VERİ MADENCİLİĞİNE GİRİŞ

Bölüm Hedefi

İşletmelerin ürettiği büyük miktardaki veriler arasında belirli bazı kuralları yada örüntüleri ortaya çıkarmak önem taşımaktadır. Veri madenciliği bu tür amaçla kullanılan yöntemler topluluğu olarak karşımıza çıkmaktadır. Veri madenciliğinin temel yöntemleri arasında sınıflandırma, kümeleme ve birliktelik kurallarının elde edilmesi sayılabilir.

2.1. Veriyi Bilgiye Dönüştürmenin Yolu

Günümüzde bilişim alanındaki gelişmelerin ne kadar baş döndürücü bir hızda geliştiğini gözlemleyebiliyoruz. Bilgisayar teknolojilerinde her gün bir başka yenilikle karşılaşıyoruz.

Sadece bilgisayarlar değil veri iletişim teknolojilerinde de hızlı bir gelişme söz konusudur. Üstelik dikkati çeken önemli noktalardan birisi söz konusu teknolojilere dayalı ürünlerin gittikçe daha ucuzlamasıdır. Kullanıcı daha yetenekli, daha hızlı ve kullanışlı bilgisayar teknolojisine kolayca sahip olabiliyor (1).

Bilişim teknolojilerindeki bu gelişme beraberinde bir sorunu da getirmiştir. Bilişim sistemleri sayesinde artık her bilgi **sayısal ortama** kaydedilmektedir.

Örneğin bir mağazada satışlar ve müşterilerle ilgili her türlü bilgi sayısal ortamda yerini almaktadır. Üstelik günlük tüm veriler sayısal ortamda saklanmaktadır. Binlerce müşterisi olan bir mağaza her gün çok sayıda veri üretmek zorunda kalmaktadır. Böylece ilgili firma bilgisayarlarında çok büyük miktarda veri birikmektedir. Bir kredi kartı firmasında yada telefon iletişimi yapan bir firmada günlük bazda biriken verilerin miktarı çok büyük olacaktır.

Bilişim teknolojisi bu devasa verileri saklamaya yeterli olabilir. Ancak bu veriler ne işe yarayacaktır ? Bu verilerden firma bazı avantajlar kazanabilecek midir ?

Biriken veri gerçek anlamda "**bilgiye**" dönüştürülebilecek midir ? Bu tür sorulara olumlu yanıt vermek mümkündür. Büyük ölçekli verilerden yararlanarak **firma karar destek sistemleri**oluşturulabilir. Bu veriler üzerinde çözümlemeler yapılarak özellikle stratejik seviyedeki kararlara destek sağlanabilir.

Veriler üzerinde çözümlemeler yapmak amacıyla çeşitli istatistiksel ve matematiksel yöntemler kullanılabilir. Ancak veri sayısı arttıkça sorunlar ortaya çıkacaktır.

Özellikle ilişkisel veri tabanları üzerinde bu çözümlemeleri yapmak zorlaşacaktır. Bu tür veriler üzerinde çözümlemeleri yapabilmek için hem yeni veri tabanı kavramlarına hem de yeni çözümleme yöntemlerine gereksinim duyulmaktadır. Veriyi yönetmek için "**Veri Ambarı**" ve verileri çözümleyerek "**yararlı bilgiye**" erişilmesini sağlayan "**Veri Madenciliği**" kavramları ortaya atılmıştır.

2.1.1. Veri Madenciliği

Veri madenciliği konusunda çeşitli tanımlar yapılmaktadır. Basit bir tanım yapmak gerekirse, **veri madenciliği**, büyük ölçekli veriler arasından belirli bir bilgiyi elde etme işidir [(1)]. Bu sayede veriler arasındaki ilişkileri ortaya koymak ve gerektiğinde ileriye yönelik kestirimlerde de bulunmak mümkün görülmektedir.

