
İstanbul

# GEDİK

Üniversitesi

İstanbul GEDİK Üniversitesi

# Veri Madenciliği

Öğr. Gör. Zeki ÇIPLAK

# Büyük Veri Nedir?

# Büyük Veri (Big Data) Nedir?

- **Klasik yöntemlerle işlenemeyecek kadar büyük** olan verilerdir.

- Büyük verileri, geleneksel yöntemlerle (örneğin ilişkisel veri tabanlarında) **saklamak, organize etmek zor**dur. Bu yüzden büyük verilere özel yeni yazılımlar geliştirilmiştir. (Spark, Hadoop vb. gibi)

- **Büyük verilerin 5 temel özelliği** vardır. Büyük veride bu beş özelliğin hepsinin olması gerekmez. Önemli olan, ancak yeni yöntemlerle işlenebiliyor olmasıdır.

# Büyük Verinin 5V'si

# Büyük Verinin 5V'si

1. **Volume** (Hacim)**:** *Büyük verinin hacmi, geleneksel yöntemlerle işlenemeyecek boyutta olmalıdır.*

2. **Velocity** (Hız)**:** *Büyük verinin üretilmesi çok hızlıdır. Özellikle sosyal medyada, anlık olarak terabaytlarca veri üretilir.*

3. **Verification** (Doğrulama)**:** *Büyük veri, güvenli ve doğrulanabilir olmalıdır. Doğru bilgiler içermelidir.*

4. **Variety** (Çeşitlilik)**:** *Büyük veri, çok çeşitlidir. Her farklı teknoloji, farklı tipte yapılandırılmış veya yapılandırılmamış veri üretir.*

5. **Value** (DEĞER)**:** *Büyük veri, en nihayetinde geleceğe dönük tahminlerde bulunabileceğimiz, karar vermeye destek olacak nitelikte değerli olmalıdır.*

# Bulut Bilişim Nedir?

# Bulut Bilişim Nedir?

- Klasik yöntemlerle işlenemeyecek olan büyük verini; yönetilebilmesi, analiz edilebilmesi ve faydalı hale getirilebilmesi için gerekli olan ortamı sağlayan özel teknolojilerdir.

- Bulut bilişim sistemini kullanarak, **anlık veri analizi** yapan şirketler, satışlarında büyük başarılar elde etmişlerdir.

