SAKARYA ÜNİVERSİTESİ

Veri Madenciliği Uygulamaları

Hafta 1

Prof.Dr. Nilüfer YURTAY

1. Genel Tanımlar

Hızla teknolojinin hayatımıza girmesiyle elektronik ortamlarda tutulan verilerin artması, gelişmelere de bağlı olarak ,bazı soruların cevaplarını aramamıza neden olmuştur.

Verilerin nasıl ve nerelerde kullanılacağı, nasıl yorumlanacağı ve bilgiye nasıl ulaşılacağı ihtiyacı ortaya çıkarmıştır.

Gelişen teknoloji ile birlikte kullanıcı talepleri de artarak çeşitlenmiştir. Bu itibarla basit raporlamalar ve standart sorguların yeterli gelmemesi de sektör çalışanlarını arayışa itmiştir."

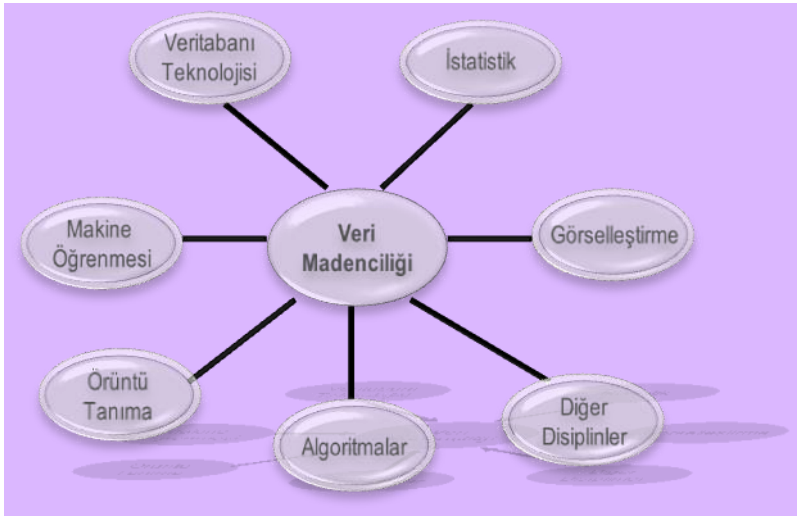
1.1 Veri Madenciliği Nedir

Verilerden üstü kapalı, çok net olmayan , önceden bilinmeyen ancak potansiyel olarak kullanılabilir bilgi ve örüntülerin çıkarılması olarak tanımlanabilir.

Veri madenciliği ;

- veri tabanı teknolojisi,
- istatistik,
- yapay zeka (*artificial intelligence*),
- makine öğrenimi (*machine learning*),
- örüntü tanıma (*pattem recognition*),
- veri görselleştirmesi (*data visualization*)

gibi pek çok teknik alan arasında köprü görevi gören çok disiplinli bir alandır(Şekil 1.1).



Şekil 1.1 Veri madenciliği ve diğer disiplinler arasındaki bağlantı

Veri madenciliği;

- Astronomi,
- Biyoloji,
- Tıp,
- Finans,

- Sigorta,
- Pazarlama gibi bir çok dalda da uygulanabilmektedir.

Dersimizin amacı, bilisim teknolojileri dünyasındaki önemini her geçen gün daha da arttıran veri madenciliği konusunu ve veri madenciliği modellerini öğrenmektir.

"İşletmelerde, operasyonel sistemlerde gerçekleşen işlemler sonucu çeşitli veriler üretilmektedir. Geleneksel ticarette geleneksel işlemler sonucu ortaya çıkan verilere, çağımızda, elektronik işlemler sonucu çıkan veriler de eklenmiştir.

İnternet, çağrı merkezleri, banka kartı, kredi kartı, üyelik kartı gibi manyetik kartlarla, binlerce müşterinin günün her saniyesinde yapabildiği onlarca işleme, alışverişe ait "terrabit" miktarlarında veri elde edilmektedir."

İşletmelerdeki bilgi sistemlerinin, isim ve kullandıkları yöntemler açısından geçirdiği aşamaları şu şekilde sıralamak mümkündür:

"**Yönetim bilişim sistemleri"nin** (Management Information System -MIS) stratejik karar verme sürecinde kullanılmaya başlamasıyla "**Karar destek sistemleri**" (Decision Support System -DSS) ortaya çıkmıştır. Karar destek sistemlen"nin istenilen bilgiyi doğru ve zamanında üretebilmesi için bu sistemlerin farklı bir yapıda kaydedilen veri ile beslenmesi gereği doğmuştur.

