VERİ BİLİMİ

VERİ MADENCİLİĞİ:

1. Veri Tabanlarında Bilgi Keşfi olarak adlandırılan paradigmaya denir

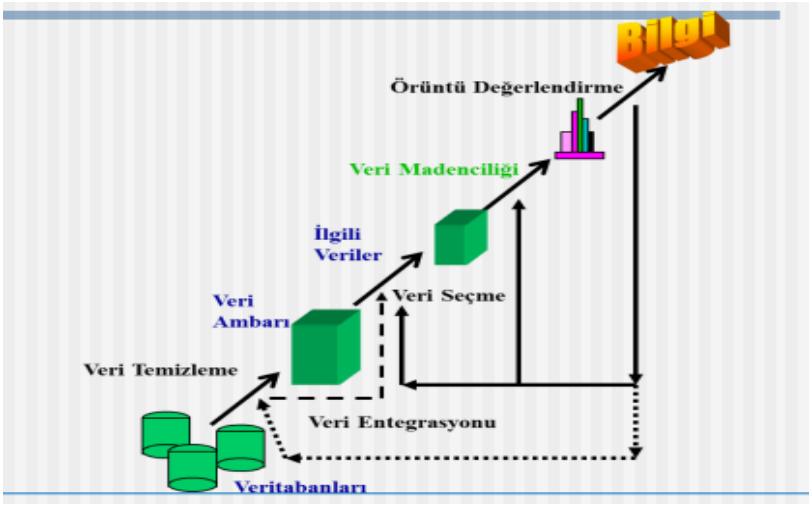
2. Büyük miktarda veri içinden, gelecekle ilgili tahmin yapmamızı sağlayacak bağıntı ve kuralların aranmasıdır

3. Bir kurumda üretilen tüm verilerin belirli yöntemler kullanarak var olan ya da gelecekte ortaya çıkabilecek gizli bilgiyi ortaya çıkarma süreci olarak değerlendirilmesidir.

Veri madenciliği ile ilgili bilgiler

- Gizli
-Önemli
-Önceden bilinmeyen
-Yararlı

BİLGİNİN KEŞFİ AŞAMALARI



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 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

VERİ MADENCİLİĞİNDE TEMEL KAVRAMLAR

Veri: kayıt altına alınmış her türlü olay, durum, fikirdir.Bunlar ham halde olup işlenmemiş olup anlamlandırılmamış kayıtlardır, Nesneler ve nesnelerin niteliğinden oluşan kümelerdir(Java class gibi)

Enformasyon: Verilerin ilişkilendirilmiş,düzenlenmiş ve işlenmiş ve formatlandırılmış hali olup anlam katılmış veriler şeklinde nitelendirilebilir

Bilgi: Bireyin enformasyonu algılaması,özümsemesi ve ondan sonuç çıkarması ile kişinin enformasyona yönelttiği anlam demektir
Örneğin Tc kimlik numarası bir enformasyonken bir kişinin sadece bir tc ile ilişkilendirilmesini bilmek bilgidir.

Bilgelik : Bilgilerin kişi tarafından toplanıp bir sentez haline getirilmesiyle ortaya çıkan bir olgudur. Yetenek, tecrübe gibi kişisel nitelikler birer bilgelik elemanıdır. karar alma ve kararın uygulanması sırasında tecrübe edilir.

Bilgi piramidi

VERİ AMBARI

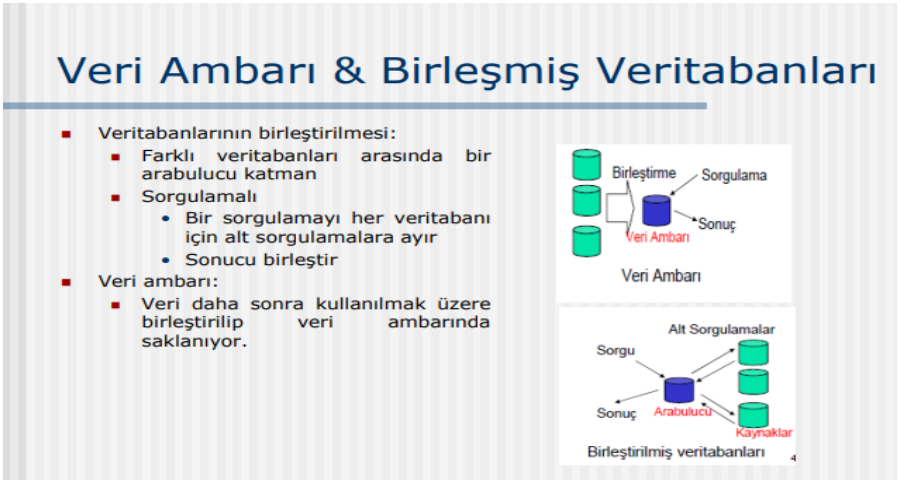
Veri Tabanı: birbiri ile ilişkili bilgilerin depolandığı alanlardır
Veri Ambarı: Veri ambarı veritabanını yormamak için oluşturulmuştur. Bir veri ambarı ilgili veriyi kolay, hızlı, ve doğru biçimde analiz etmek için gerekli işlemleri yerine getirir.Veri ambarı, işlemsel sistemlerdeki veriyi kopyalayıp, karar verme işlemi için uygun formda saklar
Data Mart: Veri ambarları bir iş probleminin tamamına yönelik bir bakış sağlarken, data mart'lar sadece belli bir kısma bakış sağlarlar.
Veri pazarları ile veriye hızlı erişim sağlayabiliriz. İkinci olarak, verinin gruplanmamış yapıda olması ve farklı iş birimlerinin farklı verileri görmesidir. Bu da bize gereksiz bir iş yükü ve güvenlik sorununa neden olmaktadır. İşte tam bu noktada, veri pazarları konuya, bölümlere uygun, veri ambarının küçük bir kopyası halinde çözüm sunmaktadır.