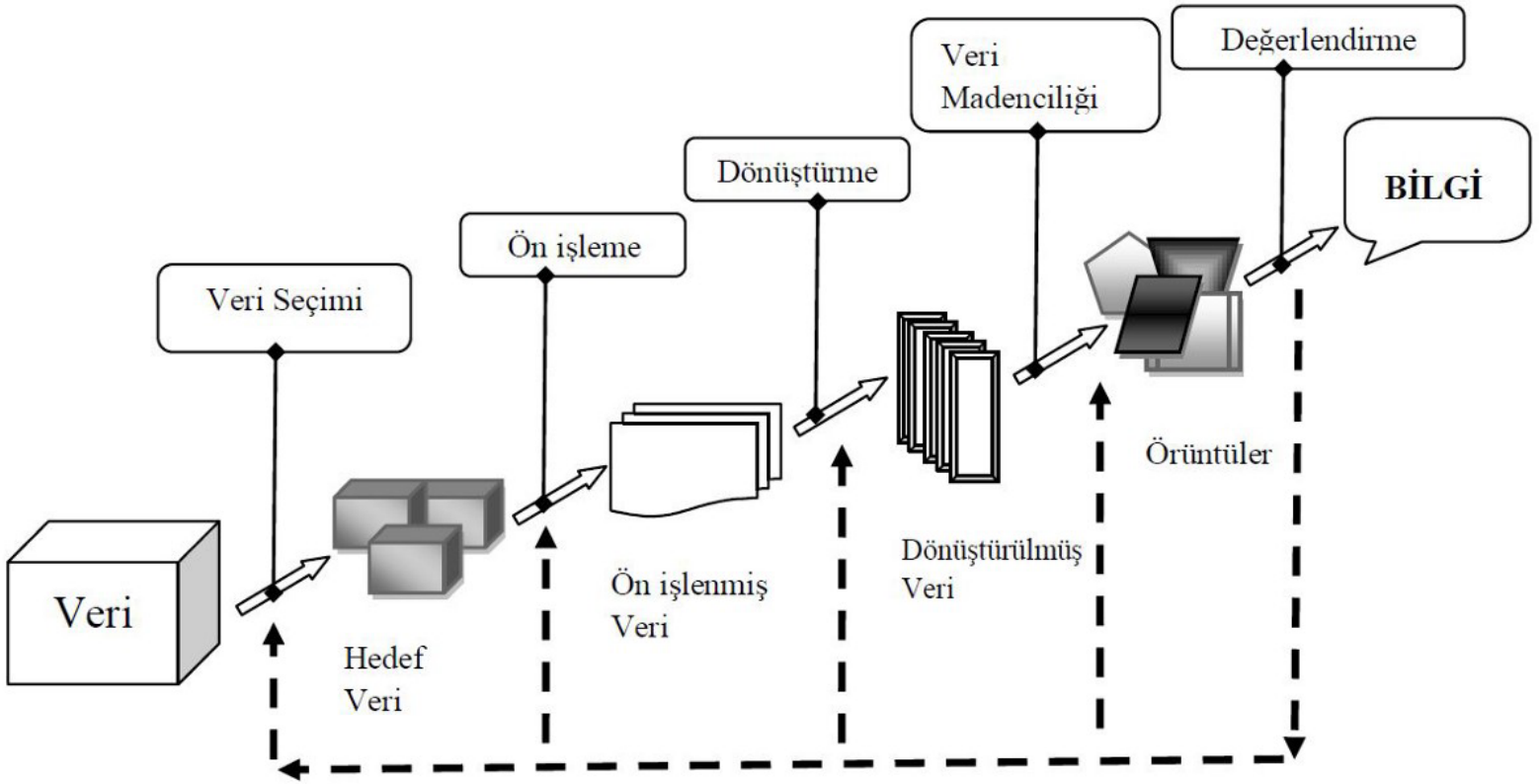
- Bu şirketlerden olan Amerikalı market zinciri **Walmart**, sattığı tüm ürünlere ait geçmiş satış verilerini kullanarak, gelecek satış tahminleri yapmış ve büyük karlar elde etmiştir.

- Anlık veri analizleri, anlık ve kişiye özel kampanyalar gibi özel uygulamalar, ancak büyük verinin hızlı ve sağlıklı işlenebilmesi ile mümkün olmuştur. Bu açıdan Bulut Bilişim'in önemi büyüktür.

# Bilgi Keşfi

# Bilgi Keşfi

- Veri Madenciliği, pratikte Bilgi Keşfi kavramı ile aynı anlamda kullanılabiliyor olsa da, aslında Bilgi Keşfinin aşamalarından biridir.



# Bilgi Keşfi

- Bilgi Keşfi, veriden faydalı bilginin keşfedilmesi sürecinin tamamıdır.

- **Bilgi Keşfinin en büyük bölümü**nü; verinin hazırlanması, dönüştürülmesi, vb. gibi ön işleme safhaları oluşturur.

- Bilgi Keşfinin bölümlerini;
  1. *Veri Madenciliği öncesi,*
  2. *Veri Madenciliği aşaması* ve
  3. *Veri Madenciliğinden sonraki süreçler*

olarak üçe ayırabiliriz.

# Bilgi Keşfinin Aşamaları

# Bilgi Keşfinin Aşamaları

1. *Amaç Tanımlama*
2. *Veri Üzerinde Yapılan Önişlemler*
3. *Model Kurma ve Değerlendirme*
4. *Modeli Yorumlama*
5. *Modelin İzlenmesi*

# 1. Amaç Tanımlama

- Veri Madenciliğine temel olan araştırma konusu ile ilgili temel bilgilerin yer aldığı safhadır.

- **Çalışmanın neden yapıldığı**yla ilgili açıklamalar ve hangi problemi çözmeye yönelik olduğu açıkça ifade edilmelidir.

- Ayrıca sonuçların **ne kadar başarılı** olduğunun, hangi metriklere göre ölçüleceği de açıklanabilir.