Ortaya çıkan verinin ve bilgi ihtiyacının farklılaşması ve artmasıyla işletmelerde halihazırda tutulan verinin operasyonel bilgi sistemlerinden ayrılma ihtiyacı doğunca "**Veri ambarları**" gündeme gelmiştir

1.2 Veri, Bilgi ve Çıkarımsal Bilgi

E-işletmelerde işlemlerden, operasyonel sistemlerden elde edilen verilerin depolandıkları ortamlar ***"veri ambarları"*** olarak adlandırılmaktadır.

Bu veriler daha sonra veri madenciliği teknikleriyle anlamlı bilgilere dönüştürülmekte ve stratejik karar verme sürecinde kullanılmaktadır.

"Veriler, insanlar, nesneler, işlemler, uygulamalar, olaylarla ilgili gerçekleri yansıtan niceliksel veya niteliksel değerlerdir. "

Örneğin, müşteriyle ilgili demografik veriler yaşı, geliri, eğitimi seviyesi, mesleği, veya hane halkında geliri olan ikinci bir kişinin olup olmadığı ile ilgili veriler olabilir.

Bilgi "information", işlenmiş verilerdir. Verilerin toplandıktan sonra sınıflandırılması, ortalama, mod, standart sapma ve benzeri istatistiksel ölçümlerle özetlenmesi, grafiksel olarak sunulması, istatistiksel ve matematiksel yöntemlerle analiz edilerek anlamlandırılması, çeşitli değişkenlerin birbirleriyle ilişkisi olup olmadığının tespit edilmesidir.

Çıkarımsal bilgi "knowledge" ise verilerden elde edilen anlamlı bilgilerden artık ileriye yönelik tahminler yapmak ve bunları eyleme dönüştürmekte kullanmak amacıyla üretilir.

Müşterilerin gelecekteki tutum ve davranışlarına ilişkin bilgi sahibi olmak, hangi müşteri kitlesine hangi kampanyanın uygun olacağını tespit edebilmek ancak çıkarımsal bilgi sahibi olunarak gerçekleştirilebilir.

1.3 Veri Kalitesi

Bir kurumun kaliteli veriye sahip olması, bir defada başlanıp bitirilecek bir projeden çok, sürekliliği olan bir disiplin karakterini sergilemektedir.

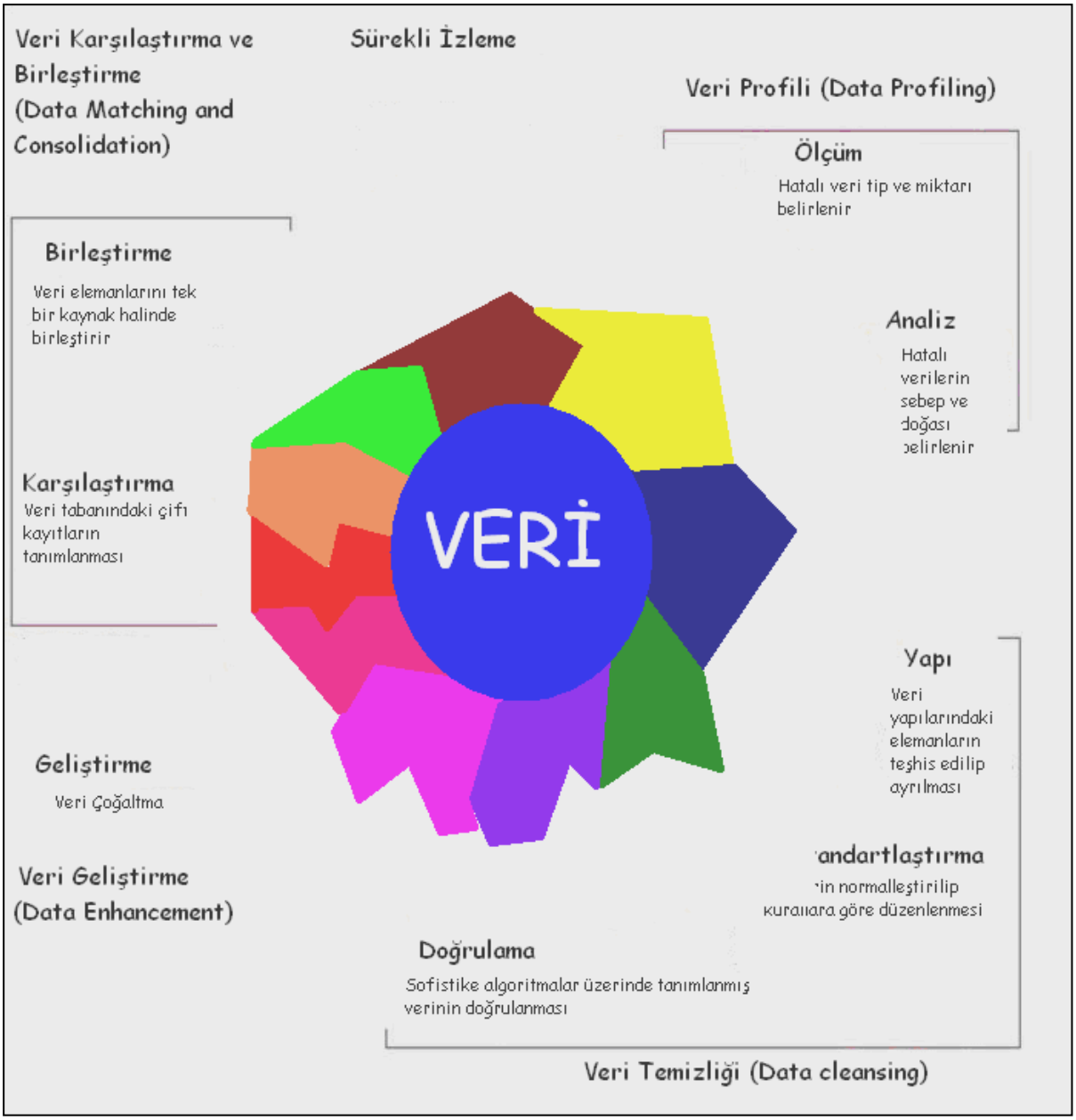
Veri kalitesi düşük bir kurumun, büyük yatırımlar yaparak hayata geçirmeye çalıştığı iş zekası projelerinin verimliliğinden bahsetmek mümkün değildir.

