analizi, sinyal işleme, görüntü analizi gibi teknikler görselleştirme amacıyla kullanılan tekniklere verilebilecek örneklerdir

VERİ MADENCİLİĞİNE ETKİ EDEN DİSİPLİNLER



Veri Madenciliğinin Etkileşimde Olduğu Disiplinler

VERİ MADENCİLİĞİ KAVRAMI

- Veri madenciliği
  - Basit ve açık olmayan, önceden bilinmeyen ve yararlı olan örüntülerin ya da bilginin çok büyük miktarlardaki veriden çıkarılması
  - Sorgulama ya da basit istatistik yöntemler veri madenciliği değildir.
  - Veri madenciliği terimi ne kadar doğru?
- KNOWLEDGE DISCOVERY FROM DATA (KDD) (VERİDEN BİLGİ KEŞFİ)
- Alternatif isimler
  - Knowledge discovery (mining) in databases (KDD), knowledge extraction, data/pattern analysis, data archeology, data dredging, information harvesting, business intelligence, etc.

- Teoride veri madenciliği bilgi keşfi işleminin bir parçasıdır
- Pratikte veri madenciliği ve veriden bilgi keşfi aynı anlamda kullanılır

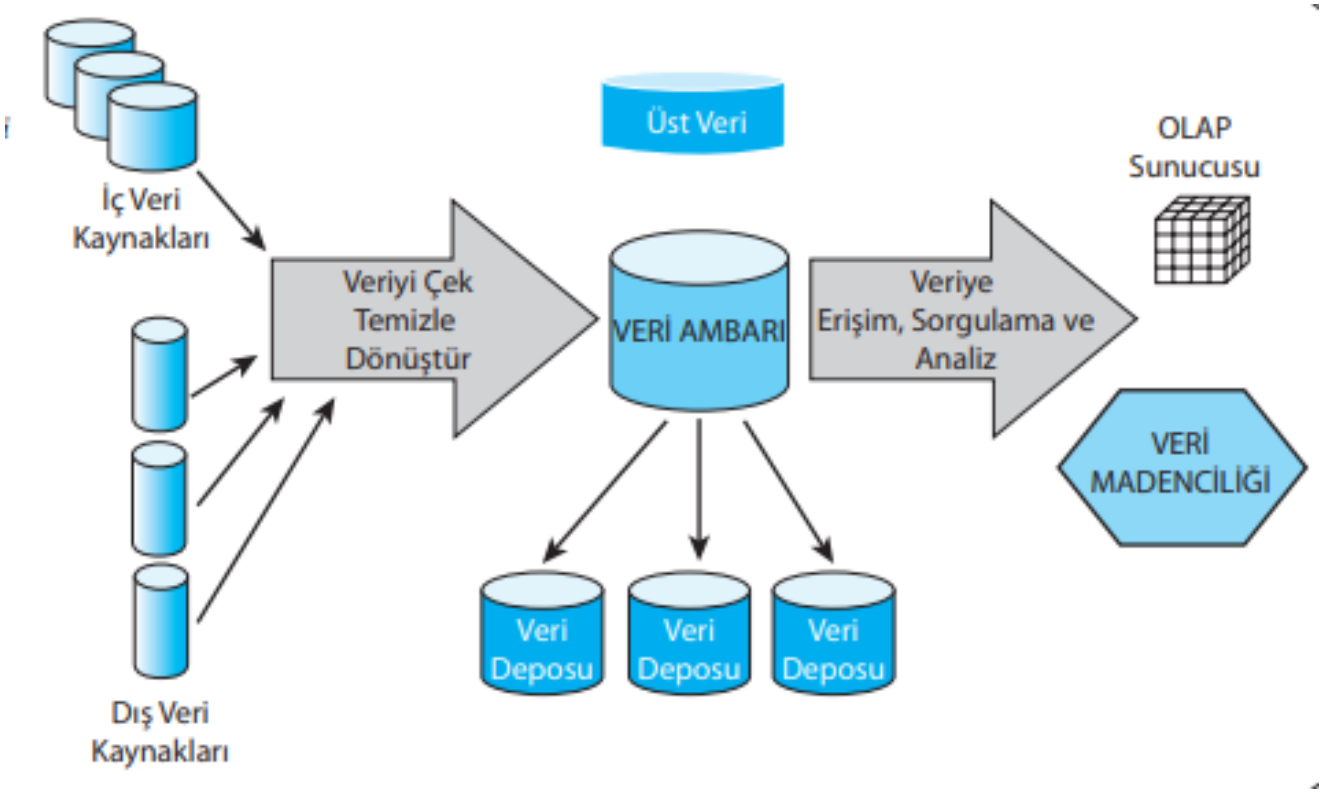
VERİ MADENCİLİĞİ KAVRAMI

- Veri madenciliği çalışmaları yapmak için var olması gereken iki temel öge veri ve veritabanıdır.
- Bununla birlikte burada sözü edilen veritabanı, işletmelerin günlük kayıtlarının yer aldığı ve işlemsel veritabanı olarak adlandırılan veri tabanları değildir.
- Daha doğru bir ifadeyle işlemsel veritabanları veri madenciliği uygulamalarında doğrudan kullanılmaz.
- Bu veritabanlarında yer alan veriler birtakım işlemlerden geçirilerek veri madenciliği için kullanılabilir, hazır hâle getirilir.
- İşte işletmelere ait veritabanlarının, belirli bir amaca göre konu odaklı olarak düzenlenmiş, veri madenciliğinde doğrudan kullanılabilir duruma getirilmiş hâli veri ambarı olarak tanımlanır.
- İşletme çalışanları günlük işlemlerini sürdürebilmek, farklı yönetim düzeyindeki yöneticiler ise operasyonel, taktik ya da stratejik kararlar verebilmek için farklı veri kaynaklarına ihtiyaç duyarlar.

VERİ MADENCİLİĞİ KAVRAMI

- İşletmelerdeki veri kaynakları iç kaynaklar ve dış kaynaklar biçiminde sınıflandırılabilir.
- İç veri kaynakları; işletmenin kendi iş süreçlerinden elde ettiği verilerin yer aldığı kaynaklardır. Örneğin, üretim bölümü, satın alma ve pazarlama bölümleri kendi günlük işlemlerine ilişkin kayıtları tutarak birer iç veri kaynağı oluştururlar.
- Bunun dışında işletmelerde günlük faaliyetleri ya da her düzeyde alınacak kararları doğrudan etkileyecek dış kaynaklı verilere de ihtiyaç duyulur. Örneğin işletmelerin içinde bulundukları sektöre ilişkin veriler, istatistik kurumlarının yayınladığı raporlar, yasal düzenlemeler, döviz kuru vb. gibi sermaye piyasasına ilişkin veriler dış kaynaklı verilere verilebilecek örnekler arasındadır. Veri ambarları, söz konusu bu iç ve dış kaynaklı verilerin biraraya getirilmesi ile oluşturulan özel veritabanlarıdır.
- Bununla birlikte verilerin birleştirilmesi gelişigüzel bir işlem değildir. Veriler farklı kaynaklardan elde edildiği için veriler arasındaki uyumsuzlukların, tutarsızlıkların giderilmesi ve verilerin amaca uygun, kullanılabilecek biçime dönüştürülmesi gereklidir