- Ek olarak, proje sonunda yapılacak olan tahminlerin veya **veriden çıkarılan bilgilerin** doğru/yanlış olması durumlarında meydana gelecek kazanım/maliyet'e ilişkin açıklamalar da yer alabilmektedir.

# 2. Veri Üzerinde Yapılan Önişlemler

- Veri Madenciliği projesinin, **en önemli safhası**dır.

- Veri Madenciliği projelerinde kullanılacak **veriler, çoğu zaman kullanılmaya hazır durumda olmazlar**. Bu yüzden bazı ön işlemlerden geçirilmesi gerekir.

- Veri Madenciliği sonucu elde edilen **yeni bilgi, ancak kaliteli verilerle** çalışıldığı zaman, yararlı olabilir. Düzgün yapılandırılmamış veri setleri, projeden istenilen sonucun alınamamasına sebep olabilir.

- Önişlemler aşamasında **gereken titizlik gösterilmezse**, bu durumda modelin kurulması aşamasında problemler çıkabilmektedir.

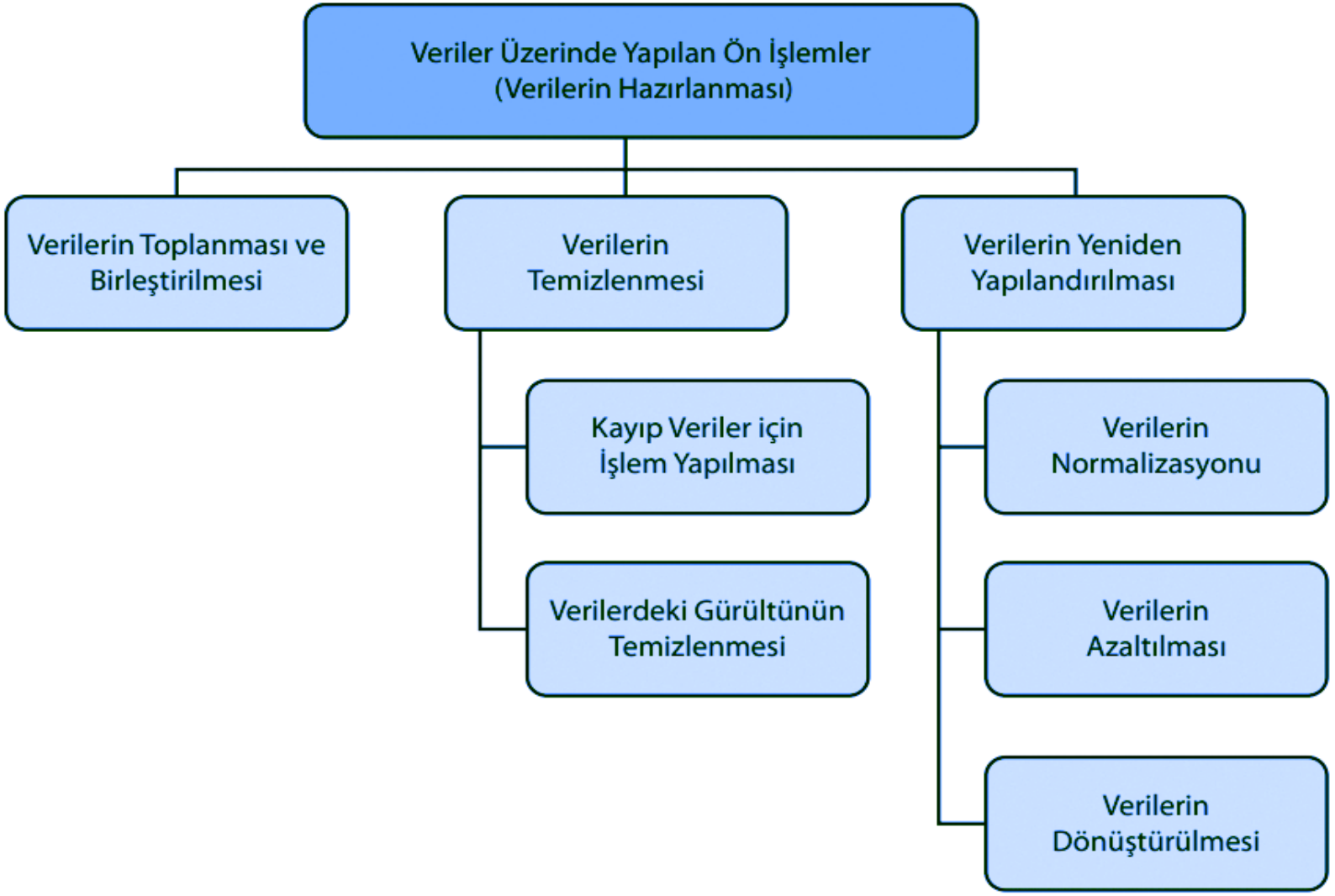
# 2. Veri Üzerinde Yapılan Önişlemler

- Önişlemler aşaması;
    1. *Verilerin toplanması ve birleştirilmesi,*
    2. *Verilerin temizlenmesi,*
    3. *Verilerin dönüştürülmesi ve yeniden yapılandırılması*

şeklinde bölümlere ayrılabilir.

- Yukarıda sayılan bölümler de, her birinin içerisinde yapılan farklı işlemlere göre, alt bölümlerde değerlendirilebilirler.

# 2. Veri Üzerinde Önişlemler



# Verilerin Toplanması ve Birleştirilmesi

- Önceki aşamada belirlenmiş **amaca uygun olan veriler**, çok çeşitli kaynaklardan toplanabilir ve bir araya getirilebilirler.

- **Veri kaynakları**, şirketin kendi iç verileri olabildiği gibi, veri pazarlayan veya ücretsiz erişime açılmış çeşitli kurumlara ait veriler de olabilmektedir.