Veri kalitesi çözümleri , temel olarak kurum içinde biriken verilerin doğru, tam ve tutarlı hale getirilmesi için gerekli yazılım araçları, metodoloji ve süreçleri kapsamaktadır.

Günümüzde her kurum, verilerinin az ya da çok kirli olduğunun farkında. Standart olarak yapılmamış veri girişleri, bazı alanlara bilgi girilmemiş olması, birden fazla tabloyu ilgilendiren bilgi girişlerinde tablolar arasındaki tutarsızlıkların oluşması, ya da çift kayıtlar/müşteri tekilleştirme projeleri, veri kalitesi projelerine duyulan ihtiyacın belirginleştiği durumlardır. Veri kalitesi projelerinde beklenen sonuçlardan birisi, veri kirliliğini mümkün olduğunca azaltabilmek.

Veri kalitesinin arttırılmasındaki başlangıç noktası Data Profiling aşamasından geçiyor. Data Profiling, verinin incelenerek, hataların bulunması, bu hatalara bağlı nedenlerin ve bunların etkilerinin araştırılması, hataların ölçümlendirilmesi ve rapor ve kokpit ortamlarında sunulması olarak nitelendirilebilir. Data Profiling ile, veritabanındaki alanların sayısal ve yüzde olarak ne kadar dolu olduğunu, ilgili alan içindeki farklı değerlerin kırılımını ya da bir tabloda olan fakat diğer tabloda olmayan verilerin yüzde olarak ne kadar yüksek olduğunu takip etmek mümkün. Bu sırada son kullanıcıların bir takım kurallar verebilmesi gerekmekte. Örneğin “Bir veritabanında Email1 ve Email2 gibi 2 farklı alan olduğu halde bunlardan sadece 1 tanesinde email adresi yazılıysa, diğer alan boş olsa bile benim için doğru bir veridir bu!” diyebilen bir kullanıcı bu kuralı kendisi tanımlayabilmelidir. Böyle bir kullanıcı proaktif yöntemlerle, veri kalitesini periyodik ve otomatize yöntemlerle takip edebilmeli ve gerektiğinde yine otomatik olarak uyarılar gönderebilmelidir. Örneğin böyle bir uyarı kurumun Data Quality Score’u olarak nitelendirilebilecek bir limitin altında kalması durumunda ilgili kişilere bir mail gönderilmesi şeklinde hayata geçirilebilir.Bu tür kuralları verecek olan kullanıcılar veri kalitesinin arttırma sürecinde temel olarak IT’den çok iş tarafındaki kullanıcılar olarak değerlendirilmelidir. Data Profiling yazılımları bir kurumda, veriden sorumlu, verinin sahibi olan iş tarafındaki birim ya da departmanlar tarafından kullanılabilecek nitelikte olmalı. Bu da bu tür yazılım araçlarının, IT’den çok son kullanıcıların kullanabileceği bir arayüze sahip olmasını gerektirmektedir.[1]