VERİ MADENCİLİĞİ KAVRAMI



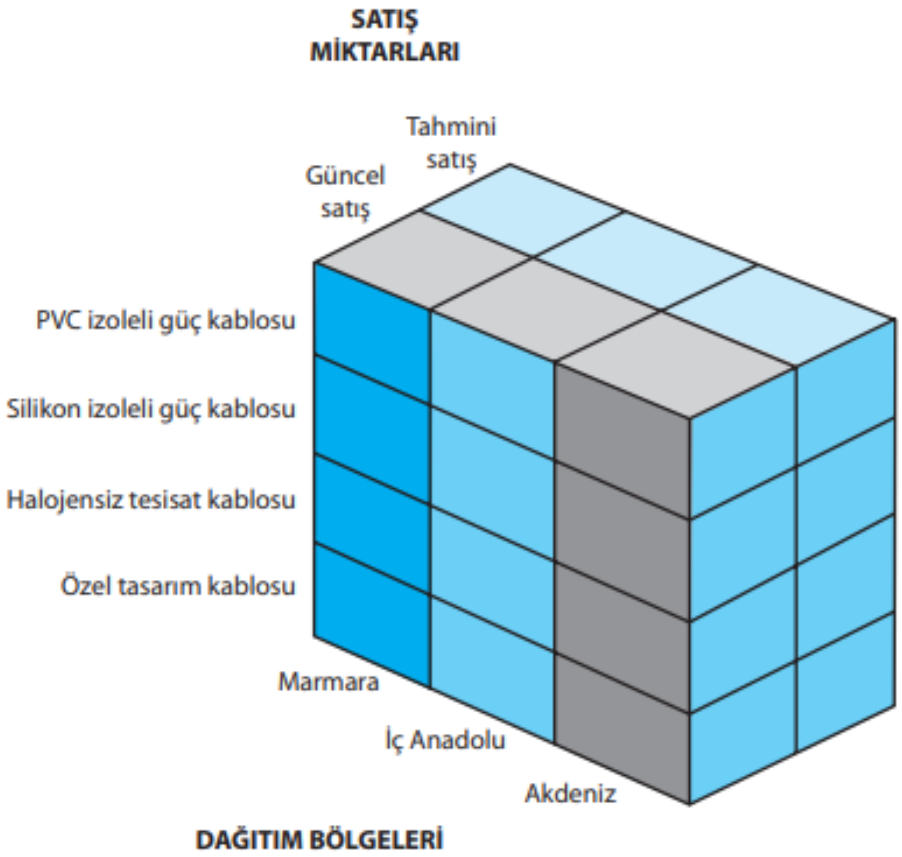
VERİ MADENCİLİĞİ KAVRAMI

- İngilizce karşılığı meta data olan üst veri, veri ambarında yer alan veriler hakkındaki tanımlamalar olup veri ambarına ilişkin veri kataloğu olarak düşünülebilir.
- Farklı kaynaklarda veri martı kavramı ile ifade edilen veri deposu (data mart) kavramı ise veri ambarının bir alt kümesi olup işletmenin yalnızca belirli bir bölümünü ya da belirli bir iş sürecini, daha özel bir fonksiyon alanını ilgilendiren parçasıdır.
- Veri ambarı tüm işletmeyi ilgilendirirken veri deposu tek bir konuya ya da özel bir amaca yönelik verileri içerir.
- İşletmeler günlük faaliyetlerine ilişkin basit sorgulamaları ve analizleri işlemsel veritabanları üzerinde kolaylıkla gerçekleştirebilirler.
- Buna karşın, çok yönlü veri analizi ve sorgulama yapmak istediklerinde normal veri analizi ve sorgulamadan farklı bir sistem kullanırlar.
- Çevrimiçi Analitik İşleme olarak adlandırılan bu sisteme kısaca OLAP (OnLine Analytical Processing) denir. OLAP uygulamaları veri ambarından çekilen veriler üzerinde gerçekleştirilir. OLAP sorgulamaları işlemsel veri tabanlarında gerçekleştirilen basit analiz ve sorgulamalardan farklı olarak, veriyi çok boyutlu biçimde analiz eder ve analiz sonucunda yöneticilere stratejik kararlarında destek olacak yararlı bilgiler sunar.
- İşletmelerin geleceklerine yönelik önemli kararlarında, karşı karşıya kaldıkları problemler basit yapıda olmayıp çok yönlü, karmaşık, analitik sorgulamalar gerektirecek yapıda ortaya çıkarlar.
- Bu tür problemlerin çözümü için günlük veri analizi ve sorgulamalar doğal olarak yetersiz kalacaktır.

VERİ MADENCİLİĞİ KAVRAMI

- OLAP işlemini gerçekleştirmek üzere veri ambarı ile etkileşim içinde olan OLAP sunucuları, karmaşık analitik sorguların kısa sürede gerçekleştirilmesine imkân veren çok boyutlu veri modelini kullanırlar.
- Örneğin farklı kablo üretimi yapan ve bunları farklı bölgelere dağıtan bir işletmenin ürettiği kablo türleri; PVC izoleli güç kablosu, silikon izoleli güç kablosu, halojensiz tesisat kablosu ve özel tasarım kablosu biçiminde olsun. Önceki ay "halojensiz tesisat kablosundan toplam ne kadar satıldığı" sorusuna cevap almak için, işlemsel veritabanlarında basit sorgulama yapmak yeterli olacaktır.
- Buna karşın her bir satış bölgesinde ne kadar halojensiz tesisat kablosu satıldığını öğrenmek ve hedeenen satış miktarı ile gerçekleşen satış miktarı sonuçlarını karşılaştırmak daha karmaşık bir sorgu yapısı olacaktır.
- İkinci sorgu türüne cevap almak amacıyla OLAP sorgulama işleminin yapılması gerekecektir.
- Bu örnekte; ürünün ne olduğu, fiyatı, maliyeti, satış bölgesi, satış zamanı vb. verinin farklı boyutlarını temsil eder. Bu nedenle işletme yöneticisi Ağustos ayında İç Anadolu bölgesinde ne kadar halojensiz tesisat kablosu satıldığını, bu miktarın bir önceki ay ile ve geçen yılki Ağustos ayı ile ve satış tahminleriyle karşılaştırmasının sonuçlarını öğrenmek için çok boyutlu veri analizine imkân sağlayan OLAP sistemini kullanabilecektir.
- OLAP, verdiğimiz örnekten çok daha karışık veri analizi ve sorgulamalarına da çok hızlı bir sürede cevap verebilme olanağı sağlamaktadır

VERİ MADENCİLİĞİ KAVRAMI



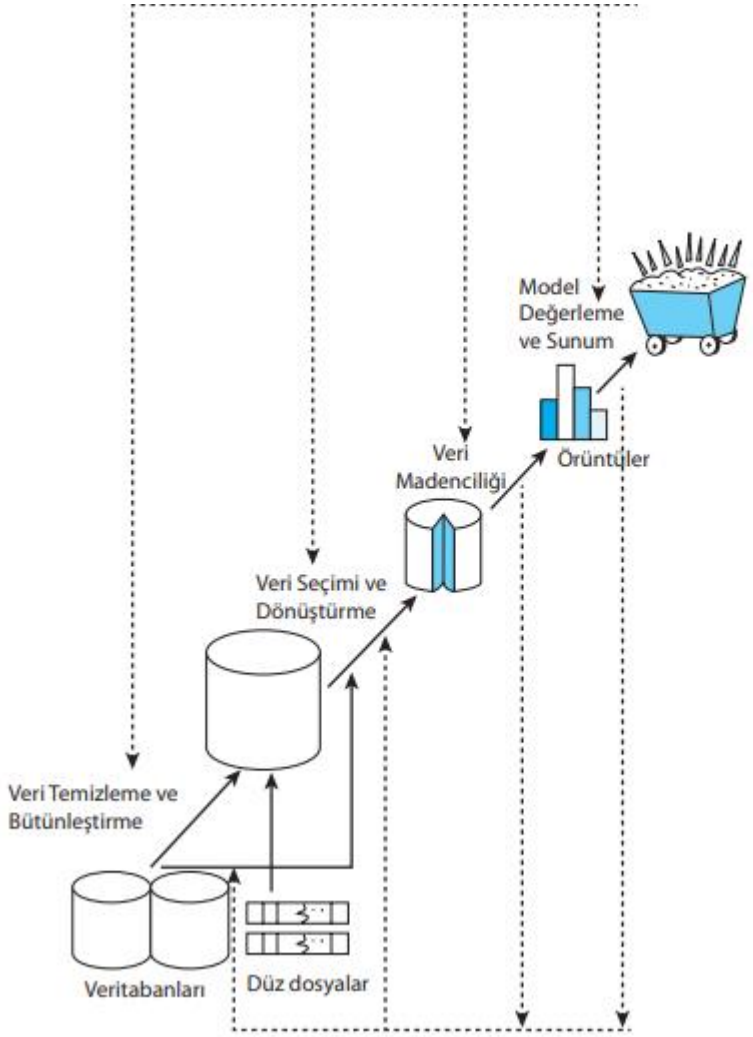
VERİ MADENCİLİĞİ KAVRAMI

- Veri madenciliği kavramı için çeşitli tanımlar yapılmıştır. Bu tanımlardan bir kısmı aşağıda verildiği gibidir:
- Veri madenciliği, büyük miktardaki veri yığınları üzerinde analiz yaparak veriler arasında var olan ve geleceğin tahmin edilmesine yardımcı olacak anlamlı ve yararlı ilişki ve kuralların bilgisayar yazılımları aracılığıyla aranması faaliyetleridir. Bu anlamda veri madenciliği, çok büyük miktardaki veriler arasındaki bağlantıları inceleyerek aralarındaki ilişkiyi ortaya çıkaran ve veritabanları içinde açıkça fark edilemeyen, gizli kalmış yararlı bilgilerin açığa çıkarılmasını sağlayan veri analizi tekniğidir.
- Veri madenciliği, çeşitli analiz araçlarını kullanarak veriler arasındaki örüntü (desen) ve ilişkileri keşfederek, bunları doğru tahminler yapmak için kullanan bir süreçtir. Veri madenciliğinin amacı, geçmiş faaliyetleri analiz ederek bu analizleri geleceğe yönelik tahminlerde temel almak ve karar vermeye destek olacak modeller oluşturmada kullanmaktır. Buna göre veri madenciliği, büyük miktarda veri içinden, gizli kalmış, değeri olan, kullanılabilir bilgileri açığa çıkarmak ve bu bilgileri özellikle stratejik kararlarda destek sağlayacak biçimde elde etmek amacıyla kullanılmaktadır.
- Veri madenciliği, veri analizi için, gelişmiş ve karmaşık araçlar kullanarak yığın veri kümeleri içinden daha önceden bilinmeyen olgu ve olayları keşfetmek ve veriler arasındaki mantıklı ilişkileri ve kalıpları ortaya çıkarmak amacıyla yapılan çalışmalardır. Burada vurgulanması gereken önemli nokta, veri madenciliği ile elde edilecek bilginin daha önceden bilinmeyen yeni keşfedilen olmasıdır. Önceden bilinmeyen bilgi, önceden tahmin bile edilemeyen bilgi anlamındadır. Bu anlamda veri madenciliği, tahmin edilen ya da farklı teknikler yardımıyla daha önceden ulaşılmış sonuçların doğruluğunu ispatlamak amacıyla kullanılan bir araç değildir. Diğer tekniklerden temel farkı, daha önce düşünülmemiş hiç akla gelmemiş sonuçları ortaya çıkarmasıdır.
- Veri madenciliği, istatistiksel ve matematiksel tekniklerle birlikte örüntü tanıma teknolojilerini kullanarak çeşitli depolama ortamlarında kayıtlı bulunan veri yığınları üzerinde gerçekleştirilen elemeler sonucunda anlamlı yeni korelasyon, örüntü ve eğilimlerin keşfedilmesi sürecidir.

VERİTABANLARINDA BİLGİ KEŞFİ SÜRECİ

- Veritabanlarında bilgi keşfi ifadesi ilk kez 1989 yılında “Veritabanlarında Bilgi Keşfi Çalışma Toplantısı”nda ortaya atılmıştır. Bu toplantıda; bilginin, veri keşfi sürecinin sonunda elde edilen ürün olduğu vurgulanmıştır.
- Veritabanlarında Bilgi Keşfi, veriden faydalı bilginin keşfedilmesi sürecinin tamamıdır. Veri madenciliği ise bu sürecin bir adımı olup veriden örüntülerin belirlenmesi ve aktarımı için özel algoritmaların uygulanması işlemlerine karşılık gelmektedir.