Veri ambarları 4 temel özelliği vardır:

Amaca Yönelik: Konuyla ilgili karar vermek için gerekli olmayan veriyi kullanmayarak konuya basit, özet bakış sağlar.Müşteri, ürün, satış gibi belli konular için düzenlenebilir.
Birleştirilmiş: Veri kaynaklarının birleştirilmesiyle oluşturulur.Veri temizleme ve birleştirme teknikleri kullanılır.
Zaman Değişkenli: Canlı veri tabanlarına göre daha uzundur
-Canlı veritabanları: en çok geçmiş 1 yıl
-Veri ambarları: geçmiş hakkında bilgi verir (geçmiş 5-10 yıl)

Değişken değil : Canlı veri tabanınkaki değişimlerden etkilenmez çünkü daha önce veri tabanından aldığı veriyi fiziksel olarak başka ortamda saklar

VERİ AMBARI İLE BİRLEŞMİŞ VERİTABANLARININ KARŞILAŞTIRILMASI



VERİ MADENCİLİĞİ İLE OLAP'IN KARŞILAŞTIRILMASI

Çevrimiçi analitik işleme (OLAP), özetlenmiş verileri sorgulamayı, ayıklamayı ve incelemeyi içeren bir veri tabanı analiz teknolojisidir. Öte yandan, veri madenciliği işlenmemiş bilgilere derinlemesine bakmayı içerir.

OLAP'ın avantajları daha geniş ve kapsamlı sonuçlar verir ve bunları kısa süreli işlemlerle gerçekleştirirken dezavantajı ise kullanıcının neyi nasıl sorması gerektiğini bilmesi gerekir ve veriden istatistiksel inceleme yapmada kullanılır

OLA NE SORUSUNA CEVAP VERİRKEN VERİ MADENCİLİĞİ NEDEN SORUSUNA CEVAP VERİR.

VERİ ÖNİŞLEME

Verinin Gürültülü olma nedenleri:

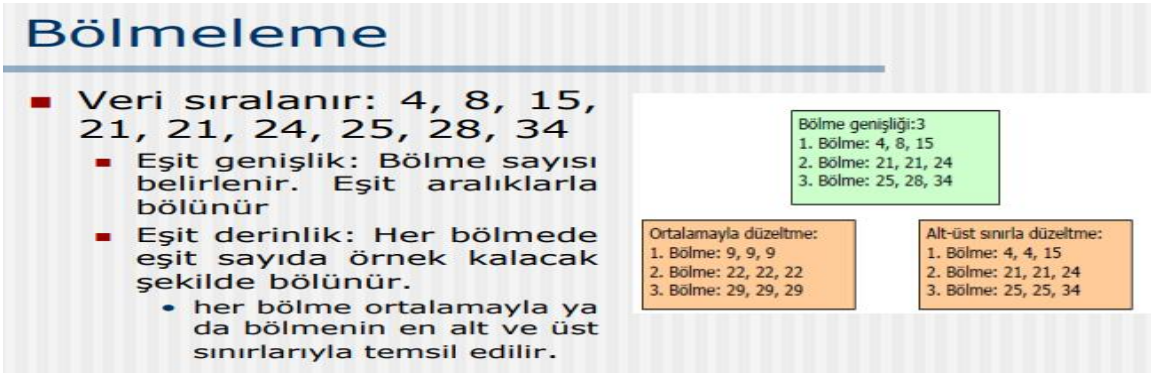
1. **Eksik veri kaydı girilmesi**(temizleme yolları: elle doldur,null veya bilinmiyor de,diğer verilerin ortalaması,olasılığı en fazla olan nitelik değerleri ile doldur)
2. **Hatalı(Gürültülü) veri kaydı girilmesi** (temizleme yolu: bölmeleme,kümeleme,eğri uydurma)
3. **Tutarsız veri kaydı girilmes**i(temizleme yolları:tutarsız veriler temizlenir)

Bu durumlara maruz kalan veri güvenilmez ve veri madenciliği sırasında kullanılamazlar

Veri önişleme aşamaları:

1. **Veriler temizlenir:** Eksik ,hatalı ve tutarsız veriler temizlenir
2. **Veriller birleştirilir:** Farklı veri kaynağındaki veriler birleştirilir
3. **Veri dönüşümü:** Normalizasyon ve biriktirme
4. **Veri azaltma:** Aynı veri madenciliği sonuçları elde edilecek şekilde veri miktarını azaltma

VERİ TEMİZLEME



KÜMELEME Benzer veriler aynı kümede olacak şekilde gruplanır ve bu kümelerin dışında kalan veriler aykırılık olarak belirlenip silinir

EĞRİ UYDURMA: Veri bir fonksiyona uydurularak doğrusal bir eğri uydurma ile bir değişkenin değeri diğer bir değişken kullanılarak bulunur

VERİ BİRLEŞTİRME

Farklı kaynaklardan veriler tutarlı olarak birleştirilmelidir aynı nitelik için farklı kaynaklarda farklı değerler olması veya farklı metrikler kullanılması,bir niteliğin başka nitelik kullanılarak hesaplanabiliyor olması birer tutarsızlıktır