[1] http://images.google.com.tr/imgres?imgurl=http://www.dsstechnology.com/Portals/0/DSSCustom/DQWheel.jpg&imgrefurl=http://www.dsstechnology.com/%25C3%2587%25C3%25B6z%25C3%25BCmler/VeriKalitesi/tabid/74/Default.aspx&usg=__lxnLy4oCGHA4Yf842g47xXFEl2k=&h=719&w=723&sz=71&hl=tr&start=2&um=1&tbnid=WhxFpLozaOLMaM:&tbnh=139&tbnw=140&prev=/images%3Fq%3Dveri%2Bkalitesi%26hl%3Dtr%26sa%3DN%26um%3D1



Şekil 1.2 Veri Kalitesi

1.4 Veri Ambarı

"Veri ambarı özneye dayalı, bütünleşmiş, zaman dilimli ve yöneticinin karar verme işleminde yardımcı olacak biçimde toplanmış olan değişmeyen veriler topluluğudur. "(W. H. Inmon)

Bu açıklama veri ambarına ilişkin önemli özellikleri göstermektedir. Dört anahtar kelime**, özneye dayalı, bütünleşmiş, zaman dilimli** ve **değişmeyen kavramları** veri ambarlarını, ilişkisel veri tabanı sistemleri, hareket işleme (*transaction processing*) sistemleri ve dosya sistemleri gibi diğer veri ambar sistemlerinden ayırmaktadır.

Veri ambarlarında tutulan veriler, işletmelerin faaliyetlerinden elde edilmiş olan, ancak farklı bir yapıda ve farklı bir fiziksel ortamda tutulan, bilgi (information, knowledge) üretmeye yönelik verilerdir.

Bu verilerden bazılarının her departmanın kendi kullanım amacına hizmet edecek şekilde ayrılmasıyla "**data mart**" olarak isimlendirilen her departmana özel veri tabanları oluşmuştur.

"**Veri madenciliği**" ise bu verilerden çeşitli teknikler, algoritmalar ve sorgulamalarla anlamlı bilgiler keşfetmektir. Veriye dayalı stratejik karar destek sistemlerinin tümü sonuçta "**iş zekası (business intelligence) çözümleri**" olarak isimlendirilmektedir.

Veri ambarı, bir kurumda gerçekleşen tüm operasyonel işlemlerin en alt düzeydeki verilerine kadar inebilen, etkili analiz yapılabilmesi için özel olarak modellenen, tarihsel derinliği olan, operasyonel sistemlerden fiziksel olarak farklı ortamdaki yapı üzerinde gerçekleşen süreçlerin toplamıdır.
Gupta (1999) ***veri ambarını*** şu şekilde tanımlamaktadır:

Veri ambarı, yapılandırılmış ve genişlemeye müsait, çeşitli operasyonlardan elde edilen verilerin nasıl aktarılacağının önceden işletmenin yapısına uygun biçimde tasarlanıp tespit edildiği, anlamlı geçmişe sahip verilerin tutulduğu, verilerin ve sorgulamaların basit işletme terimleriyle ifade edildiği ve hızlı analiz yapabilme ihtiyacına cevap veren veri depolama, erişme, sorgulama sistemleridir.

Veri Ambarlarının Kullanım Amacı :

- Müşterilerin gizli kalmış satın alma eğilimlerini tespit etmek ,
- Satış analizi ve trendler üzerine odaklanmak,
- Finansal analiz yapabilmek,
- Karar destek sistemleri için Stratejik Analiz yapabilmek.

Mevcut bilgi sistemlerinde çeşitli raporların zaten eskiden beri üretilmekte olduğu bir gerçektir. Ama bu veriler, bilgiler ve raporlar *veri ambarı* uygulamalarından önce daha çok geçmişin özetlenmesi şeklindeydi.

"Veri ambarina duyulan ihtiyaç, bilgiye olan ihtiyacın farklılaşmasından da kaynaklanmaktadır."

Bilgi edinme kültürü daha önce raporlama anlamına gelirken şimdi keşfedilen bilgi anlam kazanmıştır.

Örneğin, işletme nasıl bir strateji izlerse müşterisini elinde tutar ve rakibe gitmesini önler, hangi müşteri hangi kampanyaya olumlu cevap verir, gibi bilgilerin keşfedilmesine ihtiyaç vardır. Bu süreç ise sorulan sorulardan da anlaşılacağı gibi **reaktif** değil **proaktif** bir süreçtir.