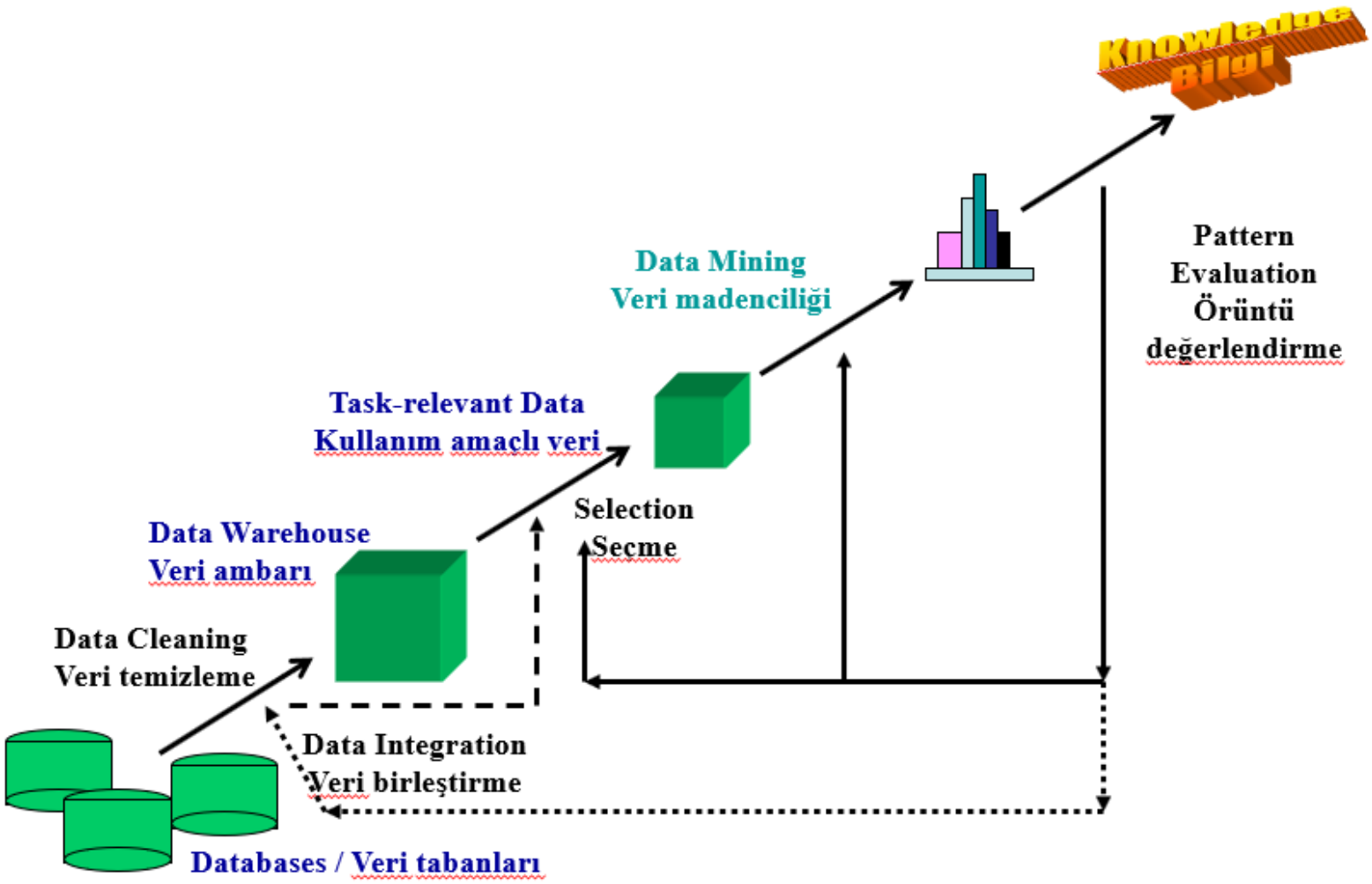
VERİTABANLARINDA BİLGİ KEŞFİ SÜRECİ



Veritabanlarında Bilgi Keşfi Sürecinin Adımları

Kaynak: *Han ve Kamber, (2012), s.7.*

VERİTABANLARINDA BİLGİ KEŞFİ SÜRECİ



VERİTABANLARINDA BİLGİ KEŞFİ SÜRECİ

- Veritabanlarında Bilgi Keşfi sürecinde, işlemsel veritabanlarında depolanmış olan verinin sorgulama ve analiz için uygun hâle getirilmesi işlemleri yürütülür. Veritabanlarında Bilgi Keşfi sürecinde izlenmesi gereken temel aşamalar aşağıdaki gibi sıralanabilir.

1. Amacın Tanımlanması
2. Veriler Üzerinde Ön İşlemlerin Yapılması
3. Modelin Kurulması ve Değerlendirilmesi
4. Modelin Kullanılması ve Yorumlanması
5. Modelin İzlenmesi Sıralanan bu aşamalara bütünsel olarak bakıldığında, veri madenciliği sürecinde;
   - Veri madenciliği öncesindeki işlemler
   - Veri madenciliği işlemleri
   - Veri madenciliği sonrasındaki işlemler biçiminde bir uygulamanın söz konusu olduğu görülebilir.

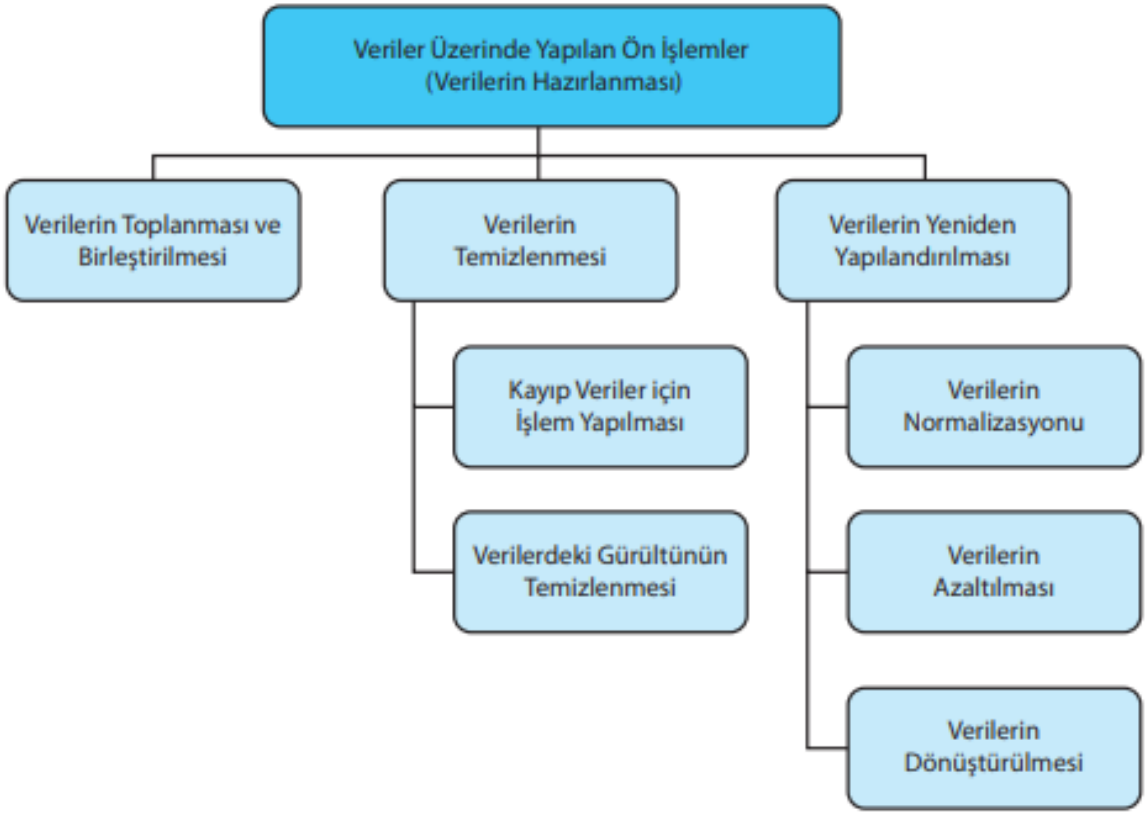
Amacın Tanımlanması

- Bu aşamada, işletmenin ya da kurumun veri madenciliğini hangi amaca yönelik olarak gerçekleştirmek istediği belirlenir.
- Söz konusu amaç bir problemi ortadan kaldırmaya odaklanmış ve açık bir biçimde ifade edilmiş olmalıdır.
- Buna ek olarak, elde edilecek sonuçların başarı düzeylerinin nasıl ölçüleceği de tanımlanmalıdır.
- Bu aşamada ayrıca, süreç sonunda yapılacak değerlendirme ve öngörülerin yanlış olması durumunda katlanılacak maliyetlere ve doğru olması durumunda elde edilecek kazanımlara ilişkin tahminlere de yer verilmelidir.

Veriler Üzerinde Ön İşlemlerin Yapılması

- Veriler üzerinde ön işlemler yapılması, verilerin veri madenciliği için hazırlanması anlamındadır.
- Veri madenciliği ile ulaşılması hedeflenen sonuçların kalitesi veritabanlarında yer alan verinin kalitesi ile yakından ilişkilidir.
- Bu nedenle veri madenciliği işlemleri öncesinde verilerin analize hazır hâle getirilmesi oldukça önemli bir aşamadır.
- Buna bağlı olarak veriler üzerinde yapılan ön işlemler, veri tabanlarında bilgi keşfi sürecinin en fazla zaman alan aşamasıdır.
- Bir sonraki aşama olan modelin kurulması aşamasında herhangi bir sorunun ortaya çıkmaması, veri üzerindeki ön işlemlerin ne kadar titizlikle yapıldığına bağlıdır.
- Ön işlemler aşamasında yeterli özenin gösterilmemesi, model kurma aşamasından ön işlemler aşamasına tekrar tekrar geri dönülmesine ve verinin yeniden düzenlenmesine neden olacaktır.
- Bu durum, ön işlemlerle verinin hazırlanması ve modelin kurulması aşamaları için harcanan enerji ve zamanın, bilgi keşfi sürecinin toplamı içinde büyük bir paya sahip olmasına neden olacaktır.
- Veriler üzerindeki ön işlemler genel olarak;
  - Verilerin toplanması ve birleştirilmesi
  - Verilerin temizlenmesi
  - Verilerin yeniden yapılandırılması biçiminde sınıflandırılabilir.