- Kullanılacak olan **verilerin**, hangi kaynaklardan toplandığı ve **doğruluk/güvenilirlik derecelerinin ne olduğu** konusu, Veri Madenciliği projesinin ulaşacağı başarıyı doğrudan etkilemektedir.

# Verilerin Temizlenmesi

| ID | Ad/Soyad | Doğum Tarihi | Boy | Kilo |
|---|---|---|---|---|
| 1 | Ali Öztürk | 4.10.2001 | 174 | 78 |
| 2 | Ebru Yılmaz | 9.05.1998 | 168 | 60 |
| 3 | Irmak Ece Aygün | 7.03.1999 | N/A | 56 |
| 4 | Orhan Uzun | NULL | 173 | 80 |
| 5 | Okan Eren | 6.06.2000 | 179 | 0 |
| 6 | Umut Ali Akan | 17.08.2002 | 185 | 90 |
| 7 | Kemal Sarı | 23.01.2003 | N/A | 89 |
| 8 | Beren Yumak | 16.02.2000 | 170 | 59 |
| 9 | Yasemin Keser | NULL | 166 | 60 |
| 10 | Rabia Esin Acar | 1997 | 1,59 | |

- Elde edilen verilerdeki kayıtların bir kısmı eksik doldurulmuş olabilir. Bazı veriler hatalı ve anlamsızlık oluşturacak şekilde de girilmiş olabilir. Bu tip verilere **Kayıp Veri** denir.

# Verilerin Temizlenmesi

- Bazı veriler de doğru olmayacak kadar uç değerler içeriyor olabilir. (*Bir insanın boyunun 5 metre olması gibi*) Bu tip değerlere de **Aykırı Veriler** denir.

- Kayıp ve Aykırı verilerin tümüne **Gürültülü Veri** denir.

- **Gürültü Verinin oluşma sebepleri;**
  - *Birçok veri setinin bir araya getirilmiş olması,*
  - *Veri girişi yapan personellerin hataları,*
  - *Özniteliklerin birimlerine dikkat edilmemesi,*
  - *Sistemsel hatalar,*
  - *Yanlış yöntemlerle veri toplanması*

vb. gibi sebeplerdir.

# Kayıp Veri Problemini Çözmek

*Kayıp veri içeren kaydı/satırı veri setinden silmek*:
Kayıp veri sayısı çok az ise ve çalışma sonucunu çok etkilemeyeceği düşünülüyorsa bu yöntem kullanılabilir.

*Kayıp verileri, tek tek doldurmak:*
Veri seti çok büyük değilse ve kayıp verilere ulaşabilmek mümkünse bu yöntem uygulanabilir.