Müşteriyi kaybettikten sonra neden kaybettiğini anlamak **reaktif** bir süreçken, bu nedenleri keşfettikten sonra, hangi müşterilerin kaybedilme riski olduğu ve bunları kaybetmemek için neler yapılacağına ilişkin bilgi elde etmek **proaktif** bir süreçtir.

1.5 Operasyonel Veri

İşletmelerin faaliyetlerine ilişkin verilerdir. Bu veriler birden fazla uygulama sonucu, çeşitli kaynaklardan üretilen, oldukça dağınık yapıda olan verilerdir.

Örneğin, sipariş kabul, sevkiyat faaliyetlerine ilişkin ürün çeşidi, fiyatı ve miktarlarıyla ilgili verilerden, stokta kalan miktar, sevk edileceği yer, alan kişi, ödeme şekline kadar çok çeşitli veriler olabilir.

1.6 Enformasyonel Veri

Operasyonel verilerin kaynaklardan çıkartılarak bir ortamda tablolaştırılmış ve kullanımı kolay bir biçime getirilmiş halidir.

Aynı müşteriyle ilgili farklı departmanlarda oluşmuş tüm verilerin bir arada ve erişilebilir durumda olmasıdır.

Tek bir müşteri, firmaya farklı kanallardan erişmiş ve farklı işlemler gerçekleştirmiş olabilir. İnternetten, şubeden veya çağrı merkezinden erişen, yani farklı kanallardan, farklı zamanlarda erişen müşterinin aynı müşteri olduğunun anlaşılabilmesi, enformasyonel veri sayesinde mümkün olmaktadır.

Kritik olan, bu verilerin farklı ortamlardan ve farklı formatlardan gelmesine rağmen anlamlı biçimde bir araya getirilmesidir.
Bilgisayar sistemleri her geçen gün hem daha ucuzluyor, hem de güçleri artıyor. İşlemciler gittikçe hızlanıyor, disklerin kapasiteleri artıyor.

Artık bilgisayarlar daha büyük miktardaki veriyi saklayabiliyor ve daha kısa sürede veriyi işleyebiliyorlar. Bunun yanında bilgisayar ağlarındaki ilerleme ile bu veriye başka bilgisayarlardan da hızla ulaşabilmek olası.

Bilgisayarların ucuzlaması ile sayısal teknoloji daha yaygın olarak kullanılıyor. Veri doğrudan sayısal olarak toplanıyor ve saklanıyor. Bunun sonucu olarak da detaylı ve doğru bilgiye ulaşabiliyoruz.

Veri madenciliği, veri ambarlarında tutulan, ilk başta çok net anlaşılamayan, adeta veriler arasında gizli saklı kalmış bilgiyi ortaya çıkartmak, bilgiyi keşfetmektir.

Veri madenciliği yaklaşımları ve araçları, ***veri ambarlarındaki*** verilerde saklı eğilimleri, eğilimlerin birbirleriyle ilişkilerini, bu ilişkilerin nedenlerini ve verilerin nasıl bir seyir gösterdiğini ortaya çıkaran yöntemler topluluğu, yaklaşımlar, modellemeler ve tekniklerdir.

