

# Veri Madenciliğine Giriş

Kurumlarda biriken veri içerisinde kurum için yararlı olan verileri bulup ortaya çıkarma işlemine veri madenciliği adı verilir.

Veriyi yönetmek için veri ambarı ve verileri çözümleyerek yararlı bilgiye ulaşılmasını sağlayan veri madenciliği kavramları ortaya atılmıştı. Veriler üzerinde çözümlemeler yapmak adına çeşitli matematiksel ve istatistiksel işlemler uygulanabilir.

## Veriler Nasıl Toplanır?

Veri toplamanın 2 temel yolu vardır. Bu yollardan ilki, sizden izin alınarak yapılan ve girmiş olduğunuz bilgiler ya da site içerisindeki davranışlarınıza dayalı olan “Açık veri toplama” yöntemidir. Diğer yöntem ise “Kapalı veri toplama” olarak adlandırılır ve genellikle sosyal ağlar, arama motorları gibi sitelerde toplanan verilerdir. Bu tür veri toplama yönteminin amacı, o site içerisindeki tüm davranışları takip edip, kullanıcının ilgi alanlarını belirlemek ve bunlara uygun pazarlama faaliyetleri yürütmektir.

## Veri Madenciliği Süreçleri Nelerdir?

**Veri Filtreleme:** Madencilikte kullanılacak verilerin belirlenme aşamasıdır.

**Veri Temizliği:** Toplanmış olan verinin içerisinde gereksiz, tutarsız ya da gürültülü olanların ayıklanması adıdır. Veri tabanında yer alan tutarsız ve hatalı veriler **gürültü** olarak ifade edilir. Eksik verilerin yerine yeni değerler belirlenmelidir. Örneğin;

- Bu verileri içeren kayıtlar veri setinden silinebilir.
- Kayıp değerlerin yerine belirlenen sabit bir değer kullanılabilir. Ancak bu sorun yaratabilir.
- Değişkenlerin tüm verileri kullanılarak ortalaması hesaplanır ve eksik değer yerine bu değer konulabilir.
- Değişkenlerin tüm değerleri yerine sadece bir sınıfa ait verilerin ortalaması hesaplanarak eksik değer yerine bu değer konulabilir.
- Verilere regresyon, karar ağacı modeli gibi uygun bir tahmin yapılarak eksik değer tahmin edilip kullanılabilir.

**Veri Birleştirme:** Farklı kaynaklardan elde edilen ve benzer niteliklere sahip ya da ilişkili veriler bu adımda birleştirilir.

**Veri İndirgeme:** Temizlenmiş ve birleştirilmiş verilerden, analize uygun olanların seçilme işlemi bu adımda gerçekleştirilir. Eğer çözümlemede elde edilecek sonucun değişmeyeceğine inanılıyorsa veri sayısı ya da değişkenlerin sayısı azaltılabilir.

Veri indirgeme çeşitli biçimlerde yapılabilir.

- Veri birleştirme ya da veri küpü
- Boyut indirgeme
- Veri sıkıştırma
- Örnekleme
- Genelleme

**Veri Dönüştürme:** Veriyi bazı durumlarda veri madenciliği çözümlerine direk katmak uygun olmayabilir. Eldeki verilerin madencilik için uygun biçime dönüştürülmesi aşamasıdır. Değişkenlerin ortalama ve varyansları birbirlerinden önemli ölçüde farklı olduğu takdirde büyük ortalama ve varyansa sahip değişkenlerin diğerleri üzerindeki baskısı daha fazla olur ve onların rolleri önemli ölçüde azaltır. Ayrıca değişkenlerin sahip olduğu çok büyük ve çok küçük değerler de çözümlemelerin sağlıklı biçimde yapılmasını engeller. Bu nedenle bir dönüşüm yöntemi uygulayarak söz konusu değişkenlerin normalleştirilmesi veya standartlaştırılması uygun bir yol olacaktır.