*Kayıp verilerin hepsi için aynı veriyi girmek:*
Veri setinde kayıp olmayan veriler birbirlerine çok yakın değerlerde ise, kayıp olan veriler için tek bir değer belirlenip, o değer kayıp veri yerine kullanılabilir.

# Kayıp Veri Problemini Çözmek

| ID | Kira Bedeli | Semt |
|---|---|---|
| 1 | 4350 | Kurtköy |
| 2 | 4150 | Kurtköy |
| 3 | 5500 | Yenişehir |
| 4 | 4500 | Sülüntepe |
| 5 | 6200 | Yenişehir |
| 6 | 6350 | Yenişehir |
| 7 | 4650 | Sülüntepe |
| 8 | | Kurtköy |

*Kayıp veri yerine, diğer verilerin ortalamasını girmek:*

Diğer verilerin ortalaması girilirken, ya tüm verilerin ya da benzer kategoride olan verilerin ortalaması alınır.

# Kayıp Veri Problemini Çözmek

*Kayıp verileri, diğer verileri kullanarak tahmin etmek:*
Tahmine dayalı istatistiksel veya makine öğrenmesi algoritmaları (regresyon, karar ağacı, sınıflandırma veya zaman serisi analizi vb.) ile kayıp veriler tahmin edilerek doldurulabilir.

# Verilerin Yeniden Yapılandırılması

- Veri Madenciliği algoritmaları, ancak veri setindeki sayısal değerleri kullanabilir. Bu amaçla kategorik değişkenler varsa, sayısal değerlere dönüştürülebilir (**one hot encoding** gibi) veya alınacak bir kararla ilgili öznitelikler veri setinden tamamen de çıkartılabilir. (Araştırınız)

- Veriler arasındaki ölçekler, birbirinden çok farklı olabilir. Örneğin, bazı özniteliklerin sayısal aralığı 0-100 iken, bazı özniteliklerin sayısal aralığı 0-10000 olabilmektedir.

- Tüm özniteliklerin aynı ölçekte olması, bulunacak katsayıların doğru değerlendirilebilmesi ve algoritmaların doğru çalışması açısından çok önemlidir. Bu amaçla **normalizasyon** ve **standardizasyon** işlemleri yapılabilir.

# Verilerin Yeniden Yapılandırılması

- **Normalizasyon işlemi**, verilerin belli bir aralığa çekilmesi işlemidir. Bu aralık 0-1 veya 0-100 gibi belirlenen bir aralık olarak seçilebilmektedir.

$$x_{\text{norm}} = \frac{x - \min(x)}{\max(x) - \min(x)}$$

- **Standardizasyon işlemi** ise, ortalama ve standart sapmaya göre verileri yeni bir ölçeğe çekme işlemidir. Standardizasyona, **Z-Skor Normalizasyon** da denmektedir.

$$x_{\text{stand}} = \frac{x - \text{mean}(x)}{\text{standard deviation } (x)}$$

# 3. Model Kurma ve Değerlendirme

- Son hale gelen **verinin, bir modele dönüştüğü** aşamadır.

- Bu aşamada **hangi algoritmanın kullanılacağı** belirlenir ve model denendikten sonra (bir sonraki aşamada), kurulan modelin en başta belirlenen **amaca uygun olup olmadığı** sorgulanır.

- Farklı algoritmalarla farklı modeller oluşturulabilir ve en iyi olduğu düşünülen model bulunana kadar bazı aşamalar tekrarlanabilir.

- Veri setinin önişleme aşaması titizlikle ele alınmazsa, bu aşamada **çokça model denemesi** olur ve bir türlü istenen sonuca ulaşılamayabilir.

# 4. Modeli Yorumlama

- Bu aşamada model seçilmiş ve **uygulamaya konmuştur**. Modelin çalıştırılmasıyla birlikte elde edilen bilgiler yorumlanmaya başlanır.

- Kazanılan bilgi, verilecek kararlara destek olacak şekilde kullanılır.

# 5. Modelin İzlenmesi

- Kurulan model ne kadar iyi olursa olsun, zaman içerisinde ihtiyaca cevap veremeyecek hale gelebilmektedir. Bu yüzden sürekli izlenmesi ve **değişiklik/düzenleme gerekiyorsa**, modele müdahale edilmesi gerekir.