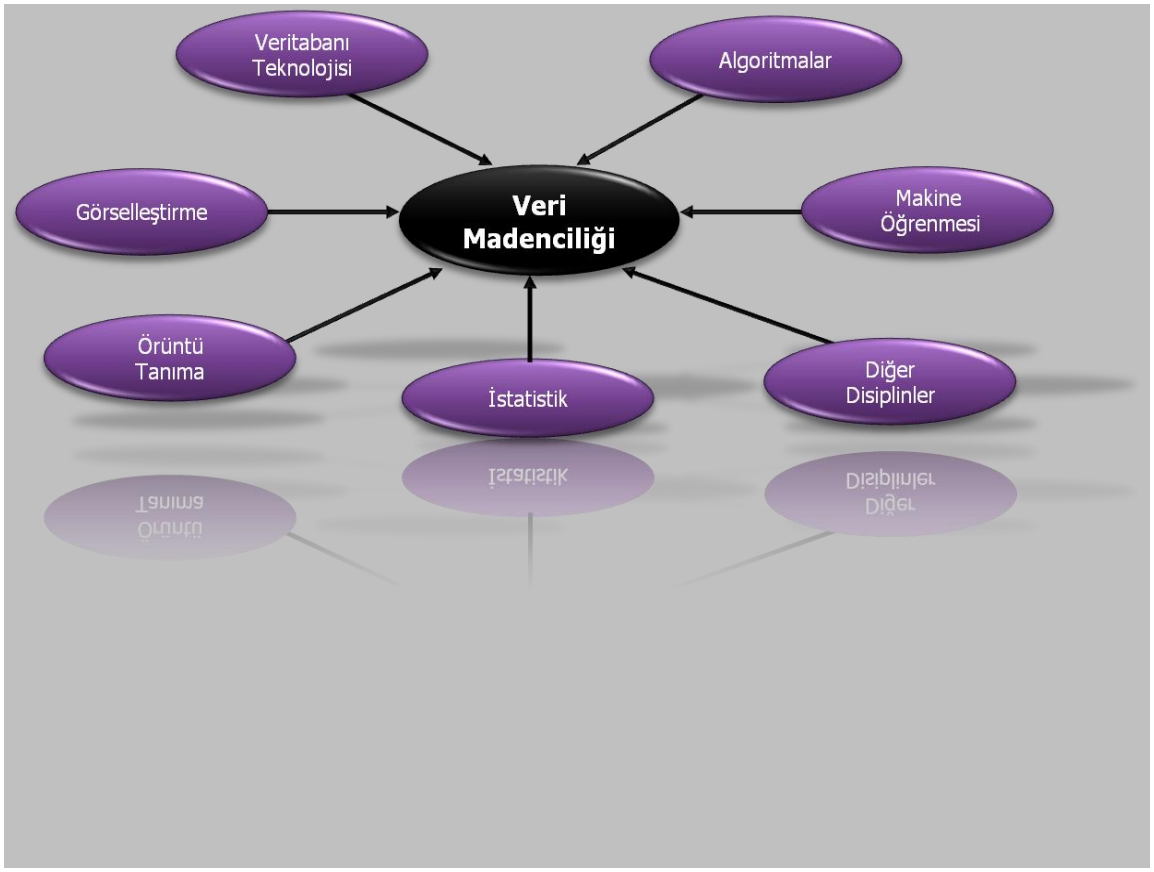
- Veri ambarlarında depolanan verilerde saklı olan bilgiyi ortaya çıkarma sürecidir.
- Verinin özelliklerinden eğilimleri anlama sürecidir.
- Çok büyük miktarlardaki veriden yeni ve anlamlı bilgiler üretmektir,
- Verinin, modellemelerle değerlendirilecek bilgiye dönüştürülmesidir.
- Veri madenciliğinin en önemli faydası bu bilginin eyleme yönelik olarak değerlendirilebilmesidir.

Veri Madenciliğinde Kullanılan Yöntemler :

- Sınıflandırma
- Kümeleme
- Görselleştirme
- İlişki kurma
- Tahmin modelleri

Veri Madenciliğinde Kullanılan Algoritmalar(Şekil 1.3) :

- Sinir Ağları (neural networks)
- Karar Ağaçları (decision trees)
- Genetik Algoritmalar
- İstatistiksel Analiz



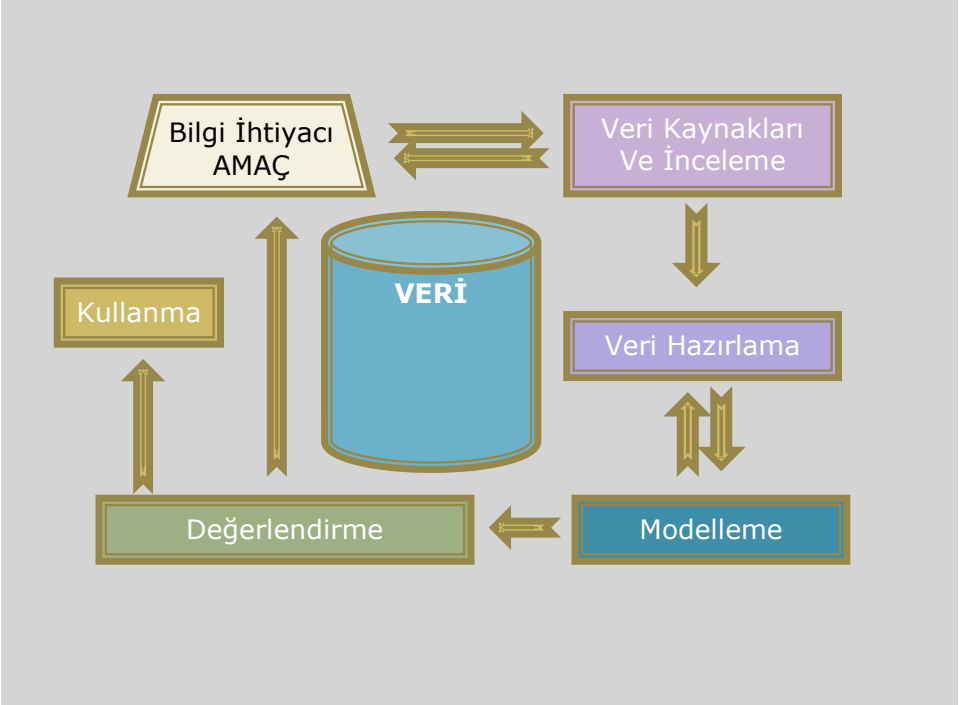
Şekil 1.3 Veri madenciliği ile iligili yöntem ve algoritmalar