Veriler Üzerinde Ön İşlemlerin Yapılması



Verilerin Toplanması ve Birleştirilmesi

- Verilerin veri madenciliğine hazırlanabilmesi için yapılması gereken ilk şey doğal olarak verilerin belirlenmesidir.
- Bu yapılırken öncelikle tanımlanan amaca ve probleme uygun verilerin neler olduğu ve bu verilerin hangi kaynaklarda yer aldığı araştırılır.
- Bu belirleme sonrası veriler bulundukları farklı kaynaklardan toplanır ve birleştirilir.
- Gerekli verilerin toplanmasında öncelikli olarak kurumun kendi veritabanı ve veri kaynaklarından yararlanılır.
- Daha önceden de belirtildiği üzere bu tür veriler iç kaynaklı verilerdir. Bunun yanı sıra, istatistiksel bilgiler, finansal raporlar, menkul kıymet değerleri gibi bilgilerin yer aldığı kamuya ait kurumsal veri tabanlarından veya veri pazarlayan farklı kuruluşların veri tabanlarından da yararlanılabilir.
- Örneklenen veri kaynakları ise dış veri kaynaklarıdır. Verilerin hangi kaynaklardan, hangi koşullar altında ve hangi yöntemlerle toplandığı önemlidir.

Verilerin Temizlenmesi

- Veritabanlarından alınan kayıtların bir kısmında, diğer kayıtlarda var olan bazı veriler eksik olabilir.
- Örneğin, işletme çalışanlarına ait kişisel bilgilerin tutulduğu kayıtlarda çoğunluğun doğum tarihi yer alırken bazı çalışanlara ait kayıtlarda doğum tarihi verisi eksik olabilir.
- Müşteriler arasında yapılan bir araştırmada yaşını, kilosunu ya da gelirini belirtmek istemeyen müşteriler olabilir.
- Veritabanı kayıtları içindeki böylesi eksik veriler kayıp veri olarak adlandırılır.
- Bunun yanı sıra kayıtlarda yer alan bir kısım veriler doğru olamayacak kadar uç değerlerde, dolayısıyla yanlış girilmiş olabilir.
- Örneğin, doğum tarihi 1974 olan bir kişi için bu değer 1074 olarak kaydedilmiş olabilir.
- 1074 değeri gibi aşırı uç değerler aykırı değer, bu şekildeki uç verilerin geneli de gürültülü veri olarak nitelendirilir.
- Bunların dışında kayıtlarda yer alan bazı veriler (örneğin, olmayan bir ürün adı ya da stok numarası gibi) tamamıyla yanlış ya da anlamsız olabilir.
- Verilerin temizlenmesi aşamasında dikkat edilmesi gereken diğer bir konu veriler arasındaki uyumsuzluktur.

Verilerin Temizlenmesi

- Verilerin temizlenmesi, kayıp ya da eksik değerleri tamamlamak, aykırı değerleri belirleyerek gürültüyü ortadan kaldırmak ve verilerdeki tutarsızlıkları, uyumsuzlukları gidermek için kullanılan birçok yaklaşımı ve tekniği kapsar. İzleyen kesiminde bu yaklaşımlardan bir kısmı hakkında bilgi verilmiştir. Kayıp verilerin neden olacağı olumsuzlukları ortadan kaldırmak amacıyla kullanılan yaklaşımlar:
- a. Kayıp veri içeren kaydı veri kümesinden çıkarmak: Bu yaklaşım, kayıp veri içeren kayıt sayısının toplam kayıt sayısı içinde çok küçük bir orana karşılık gelmesi ve kayıp verilerin sonuçlara önemsenmeyecek bir etki yapması durumunda kullanılabilecek bir yaklaşımdır. Kayıp veri içeren kayıt sayısının çok olması durumunda, sonuçları olumsuz etkileyeceğinden bu yaklaşım önerilmez.
- b. Kayıp verileri tek tek yazmak: Veri kümesi küçük, kayıp verilere ulaşmak mümkün, yeterince zaman mevcut ve kayıp verilere mutlaka ihtiyaç duyuluyorsa bu yaklaşım kullanılır. Söz konusu bu durumların dışında bu yaklaşımı kullanmak gereksiz zaman kaybına neden olacaktır.

Verilerin Temizlenmesi

- c. Kayıp verilerin hepsi için aynı veriyi girmek: Örneğin, yapılan bir hane halkı araştırmasında bazı bireyler gelirlerini belirtmekten kaçınmış olabilir. Ya da bir işletmenin müşterileri arasında yaptığı bir araştırmada bazı müşteriler doğum tarihi bilgisini yazmamış olabilir.
- Bu durumda gelir bilgisinin bulunmadığı tüm kayıtlar için, yok anlamında Y harfi, doğum tarihinin eksik olduğu tüm kayıtlar içinse, eksik anlamında E harfi kayıp veri yerine yazılabilir. Kayıp verilerin bu yaklaşımla giderilmesi farklı sonuçlar verebilir. Gelir bilgisinin Y olması ya da doğum tarihinin E olması belirleyici ya da ayırt edici bir özellikmiş gibi görünebilir. Diğer bir ifadeyle bu veriler kullanılan veri madenciliği algoritmasını yanıltabilir.
- Bunun tersine bu yaklaşım bazı durumlarda veri madenciliğinin gerçek amacına hizmet ederek gizli bilgilerin keşfedilmesini de sağlayabilir. Örneğin, doğum tarihi girmemiş olan müşterilerin, en çok para harcadıkları ürünlerin yaşlanma karşıtı ürünler olduğu sonucu elde edilebilir. Kayıp verilerin aynı veri değeri ile temsil edilmesinde kayıp veriyi temsil etmek üzere sayısal bir değer de atanabilir.
- Örneğin, 9 rakamı veri girişi yapılmadığı anlamında kullanılabilir. Veri madenciliği algoritması bu değeri göz ardı ederek yok sayabilir ve analiz buna göre gerçekleştirilebilir. Burada dikkat edilmesi gereken bir nokta, ister sayısal ister alfabetik kodlama yapılsın kayıp verileri temsil için seçilen değerin, analizde anlamlı olacak (diğer verileri temsil eden) başka bir değere karşılık gelmediğinden emin olunmasıdır.

Verilerin Temizlenmesi

- d. Kayıp veri yerine tüm verilerin ortalama değerinin girilmesi: Örnek olarak ücret verisi eksik olan kayıtlar için, diğer kayıtlarda yer alan ücret verilerinin ortalaması yazılabilir. Ortalama değeri bulunurken tüm verilerin ortalaması yerine belirlenen bir sınıfın ortalamasınının alınması daha uygun olacaktır. Daha açıklayıcı olması açısından Tablo 1.2'de görülen bir muhasebe bürosu çalışanları kira bilgileri örnek verisi üzerinden devam edelim. Tablo 1.2'de 6 sıra numarasına sahip Alanönü mahallesinde ikamet eden 9 yıllık çalışanın ödediği kira miktarı bilgisi elde edilememiş olsun. Bu kayıp veri için doğrudan aritmetik ortalama değeri bulunacak olursa verileri girilmiş olan diğer beş kişinin kira miktarları ortalaması hesaplanır. Bu değer, veri için 901 olarak hesaplanabilir. Bu yaklaşım yerine verileri kendi arasında sınıflayıp, sınıf ortalaması alma yoluna da gidilebilir. Bu durumda önce uygun sınıfın belirlenmesi gerekecektir. Sınıf olarak Alanönü mahallesinde ikamet eden çalışanlar seçilirse bu grupta yer alan iki çalışanın kira ortalama değeri 865 olacaktır

Sıra No	Çalışma Süresi	Kira Miktarı	Mahalle
1	5	850	Yenibağlar
2	6	675	Emek
3	3	780	Alanönü
4	9	950	Alanönü
5	14	1250	Batıkent
6	9		Alanönü

Verilerin Temizlenmesi

- e. Kayıtlarda yer alan diğer değişkenler yardımıyla kayıp verilerin tahmin edilmesi: Veri kümesinde yer alan ve eksik olmayan kayıtlardaki veriler kullanılarak kayıp veriler tahmin edilebilir.
- Kayıp verilerin tahmininde, regresyon analizi, zaman serileri analizi, Bayesyen sınıflandırma, karar ağaçları, maksimum beklenti vb. biçiminde sıralanan teknik ya da yöntemler kullanılabilir

Verilerin Temizlenmesi

- Verilerdeki gürültünün temizlenmesi amacıyla kullanılan yaklaşımlar ise aşağıdaki gibidir:
- a. Bölümleme yöntemiyle gürültünün temizlenmesi: Bu yöntemde üzerinde analiz yapılacak veriler önce küçükten büyüğe doğru sıralanır.
- Daha sonra veriler eşit sayıda eleman içeren gruplara bölünür. Her grupta bulunan verilerin ortalama değeri ya da medyan değeri bulunarak grupta yer alan tüm veriler ortalama ya da medyan değeri ile değiştirilerek düzeltme yapılır.
- Örneğin; 24, 18, 7, 27, 31, 24, 11, 37, 28 biçimindeki bir veri seti üzerinde bölümleme yöntemiyle düzeltme yapmak istendiğinde, öncelikle verilerin küçükten büyüğe doğru sıralanması gerekir.
- Sıralı veri 7, 11, 18, 24, 24, 27, 28, 31, 37 olacaktır. Veri eşit sayıda birim içerecek biçimde gruplara bölündüğünde izleyen yapı oluşacaktır.

Verilerin Temizlenmesi

Bölüm 1:	7	11	18
Bölüm 2:	24	24	27
Bölüm 3:	28	31	37

Bu aşamada ilgili bölümlendirmeye göre her grup ortalaması tespit edilerek birimlerin gerçek değerleri yerine kullanılabilir. Bu durumda izleyen veri yapısı ortaya çıkacaktır.

Bölüm 1:	12	12	12
Bölüm 2:	25	25	25
Bölüm 3:	32	32	32

Buradan görüldüğü üzere her grupta yer alan orijinal değerler o grubun ortalama değeri ile değiştirilerek düzeltme sağlanmıştır (Verilerin düzeltilmesi amacıyla ortalama değeri yerine medyan değerinin kullanımı da tercih edilebilirdi).