- **Min-Max Normalleştirme:** Verileri 0 ile 1 aralığındaki değerlere dönüştürme işlemidir.
- **Z-score Standartlaştırması:** Verilerin ortalaması ve standart sapma göz önüne alınarak yeni değerlere dönüştürülmesi esasına dayanmaktadır.

**Madencilik Çalışması:** Bu aşamada, daha önceden hazırlanmış olan veriler üzerinde, amaca uygun veri madenciliği algoritmaları uygulanır.

**Yorumlama Doğrulama ve Sunum:** Veri madenciliği uygulaması gerçekleştirildikten sonra, elde edilen sonuçlar yorumlanır ve bu sonuçların doğrulukları ile ilgili araştırma yapılır. Doğrulama işlemi, farklı uygulamalardan elde edilen sonuçların karşılaştırılması ile gerçekleştirilir. Sonuçlar çoğu zaman grafiklerle desteklenir. Örneğin hiyerarşik kümeleme modeli uygulanmış ise sonuçlar **dendrogram** adı verilen özel bir grafik ile sunulur.

## Veri Madenciliği Yöntemleri Nelerdir?

Veri madenciliği konusunda çok sayıda yöntem ve algoritma geliştirilmiştir. Bu yöntemlerin bir çoğu istatistiksel tabanlıdır.

**Sınıflama:** Veri madenciliğinde en yaygın kullanılan yöntemlerden biri olup veri tabanlarındaki gizli örüntüleri ortaya çıkarmakta kullanılır. Bu yöntemde, eldeki verinin niteliklerini inceleyip, daha önceden belirlenmiş olan sınıflardan uygun olanına aktarmaktır.

**Birliktelik Kuralları:** Veri madenciliğinde yaygın olarak kullanılan yöntemlerden biridir. Bu yöntem, büyük boyutlardaki veri tabanlarında yer alan ve birbirleriyle bağlantılı verileri ve aralarındaki bağlantıları belirlemeyi amaçlar. Özellikle pazarlama alanında kullanılır.

**Kümeleme:** Bu veri madenciliği yönteminde amaç, verilerin kendi aralarındaki ilişkilere göre alt sınıflara ayrılmasını sağlamaktır.

**Tahminleme:** Bir veri seti içerisinde eksik durumda olan sayısal verilerin tahmin edilmesine dayanan veri madenciliği yöntemidir.

**Aykırlık Analizi:** Veriler içerisinde aşırı sapma olanların tespit edilmesi yöntemidir. Bu yöntemle, kredi kartlarından gerçekleştirilen olağandışı harcamalar tespit edilmektedir.

## Veri Madenciliği Hangi Alanlarda Kullanılır?

Veri madenciliğinin kullanım alanları ile ilgili herhangi bir sınırlama yoktur ve verinin biriktiği her yerde veri madenciliği yapılabilir. Günümüzde veri madenciliğinin en yaygın olarak kullanıldığı alanlara bakacak olursak;

- Bankacılık:
  - Kredi kartı dolandırıcılıkların ve sahtekarlıklarının belirlenmesi.
  - Kredi kartı kullanım alışkanlıklarına göre müşterilerin belirlenmesi.
  - Kredi taleplerine ilişkin değerlendirmelerin yapılması.
- Pazarlama:
  - Bireylerin satın alma alışkanlıklarının belirlenmesi.
  - Müşteri özellikleri arasındaki bağlantıların ortaya konulması.
  - Satış tahmini.

- Müşteri değerlendirme.
- Pazar sepeti analizi.
- Müşteri ilişkileri yönetimi.
- CRM:
  - Müşteri sadakatının artırılması.
  - Pazarlama kampanyalarından en yüksek seviyede yarar sağlama çalışmalarının yapılması.
- E-Ticaret:
  - Sunuculara yapılan saldırıların tespit edilmesi.
  - e-CRM uygulamalarının yönetimi
  - Web sitesinde gezinen kullanıcıların davranışlarının belirlenmesi.
- Sigortacılık
  - Sigorta risk gruplarının belirlenmesi.
  - Sigorta dolandırıcılıklarının tespiti.