1.7 Veri Madenciliği Süreci

Süreç aşağıdaki gibi özetlenebilir:

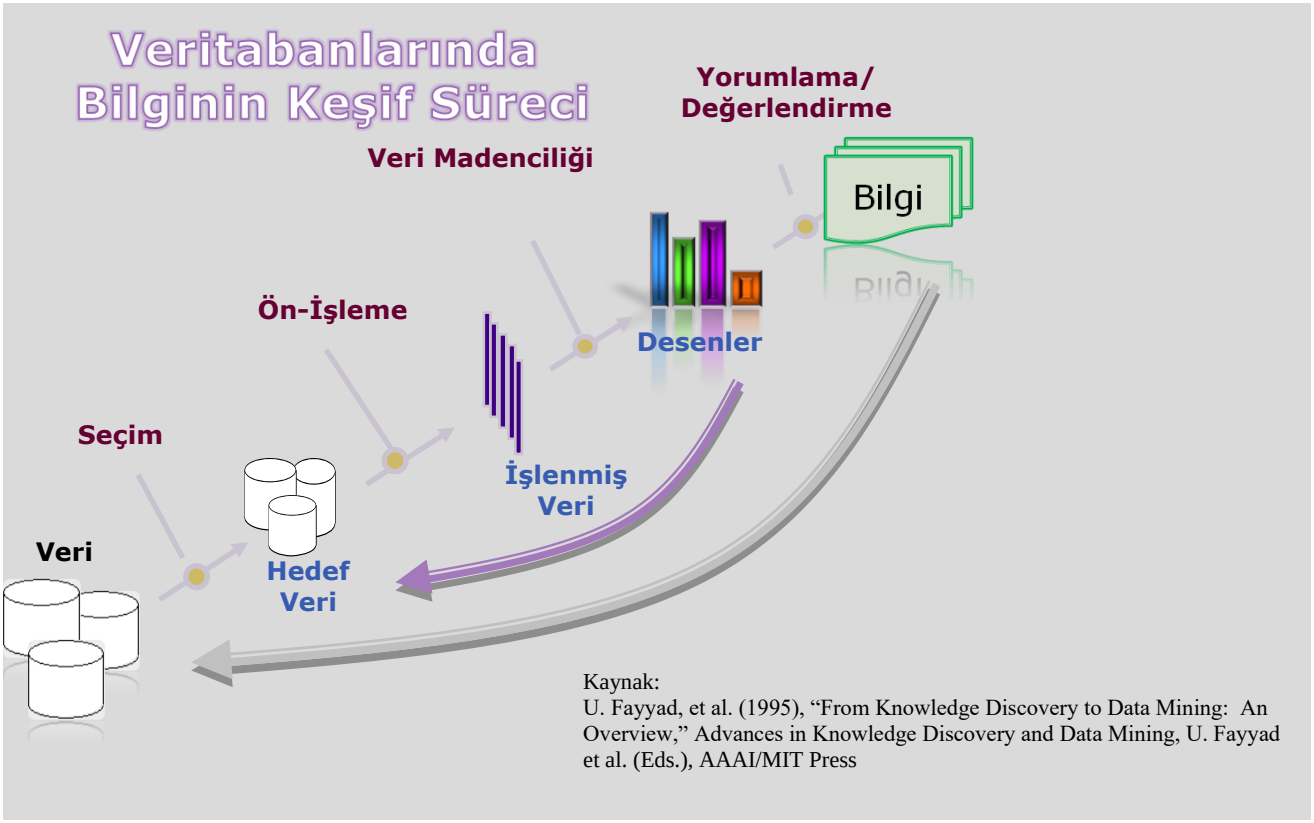
- Problemin Tanımı
- Kullanılacak verilerin seçilmesi ve hazırlanması
- Verilerin bulunması ve analizi
- Modelin oluşturulması
- Modelin geçerliliğinin testi
- Bilginin üretilmesi, eylem planına dönüştürülmesi ve sonuçların ölçülüp değerlendirilmesi

Veri standart süreci şekil 1.4 de özetlenmiştir.