Verilerin Temizlenmesi

b. Sınır değerleri kullanılarak gürültünün temizlenmesi: Bu yöntemde de veriler önceki yöntemde olduğu gibi küçükten büyüğe doğru sıralanarak eşit bölümlere ayrılır. Daha sonra, her bölümün en küçük ve en büyük değerli verileri sınır değerleri olmak üzere bölüm içindeki her bir değer üst sınır ya da alt sınır değerlerinden hangisine yakınsa o sınır değeri ile değiştirilir. Bu durumu bir önceki örnekteki verileri kullanarak gösterirsek;

Bölüm 1:	7	11	18
Bölüm 2:	24	24	27
Bölüm 3:	28	31	37

Sınır değerleri ile düzeltme yapıldıktan sonra veriler;

Bölüm 1:	7	7	18
Bölüm 2:	24	24	27
Bölüm 3:	28	28	37

Verilerin Temizlenmesi

Bu yöntemde kullanılabilecek diğer bir yaklaşımda, veri kümesi içindeki en büyük veri ile en küçük veri değerlerinin birbirinden farkının, kümedeki eleman sayısına bölünmesiyle elde edilen değerin o küme elemanlarına atanmasıdır. Buna göre örneğimizde

Bölüm 1:	7	11	18 = (18-7)/3 = 3,67
Bölüm 2:	24	24	27 = (27-24)/3 = 1
Bölüm 3:	28	31	37 = (37-28)/3 = 3

olacaktır. Bunun sonucunda her bölümdeki değer hesaplama sonucu bulunan değerler değiştirilerek;

Bölüm 1:	3,67	3,67	3,67
Bölüm 2:	1	1	1
Bölüm 3:	3	3	3

biçiminde düzeltilmiş olur

Verilerin Temizlenmesi

- c. Kümeleme yöntemiyle düzeltme yapılması ve gürültünün temizlenmesi: Bu yaklaşım aykırı değerlerin ortaya çıkarılması ve düzeltilmesinde kullanılır. Buna göre, veri setinde yer alan veriler birbirlerine olan benzerlik ve yakınlıklarına göre kümelere ayrılır. Bu kümeleme işlemi sırasında uç değer olarak kabul edilen bazı veriler hiçbir küme içinde yer alamayacaktır. Bu şekilde belirlenen her bir aykırı değere, en yakın olduğu kümenin ortalama değeri veya en küçük ya da en büyük değeri atanarak aykırı veriler temizlenmiş olur.
- d. Regresyon yöntemiyle düzeltme yapılması ve gürültünün temizlenmesi: Verilerde gürültünün temizlenmesi amacıyla kullanılabilecek diğer bir yöntem, değişken değerlerini bir fonksiyon yardımıyla ilişkilendiren regresyon yönteminin kullanılmasıdır. Doğrusal regresyon iki nitelik ya da iki değişken arasındaki en uygun doğruyu bulmayı içerir. Bu nedenle bir nitelik (ya da değer) diğerinin tahmin edilmesinde kullanılabilir. Çoklu doğrusal regresyon doğrusal regresyonun genişletilmiş biçimi olup ikiden fazla nitelik (değişken) söz konusu olduğunda kullanılır ve analiz çok boyutlu düzlemde gerçekleştirilir.

TEMEL KAVRAMLAR

Verilerin Yeniden Yapılandırılması

- Veri madenciliği amacıyla kullanılan model, teknik ve algoritmalar belirli yapılardaki veriler üzerinde uygulanabilir.
- Örneğin, bir kısım algoritmalar yalnızca sayısal değerler üzerinde çalışırken bir kısım algoritmalar da yalnızca kategorik değerler üzerinde çalışır.
- Bunun dışında bazı algoritmalar ise yalnızca 0 ve 1'lerle temsil edilen veriler üzerinde çalışır. Bu nedenle eldeki verilerin kullanılacak algoritmaya uygun hâle getirilmesi, diğer bir ifadeyle yeniden yapılandırılması gerekir. Bu amaçla gerçekleştirilen işlemler; verilerin normalizasyonu, verilerin azaltılması ve verilerin dönüştürülmesi başlıkları altında incelenebilir.
- a. Verilerin normalizasyonu: Farklı değerlerdeki verilerin 0,0-1,0 gibi aralıklardaki değerlerle temsil edilmesi işlemine normalizasyon denir. Normalizasyon işlemi için kullanılabilen yöntemlerden bir kısmı; min-maks normalizasyonu, sıfır-ortalama normalizasyonu ve ondalıklı normalizasyon biçiminde sıralanabilir.
- b. Verilerin azaltılması: Bellek kapasitelerinin artmış olması ve bilgisayar sistemlerinin ucuzlaması sonucunda veri tabanlarında gerekli olsun ya da olmasın çok miktarda veri tutulmaktadır. Bu aşırılık veri ön işlemleri aşamasında veri analizi çalışmalarını zorlaştırmaktadır. Bu nedenle veriler yapılandırılırken gerçekleştirilen bir diğer işlem de verilerin temel özelliklerini kaybetmeden miktar olarak azaltılmasıdır. Verilerin azaltılması, veri kümesi içinde gereksiz olduğu düşünülen verinin kaldırılması biçiminde olabileceği gibi daha çok birden fazla değişkenin birleştirilerek tek bir değişkenle ifade edilmesi biçiminde gerçekleştirilir. Bu işleme veri indirgeme işlemi de denilmektedir. Verilerin azaltılması amacıyla geliştirilen çeşitli yöntemler bulunmaktadır. Bu yöntemlerden bazıları; boyut sayısını azaltma, veri sıkıştırma, temel bileşenler analizi, faktör analizi biçiminde sıralanabilir.

Verilerin Yeniden Yapılandırılması

- c. Verilerin dönüştürülmesi: Bu aşama, analize konu olan veri kümesinin gerekli veriyi içermesi ancak verinin kullanılan algoritmaya uygun yapıda olmaması durumunda gerçekleştirilir. Verilerin gösterim biçimi kullanılan algoritmanın etkinliği üzerinde çok önemli bir paya sahiptir. Buna göre verilerin dönüştürülmesi, algoritmada doğrudan kullanıcak biçimde verinin kendi içinde yeniden düzenlenmesini ifade etmektedir.
- Örnek olarak günlük işlemlerin kaydedildiği işlemsel veritabanlarındaki verilerin büyük çoğunluğu sayısal veriler olup sürekli değerler alır. Veri madenciliğinde kullanılan bazı algoritmalar her bir veriyi ayrı bir değişken olarak ele aldığından buna göre işlem yapar. Bu durumu açıklamak üzere, bir işletmede çalışanların maaşlarının 1.500 TL ile 5.000 TL arasında değiştiğini varsayalım. Bu durumda bazı veri madenciliği algoritmaları, söz konusu alt limit ve üst limit değerlerini ve bu iki değer arasındaki tüm değerleri ayrı değişkenler olarak ele alacaktır. Bunun sonucu olarak veriler üzerindeki işlem süresi artacak, elde edilecek sonuçlar gereğinden ayrıntılı ve uzun olacaktır. Bu nedenle sürekli nitelikteki bu tür verilerin kesikli ve kategorik veri biçimine dönüştürülmesi gerekir. Bu amaçla, belirlenen farklı değerler arasındaki maaşlar, düşük, orta, yüksek biçiminde kategorize edilebilir.

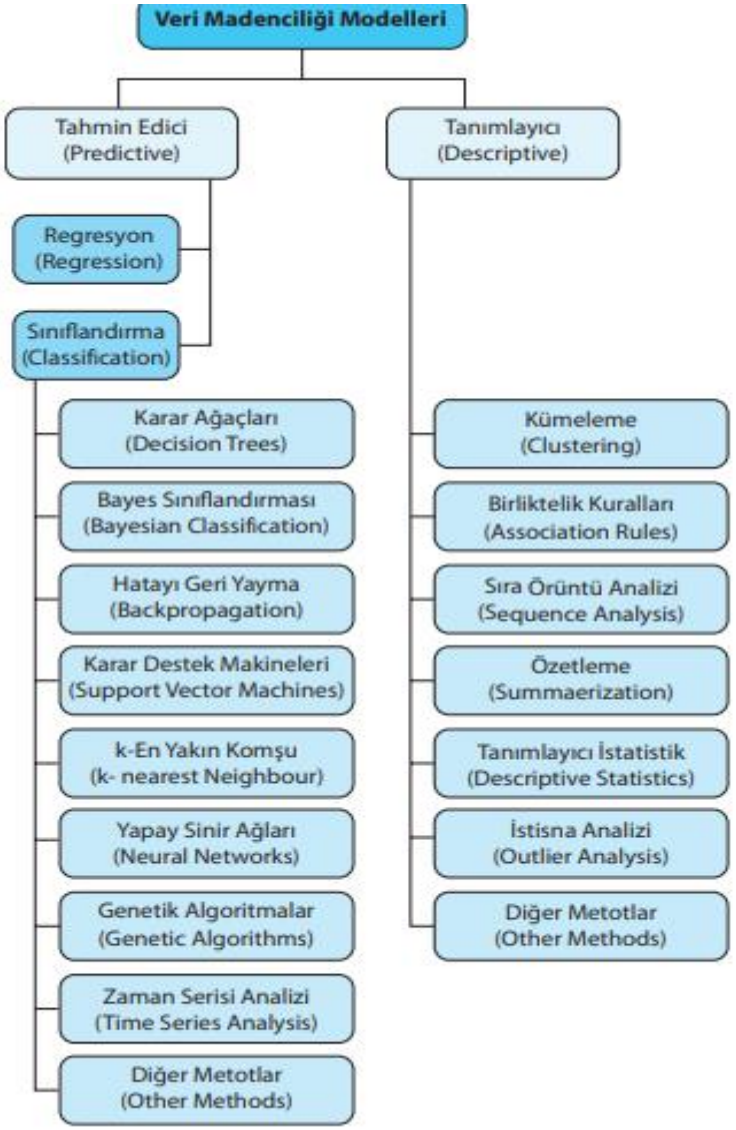
Modelin Kurulması ve Değerlendirilmesi

- Bu aşama, veri madenciliği modelinin kurulduğu ve geçerli bir model olup olmadığının değerlendirildiği aşamadır.
- Tanımlanan amaca ulaşmada kullanılacak en uygun modelin belirlenmesi için, çok sayıda modelin denenmesi gerekebilir.
- Bu nedenle, veri hazırlama ve model kurma aşamaları, en iyi olduğu düşünülen modele ulaşılıncaya kadar tekrarlanır.
- Kurulan modelin geçerli bir model olup olmadığı da çeşitli açılardan sınanmalıdır. Yanlış model kurulması ya da modelde kullanılan verilerin tutarsız, eksik ya da sıra dışı değerlerden oluşması modelin geçerliliğini etkileyen önemli nedenlerdir.