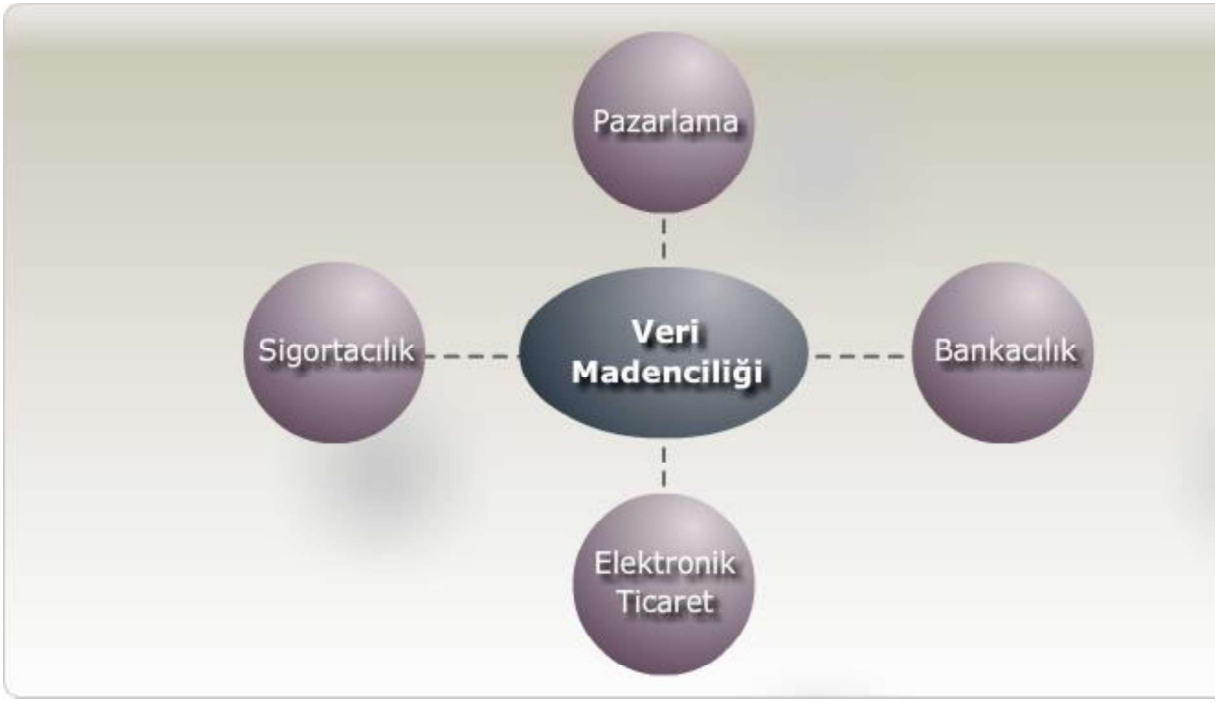
Bunun anlamı, veri madenciliği, bir kurumda üretilen tüm verilerin belirli yöntemler kullanarak var olan yada gelecekte ortaya çıkabilecek gizli bilgiyi su yüzüne çıkarma süreci olarak değerlendirilebilir. Bu açıdan bakıldığında, veri madenciliği işinin kurumların karar destek sistemleri için önemli bir yere sahip olabileceğini söyleyebiliriz.

Veri madenciliği aslında **klasik istatistiksel uygulamalara** çok benzer. Ancak klasik istatistiksel uygulamalar yeterince düzenlenmiş ve çoğunlukla özet veriler üzerinde çalıştırılır. Ayrıntı bilgi olsa bile, burada kayıtlar binlerce olabilir.

Veri madenciliğinde ise milyonlarca ve hatta milyarlarca veri ve çok daha fazla değişken ile ilgilenilir. Veri sayısı çok olunca, bazı özel analiz algoritmaların geliştirilmesi gerekmiş, ayrıca verinin saklandığı ortamların da örneğin veri ambarı biçiminde yeniden düzenlenmesini gerekli kılmıştır.