Şekil 1.4 Veri standart SÜreci

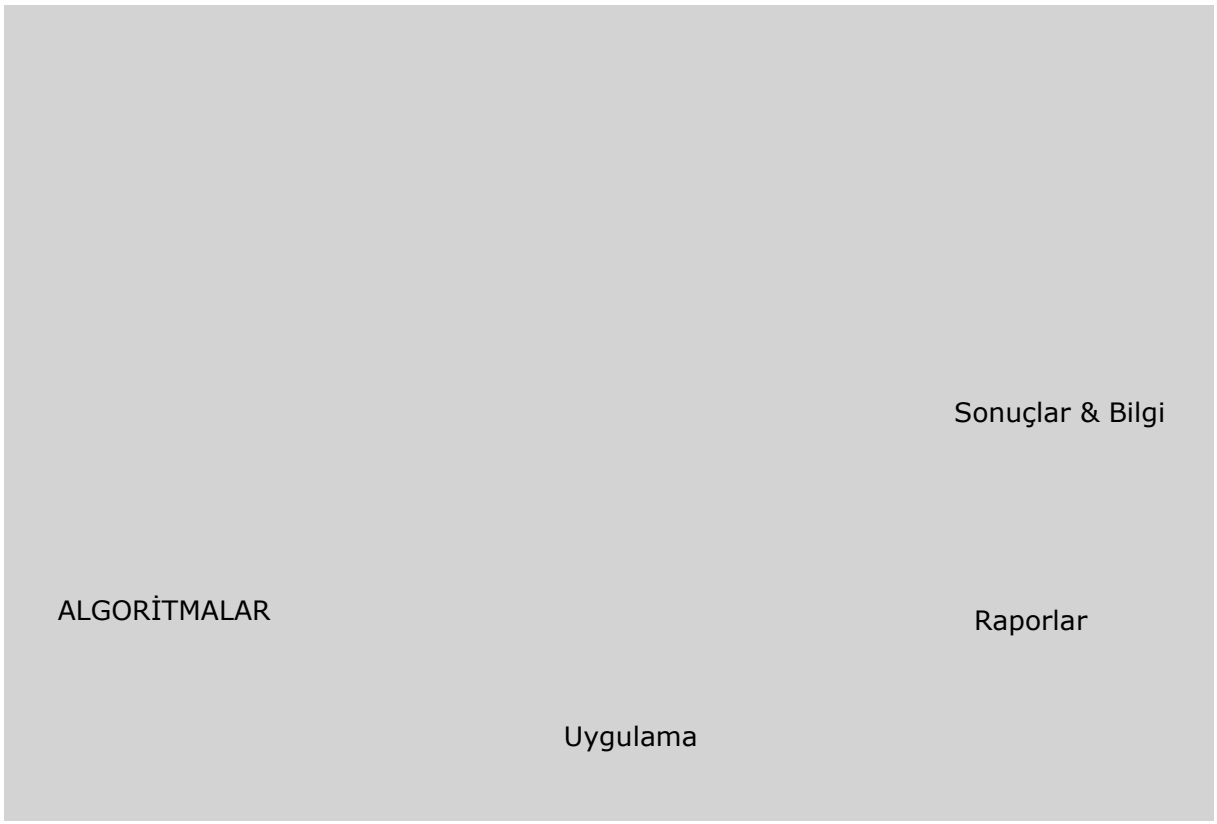
Veri tabanlarındaki bilginin keşif süreci de Şekil 1.5 de özetlenmiştir.



Şekil 1.5 Veritabanlarındaki bilginin keşif süreci

Veri madenciliği çok boyutlu bir görünüme sahiptir(şekil 1.6):

- **Kullanılan Veriler**
  - İlişkisel, veri ambarı, muamele verisi, nesneye yönelik –ilişkisel, seriler, zaman, uzaysal veri, metin, çoklu-ortam, heterojen veritabanları, WWW
- **Keşif Edilecek Bilgi**
  - Karakterizasyon, discriminasyon (ayırım), ilişki (bağlantı), sınıflandırma, gruplama, eğilim/sapma, aykırı değer (outlier), vs.
- **Kullanılan Teknikler**
  - Veritabanına yönelik, veri ambarı (OLAP), makina öğrenmesi, istatistik, görselleştirme
- **Uygulama Alanları**
  - Perakende, haberleşme, bankacılık, sahtekârlık analizi, biyolojik veri analizi, borsa analizler, Web madenciliği vb.



Şekil 1.6 Veri madenciliğinin çok boyutluluğu