VERİ MADENCİLİĞİNDE KULLANILAN MODELLER

- Veri madenciliği farklı görevleri yerine getirmek amacıyla pek çok farklı algoritmayı kullanır. Aydın (2007) “Algoritmalar veriyi inceler ve incelenen verinin özelliklerine en uygun modeli belirler” ifadesini kullanmaktadır.
- Veri madenciliğinde kullanılan algoritma, teknik ve modeller sonuçta birer bilgisayar yazılımıdır. Bu yazılımlar matematiksel altyapı ve algoritma adımları olarak birbirlerinden farklı olsalar da ortak olan bazı özellikleri vardır.
- Bu özelliklerin başında veri madenciliği yazılımlarının öğrenme özelliği gelir.
- Söz konusu bu yazılımlar kendilerine verilen örnek veriler üzerinde inceleme yaparak kullandıkları algoritmalarla bu verilerden bazı sonuçlar ve kurallar çıkarırlar.
- Yazılımın veriler üzerinde yaptığı bu inceleme işlemine öğrenme adı verilir.
- Daha sonra yazılım bu çıkarımları verilerin kalan kısmına uygulayarak ne kadar öğrendiği konusunda kendini sınar.
- Bu sınama sonucunda eğer gerekli görürse başlangıçta yaptığı çıkarımlarını yeniler. Yenilenen çıkarımlar (sonuçlar, kurallar) üzerinde yapılan ayrı bir işlemle doğrulama gerçekleştirilir. Doğrulama işleminden sonra ise aşırı öğrenme olup olmadığı da kontrol edilir. Aşırı öğrenme algoritmanın çıkardığı kuralların sadece üzerinde çalıştığı veriler için geçerli olmasını, dışarıdan başka verilere uygulandığında ise geçersiz olması durumunu ifade eder. Aşırı öğrenme durumunda, mevcut veriler üzerinde uygulandığında doğru sonuç veren çıkarım ve kurallar, dışarıdan gelen yeni veriler üzerinde tam sonuç veremeyecektir (Silahtaroğlu, s.50).
- Veri madenciliğinde kullanılan modeller;
    - Tahmin edici modeller
    - Tanımlayıcı modeller olmak üzere temelde iki başlık altında incelenebilir.

VERİ MADENCİLİĞİNDE KULLANILAN MODELLER



Veri Madenciliği Modellerinin Sınıflandırılması

Kaynak: *Kaya ve Köymen, (2008), s.161*

Tahmin Edici Modeller

- Tahmin edici modeller; eldeki verilerden hareketle bir model geliştirilmesi ve geliştirilen bu model kullanılarak önceden sonuçları bilinmeyen veri kümeleri için sonuçların tahmin edilmesini amaçlar. Kısaca bilinenden yola çıkarak bilinmeyeni tahmin etme çabasıdır. Tahmin edici modeller özellikle karar verme süreci açısından büyük önem taşır.
- Örneğin bankalar, müşterilerinin önceki dönemlerde kullanmış oldukları kredilere ilişkin verilerine kendi veritabanlarından ulaşabilirler.
- Bu verilerden hareketle, müşterilerinin daha sonraki kredi taleplerinde kredi borcunu geri ödeyip ödemeyeceği, ya da ödemelerde düzenli olup olmayacağı konusunda tahminlerde bulunabilirler.
- Başka bir örnek olarak bir hastanede herhangi bir hastalığa ilişkin verilerin kaydedildiği veritabanı üzerinde tahmin edici modellerin uygulanması verilebilir.
- Buna göre hastalığa ilişkin geçmiş olaylardan elde edilen tıbbi veriler ve hastanın durumu bir arada değerlendirilerek bir tahmin modeli oluşturulabilir. Bu model kullanılarak, hastaneye yeni gelen bir hastanın hastalığına ilişkin tahmin, testler sonrası oluşan tıbbi veriler kullanılarak yapılabilir.

Tahmin Edici Modeller

- Tahmin edici modellere ilişkin yazılımlardaki öğrenme, daha çok bir insanın öğrenme biçimine benzetilebilir. Tahmin edici modeller de kendisine verilen veritabanını inceler ve bu veritabanındaki temel unsurları birbirine benzeterek tanımlamaya, onları isimlendirmeye ve sınıflamaya çalışır. Burada öğrenme işlevinin denetimli ve denetimsiz öğrenme olarak ikiye ayrıldığını söylemek gerekir.
- Denetimli öğrenmede, öğrenici konumunda olan algoritmaya, nesneler, nesnelerin özellikleri ve yine bu nesnelerin tanımlanmış, daha sonra tahmin edilmesi istenecek olan değişkenleri verilir.
- Veri madenciliğindeki nesneler veritabanındaki her bir kayıttır. İlgili algoritma ya da yazılım veritabanına girilmiş olan nesnelerin özelliklerini değerlendirir. Veri madenciliği algoritmaları, veritabanındaki ya da daha doğru bir ifadeyle veri madenciliği için oluşturulmuş veri ambarındaki nesnelerin özelliklerini nesnelerin isimleriyle ilişkilendirerek bu nesnelerin birbirinden farklı ya da benzer, aynı sınıftan nesneler olduklarını bulur ve öğrenir. Daha sonra, kendisine verilen değişik özellikleri değerlendirerek bu özelliğe sahip olan nesnenin ismini tahmin eder.
- Denetimli öğrenmenin tersi durumuna ise denetimsiz öğrenme denir. Denetimsiz öğrenmede nesnelerin özellikleri verilirken tahmin için kullanılacak herhangi bir parametre diğer bir ifadeyle nesnelerin isimleri verilmez. (Silahtaroğlu, s.52) Tahmin edici modeller kendi içinde regresyon (eğri uydurma) modelleri ve sınıflandırma modelleri biçiminde ikiye ayrılır.
- Regresyon Modelleri: Bilindiği gibi regresyon, bağımsız değişkenler ile bağımlı değişkenler arasındaki ilişkiyi en iyi tanımlayan fonksiyonu elde etmek için uygulanan istatistiksel tekniktir.
- Regresyon analizinde model, değişkenler arasındaki ilişkinin net bir biçimde gösterilebildiği bir fonksiyon ile temsil edilir. Sınıflandırma Modelleri: Sınıflama, veri sınıfı ve kavramlarını tanımlama ve ayırt etmeyi sağlayan bir model kümesini bulma sürecidir. Sınıflandırmada, veriler istatistik ve/ veya makine öğrenimi yöntemleri kullanılarak önceden belirlenen sınıflara atanır. Sınıflama modelleri, sınıflar önceden incelenen veriler aracılığıyla oluşturulduğundan, denetimli öğrenme modelleridir.
- Regresyon ve sınıflandırma modellerinden en yaygın kullanılanlar; karar ağaçları, yapay sinir ağları, genetik algoritmalar, zaman serisi analizi, k-en yakın komşu ve Bayes sınıflandırması biçiminde sıralanabilir.

Tahmin Edici Modeller

- İlerleyen derslerde bu kavramlar kısaca açıklanacaktır.
- 1. Karar ağaçları: Karar ağaçları, sınıflandırma problemlerinde en çok kullanılan algoritmalardan biridir. Bunun nedeni, karar ağaçlarının yapılandırılmasının ve anlaşılmasının diğer algoritmalara göre daha kolay olması ve veritabanı sistemleri ile daha kolay uyum sağlayabilmesidir.
- Karar ağaçları biçiminde geliştirilen veri madenciliği modeli, kökleri yukarıda, ters çevrilmiş bir ağaca benzetilebilir. Ağaç karar verme noktalarını temsil eden düğümler ve bu düğümleri birbirine bağlayan dallardan oluşur.
- En üstte yer alan düğüm kök düğüm olarak adlandırılır. Kök düğümde bazı özellikler test edilerek bu testin farklı sonuçlarına göre kök düğümden farklı yönlere dallar oluşturulur. Her bir dal yeni bir karar düğümüne bağlanır ve burada yeni birtakım özellikler test edilerek bu düğümlerden de yeni dallar türetilir. Ağaç yapısının en altında ise artık kendisinden yeni bir dal türemeyecek ve bu nedenle yaprak olarak adlandırılan düğümler bulunur. Buna göre veritabanındaki tüm kayıtlar bir ağaç yapısı biçiminde düzenlenerek ağaçta yer aldıkları dala göre sınıflandırılmış olur.
- 2. Yapay sinir ağları: Yapay sinir ağları karmaşık hesaplamaları gerçekleştiren biyolojik sinir sistemlerini model alır. Bu anlamda biyolojik sinir sistemlerinin simülasyonudur. Biyolojik sinir sistemlerinde öğrenme, sinir hücreleri arasındaki etkileşim ile gerçekleşir. Biyolojik sinir sistemlerinin öğrenme özelliği, tanımlanan görevden bağımsız olarak esnek yapıda, karmaşık verilerin işlenmesinde hesaplamaya dayalı modellerin oluşturulmasına esin kaynağı olmuştur. Yapay sinir ağları, özellikle bağımlı ve bağımsız değişkenler arasındaki karmaşık ve doğrusal olmayan ilişkileri modelleyebilmesi açısından tercih edilir. Bununla birlikte, bu yöntemle oluşturulan modellerin yorumlanması diğerlerine göre daha zordur.

Tahmin Edici Modeller

- 3. Genetik algoritmalar: Genetik algoritmalar karmaşık eniyileme problemlerinin çözümünde kullanılan bir teknolojidir. Dolayısıyla aslında doğrudan bir veri madenciliği modeli değildir. Bununla birlikte veri madenciliğinde de kullanılabilen bir eniyileme yöntemidir.
- Genetik algoritmalar da yapay sinir ağları gibi biyolojik mekanizmalardan esinlenerek geliştirilmiş algoritmalardır. Genetik algoritmalar doğada gözlenen evrim sürecine benzer bir yapıda ele alınan problemi, sanal olarak evrimden geçirerek çözmektedir.
- Problemin çözümü için öncelikle, nüfus olarak tanımlanan ve kromozomlar tarafından temsil edilen bir dizi sonuç (bir çözüm kümesi) belirlenir. Bir nüfustan alınan sonuçlar, bir öncekinden daha iyi olması beklenen yeni bir nüfusu oluşturmak için kullanılır. Yeni nüfusların seçiminde her yeni bireyin problem için çözüm olup olmadığına uygunluk fonksiyonları kullanılarak karar verilir.
- Burada sözü edilen kromozomlar veritabanındaki her bir kayıttır ve bu kromozomlar üretilecek yeni sonuçlar hakkında birtakım bilgiler içerirler. Dolayısıyla bu bilgilerin kullanılabilmesi için kromozomların kullanılabilir biçimlere dönüştürülmesi gerekir; bu işleme kromozomların çözümlenmesi denir.
- 4. Zaman serisi analizi: Zaman serisi analizi, zaman değişkeni ile ilişkilendirilmiş verilerin tahmin edilmesi problemlerinde kullanılır. Zaman serisi analizlerinin kullanıldığı en yaygın alan borsa işlemleridir. Bir hisse senedinin veya borsa endeksinin gelecek değeri tahmini zaman serisi problemlerine örnek oluşturur.
- Zaman serisi problemlerinin çözümünde istatistiksel ve istatistiksel olmayan birçok veri madenciliği algoritması kullanılmaktadır. Tahmin modellerinin oluşturulmasında geçmiş verilerden yararlanılması nedeniyle bu modeller denetimli öğrenme modellerindendir.

Tahmin Edici Modeller

- 5. k-en yakın komşu: k-en yakın komşu algoritması sıklıkla kullanılan bir algoritmadır. Temel olarak algoritma sınıfları belli olan bir örnek kümesindeki gözlem değerlerini inceler. Daha sonra elde edilen bu bilgi sisteme eklenen verinin ait olduğu sınıfın tespitinde kullanılır. Sınıflandırma yapılırken veritabanındaki her bir kaydın diğer kayıtlarla olan uzaklığı hesaplanır. Ancak, bir kayıt için diğer kayıtlardan sadece k adedi göz önüne alınır. Algoritmanın isminden de anlaşılacağı gibi bu k adet kayıt, başka bir ifadeyle veritabanındaki nokta, mesafesi hesaplanan noktaya diğer kayıtlara nazaran en yakın olan kayıtlardır. Bu yöntem coğrafi bilgi sistemlerinde çok kullanılır, belirlenen bir noktaya en yakın şehir, istasyon vb. belirlenmesi aslında k-en yakın komşu algoritmasının temelini oluşturur. Algoritmada k değeri önceden seçilir; değerinin yüksek olması birbirlerine benzemeyen noktaların bir araya toplanmasına, çok küçük seçilmesiyse birbirine benzediği, yani aynı sınıfın noktaları oldukları hâlde, bazı noktaların ayrı sınıflara konmasına ya da o tür noktalar için ayrı sınıfların açılmasına neden olur (Silahtaroğlu, s. 118). Gözlem değerlerinin arasındaki uzaklıkların hesaplanmasında “Öklid” uzaklık formülü kullanılır.
- 6. Bayes sınıflandırması: Bayes sınıflandırma yöntemi, elde var olan, mevcut sınıflanmış verileri kullanarak yeni bir verinin mevcut sınıflardan herhangi birine girme olasılığını hesaplayan yöntemdir. İstatistikte kullanılan Bayes kuralına dayalı olarak geliştirilmiş algoritma ve sınıflandırma teknikleri bu isimle anılır (Silahtaroğlu, s. 97).

Tanımlayıcı Modeller

- Tanımlayıcı modeller verilerdeki örüntü veya ilişkileri tanımlar. Bu modeller tahmin edici modellerin aksine analiz edilen verilerin özelliklerini incelemek için kullanılan modellerdir.
- Tahmini modellerde kullanılan yazılımlar kendilerine verilen veritabanını bir bütün olarak düşünür ve öğrenme işlemini de bu bütünü temel alarak gerçekleştirir. Buna karşın tanımlayıcı modellerde, veritabanındaki kayıtlar arasında bir bağlantı, ilişki kurulmaya çalışılır. Böylece bir veritabanındaki kayıtlar arasında çok rastlanan kurallar ortaya çıkarılır. Sepet analizi olarak adlandırılan ve İnternet üzerinden alışveriş yapılan sitelerde, alışveriş sepetindeki ürünler arasındaki ilişkiyi ortaya çıkarıp, müşterinin herhangi bir ürünü seçmesinin ardından müşteriye ilgisini çekecek bir başka ürünün önerilmesi, tanımlayıcı modeller kullanılarak yapılan veri madenciliği örneğidir.
- Bir diğer örnek olarak sigorta poliçesini yenilememiş müşterilerin benzer özelliklerini belirleyecek bir kümeleme çalışması verilebilir. En yaygın kullanılan tanımlayıcı modeller; kümeleme, birliktelik kuralları, sıra örüntü analizi ve özetleme biçiminde sıralanabilir.
- 1. Kümeleme: Kümeleme, verileri birbirlerine olan benzerliklerine göre anlamlı ve/ veya kullanışlı gruplara ayırmaktır. Eğer amaç anlamlı kümeler oluşturmaksa o zaman kümeler verilerin doğal yapısını yansıtmalıdır.
- Bazı durumlarda ise kümeleme veri özetleme gibi farklı amaçlar için kullanışlı bir başlangıç noktası oluşturmaktadır. Kümeleme analizi bir hedef değişken içermediğinden, diğer bir ifade ile veriler bağımlı bir değişkene göre değil öznitelik değerlerine göre gruplandırıldığından, daha önce sözü edilen sınıflama analizinden farklı bir yaklaşımdır. Kümeleme analizinde, hedef değişkenin değerini belirlemeye yönelik sınıflama, tahmin etme veya kestirim yapılmaya çalışılmaz. Bunun yerine verinin tamamını bölümlere ayırmak için homojen alt gruplar veya kümeler araştırılır. Bu işlem gerçekleştirilirken kümeler içindeki verilerin benzerliği göz önüne alınır.

Tanımlayıcı Modeller

- Oluşturulan kümeler önceden tanımlanmadığından ve verinin özelliklerine göre belirlendiğinden kümelerin anlamı konuyla ilgili bir alan uzmanı tarafından yorumlanmalıdır.
- Verilerin kümeleme analizine göre modellenmesinde matematik, istatistik, makine öğrenimi ve yapay zekâ gibi birçok alandan yararlanılır.
- Kümeleme sürecinde bağımlı ve bağımsız değişkenler arasında bir bağ kurmak söz konusu olmadığından, kümeleme yaklaşımı makine öğreniminde denetimsiz öğrenme başlığı altında yer alır.
- Diğer bir ifadeyle kümelemedeki öğrenmenin denetimsiz öğrenme olmasının nedeni önceden belirlenmiş sınıfların olmayışıdır. Önceden belirli sınıflar olsaydı, bu durumda kullanılan model zaten bir sınıflandırma modeli olacaktı. Önceden sınıflar belirli iken, yani kadın ve erkek diye iki ayrı sınıf varken yapılan (algoritmik) öğrenmeye denetimli öğrenme; herhangi bir sınıf ismi verilmeden yapılan öğrenmeyeyse denetimsiz öğrenme denilir.
- Örneğin, veritabanındaki kayıtlarda her kaydın yanına kadın veya erkek bilgisi yazılıyor olsun, bu durumda veritabanı üzerinde yapılan herhangi bir (kadın veya erkek olduğuna dair) kural çıkarma işlemi denetimli öğrenmedir.
- Ancak aynı veritabanında, kayıtların yanında kadın mı erkek mi olduğu bilgisi yok iken yapılan kural çıkarma işlemi denetimsiz öğrenmedir. Bu işlem aynı zamanda veritabanını (iki) kümeye ayırma, yani kümeleme işlemidir. Burada kadın/erkek gibi bir etiket ya da sınıf olmayacağı için kümeleme kayıtlar arasındaki benzerlik veya mesafe ölçütüne göre yapılır. İki verinin benzerliğinden kasıt ise aralarındaki mesafenin ölçülmesi ve değerlendirilmesidir.
- Bu değerlendirme, veritabanındaki diğer verilere kıyasla iki verinin ne kadar yakın ya da benzer oldukları açısından yapılabileceği gibi önceden belirlenmiş kısıtlar eşik değerleri çerçevesinde de yapılabilir (Silahtaroğlu, s. 59-60).

Tanımlayıcı Modeller

- Birliktelik kuralları: Birliktelik kuralları veriler arasındaki güçlü birliktelik özelliklerini tanımlayan örüntüleri keşfetmek için kullanılan analiz yöntemidir. Birliktelik kuralı, belirli türdeki veri ilişkilerini tanımladığı için tanımlayıcı modeller içinde yer almaktadır. Herhangi bir ürün alındığında bir başka ürünün de satın alınması bir birliktelik kuralı verir. İş dünyasında birliktelik analizi, pazar sepeti veya benzeşme analizi olarak da adlandırılır ve müşterilerin satın alma alışkanlıklarını analiz ederek, ilgili ürünler arasındaki potansiyel çapraz satış olanaklarını tanımlamak için kullanılır. Örneğin; "Bira satın alan müşteriler %80 olasılıkla cips de satın alırlar" ya da "Düşük yağlı peynir satın alan müşteriler %90 olasılıkla yağsız yoğurt da satın alırlar" biçimindeki sonuçlara birliktelik kuralları analizi ile ulaşılabilir. Raf düzenlemeleri bu sonuçlar temel alınarak yapıldığında satış oranları arttırılabilir.
- 3. Sıra örüntü analizi: Sıra örüntü analizi birliktelik kurallarına benzer bir yapıda olup aynı zamanda olayların zaman sıralarıyla ilgilenir. Birliktelik kurallarında sözü edilen pazar sepeti analizinde, ürünlerin müşteri tarafından aynı anda alınmasıyla ilgilenilirken sıra örüntüleri analizinde belirli bir zaman aralığında satın alınan ürünler arasındaki ilişkilerle ilgilenilir. "A ameliyatı olan bir hastada, 10 gün içinde %40 olasılıkla B enfeksiyonu oluşacaktır", "Menkul Kıymetler Borsası endeksi düşerken A hisse senedinin değeri %20'den daha fazla artacak olursa, üç iş günü içinde B hisse senedinin değeri %60 olasılıkla artacaktır" ya da "Çekiç satın alan bir müşteri, ilk üç ay içerisinde %15, bu dönemi izleyen üç ay içerisinde %10 olasılıkla çivi satın alacaktır" biçiminde sıralanabilecek ilişki tanımlamaları, sıra örüntü analizi ile tanımlanabilecek ilişkilere örneklerdir.
- 4. Özetleme: Karakterizasyon veya genelleştirme olarak da adlandırılan özetleme, verileri basit tanımları yapılmış alt gruplar içine yerleştirme işlemidir. Özetleme veritabanı hakkında betimleyici bilgileri ortaya çıkarır ve verilerden elde edilen ortalama veya standart sapma gibi tüm veriyi temsil eden göstergelerin hesaplanmasını ifade eder. Özet bilgiler, veritabanı fonksiyonları ve tanımlayıcı veri madenciliği teknikleri kullanılarak elde edilebilir.

VERİ MADENCİLİĞİNİN DİĞER VERİ ANALİZİ YAKLAŞIMLARI İLE KARŞILAŞTIRILMASI

- Veri madenciliği ile veri analizi amacıyla kullanılan diğer yaklaşımlar farklı açılardan karşılaştırılabilir. Buna göre veri madenciliği ile geleneksel istatistiksel analiz, veri sorgusu, SQL (Yapılandırılmış Sorgu Dili), OLAP (Çevrimiçi Analitik İşleme) gibi diğer yaklaşımlar karşılaştırıldığında izleyen kesiminde verilen farklılıklar olduğu görülmektedir. Geleneksel istatistiksel analiz ile veri madenciliği arasındaki temel farklar aşağıdaki gibi sıralanabilir:

• İstatistiksel analizde analize genellikle bir hipotez kurularak başlanırken veri madenciliği ile analizde herhangi bir hipoteze gerek duyulmaz.

• İstatistikçiler hipotezlerini eşleştirmek için kendi eşitliklerini geliştirmek zorunda oldukları hâlde, veri madenciliği algoritmaları eşitlikleri otomatik olarak geliştirir.

• İstatistiksel analizler genellikle sayısal veriler üzerinde gerçekleştirilirken veri madenciliği sayısal verilere ek olarak metin, ses vb. gibi farklı veri türleri üzerinde de işlem yapabilir.

• İstatistiksel analizde kirli veri analiz sırasında bulunur ve filtre edilirken veri madenciliği temizlenmiş veri üzerinde gerçekleştirilir

VERİ MADENCİLİĞİNİN DİĞER VERİ ANALİZİ YAKLAŞIMLARI İLE KARŞILAŞTIRILMASI

- İstatistiksel analizde bulunan sonuçlar kolaylıkla yorumlanabilirken veri madenciliğinin sonuçlarını değerlendirmek ve yorumlamak aynı derecede kolay olmayıp uzman istatistikçilere gereksinim duyulur. Veri sorgusu, OLAP ve veri madenciliği, kullanım amacına göre karşılaştırıldığında;
- Veri sorgusu aranan ulaşılmak istenen bilginin ne olduğu bilindiği durumda ve büyük veri tabanı ile çalışılmak istendiği durumlarda,
- OLAP büyük veritabanlarında veriler arasındaki basit ilişkilerin keşfedilmek istendiği durumlarda,
- Veri madenciliği veriler arasında var olan fakat açıkça gözlenemeyen örànta ve ilişkilerin keşfedilmesi istendiği durumlarda kullanılır. SQL, OLAP ve veri madenciliği, keşfedilmek istenen bilgi tipine göre karşılaştırıldığında ise,
- Seçilen kayıtlara ait ortalama ve toplam değer gibi özet bilgiler sığ bilgi olarak tanımlanır. Bu tür bilgilere ulaşmak için SQL kullanımı yeterlidir.
- Farklı özelliklerin ortaya çıkma sıklığı hakkındaki bilgi çok boyutlu bilgi olarak nitelendirilir. Bu tür bilgiye ulaşma işlemini veri küpü üzerinden OLAP yapabilir.
- Önceden tahmin edilemeyen örüntü ve ilişkiler gizli bilgi olarak ifade edilebilir ve bu örüntü ve ilişkiler veri madenciliği için başlangıç olabilir.
- Sadece önsel teknik veya metabilginin kullanımıyla keşfedilebilecek gizli örüntüler ve ilişkiler hakkında bilgi ise derin bilgi olarak tanımlanabilir ve bunlar da veri madenciliğinin araştırma sınırları içinde yer alır (Koyuncugil ve Özgülbaş, s. 25).

VERİ MADENCİLİĞİNİN UYGULANDIĞI ALANLAR

- Kâr amacı güden ya da gütmeyen tüm kuruluşlarda, kurumun yaşamını sürdürebilmesi öncelikli amaçlardan biridir. Bu amacın başarılabilmesi ise her gün değişen ve yenilenen koşullara uyum sağlayabilme becerisi ile sağlanabilecektir. Bu nedenle yalnızca deneyim ve önsezilere dayanarak kararlar vermek beraberinde yüksek riski de getirecektir. Bu riski azaltmanın yolu ise doğru karar destek sistemlerinden yararlanmaktır.
- Doğru karar destek sistemlerinin oluşturulması söz konusu olduğunda ise veri madenciliği teknikleri çok önemli araçlar olarak karşımıza çıkmaktadır. Bu araçların kullanımıyla, kurumların ve işletmelerin etkin kararlar almak için ihtiyaç duydukları bilgilere erişmeleri de mümkün olacaktır.
- Veri madenciliğinin uygulandığı alanlar kesin çizgilerle sınırlandırılamaz. Büyük miktarda verinin üretildiği ve kaydedildiği ve karar verme sürecine ihtiyaç duyulan tüm alanlarda veri madenciliği uygulamaları yapmak mümkündür.
- Veri madenciliğinin yoğun ve başarılı bir biçimde kullanıldığı başlıca alanlar; pazarlama, finans (bankacılık, sigortacılık, borsa), parekendecilik, sağlık, telekomünikasyon, endüstri ve mühendislik, eğitim, tıp, biyoloji, genetik, kamu, istihbarat ve güvenlik biçiminde sıralanabilir.

Pazarlama Alanındaki Uygulamalar

- Veri madenciliğinin en çok kullanıldığı alanların başında pazarlama alanının geldiği söylenebilir. Yapılan çalışmalar incelendiğinde, pazarlama alanında yapılan veri madenciliği uygulama konuları izleyen biçimde sıralanabilir.
- Müşterilerin satın alma örüntülerinin belirlenmesi
- Benzer özellikler gösteren müşterilerin bulunması
- Müşterilerin demografik özellikleri arasındaki bağlantıların belirlenmesi
- Benzer gelir grupları ilgi alanları harcama alışkanlıklarının ortaya çıkarılması

Pazarlama Alanındaki Uygulamalar

- Benzer müşterileri otomatik olarak gruplayarak pazar dilimlerinin tanımlanması ve bu bilginin pazarlama kampanyalarında kullanılması
- Mevcut müşterilerin elde tutulması yeni müşterilerin kazanılması
- Satış tahmini yapılması
- Müşteri ilişkileri yönetimi
- İnternet üzerinden satış yapan işletmeler için kullanıcı profillerinin belirlenmesi
- Web sayfalarının kullanıcı bilgilerine göre kişiselleştirilmesi

Finans Alanındaki Uygulamalar

- Veri madenciliğinin sıklıkla kullanıldığı bir diğer alan bankacılık, sigortacılık ve borsa olarak sıralayabileceğimiz finans sektörüdür. Finans sektöründeki veri madenciliği uygulama konuları da izleyen biçimde sıralanabilir.
- Farklı finansal göstergeler arasındaki gizli korelasyonların bulunması
- Müşteri kaybı analizi
- Kredi kartı dolandırıcılıklarının belirlenmesi
- Müşteriler arasındaki benzerliklerin belirlenmesi
- Kredi kartı harcamalarına göre müşteri gruplarının belirlenmesi
- Kredi kartı ve kredi taleplerinin değerlendirilmesi
- Risk analizi ve risk yönetimi
- Hisse senedi fiyatlarının tahmin edilmesi
- Yatırımların modellenmesi
- Sigorta dolandırıcılıklarının belirlenmesi
- Sigorta yaptıran müşteriler içinde riskli müşteri grubunun belirlenmesi

Sağlık Alanındaki Uygulamalar

Veri madenciliğinin sağlık alanındaki uygulamaları;

- Yeni ilaçların geliştirilmesi
- Piyasada var olan ilaçların etkilerinin belirlenmesi
- Hastalara uygulanan test sonuçlarının tahmin edilmesi
- Hastalıkların önceden teşhis ve tedavi edilmesi biçiminde sıralanan konularda yapılmış olup önemli etkileri olan uygulama alanlarından biridir

Endüstri ve Mühendislik Alanındaki Uygulamalar

Endüstri ve mühendislik alanında veri madenciliğinden;

- Kurum kaynaklarının optimal kullanımı
- Üretim süreçlerinin kontrol edilmesi
- Kalite kontrol analizlerinin gerçekleştirilmesi
- Sistem performanslarına etki eden faktörlerin ve kuralların belirlenmesi konularında yararlanılmaktadır.

Eğitim Alanındaki Uygulamalar

Eğitim alanında yapılan veri madenciliği uygulama konuları ise izleyen biçimde sıralanabilir.

- Öğrenci verilerinin analiz edilmesi
- Öğrenci başarı ve başarısızlık nedenlerinin tespit edilmesi
- Öğrenci başarılarının arttırılması
- Eğitim-öğretim ortamlarındaki aksaklıkların tespit edilmesi
- Daha etkili eğitim öğretim ortamlarının oluşturulması