2.1.2. Uygulama Alanları

Veri madenciliğinin günümüzde yaygın bir kullanım alanı bulunmaktadır. Örneğin pazarlama, bankacılık ve sigortacılık gibi alanlarda ve elektronik ticaret ile ilgili alanlarda yaygın şekilde kullanılmaktadır.



Pazarlama

- Müşterilerin satın alma alışkanlıklarının belirlenmesi,
- Müşterilerin demografik özellikleri arasındaki bağlantıların ortaya konulması,
- Mevcut müşterilerin elde tutulması, yeni müşterilerin kazanılması,
- Pazar sepeti analizi
- Müşteri ilişkileri yönetimi
- Müşteri değerlendirme
- Satış tahmini

Sigortacılık

- **Yeni poliçe talep edecek müşterilerin tahmin edilmesi,**
- **Sigorta dolandırıcılıklarının tespiti,**
- **Riskli müşteri gruplarının belirlenmesi.**

Bankacılık

- **Farklı finansal göstergeler arasında gizli ilişkilerin ortaya konulması,**
- **Kredi kartı dolandırıcılıklarının ve sahtekarlıkların belirlenmesi,**
- **Kredi kartı harcamalarına göre müşteri gruplarının belirlenmesi,**
- **Kredi taleplerinin değerlendirilmesi.**

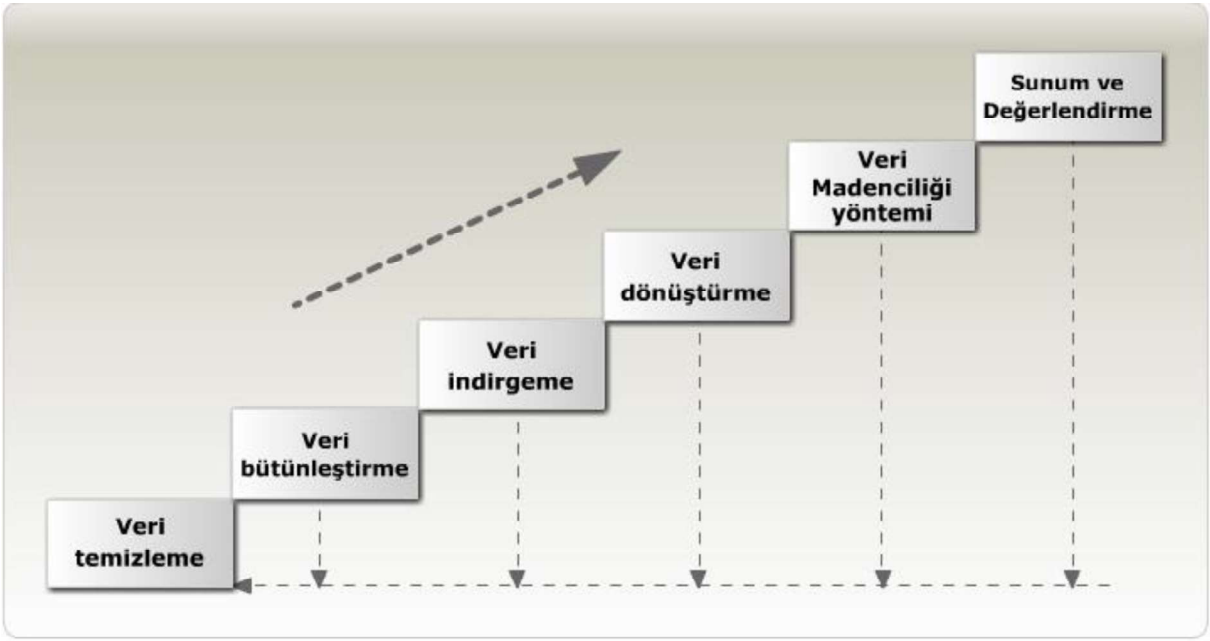
Elektronik Ticaret

- **Saldırıların çözümlenmesi**
- **e-CRM uygulamalarının yönetimi**
- **Web sayfalarına yapılan ziyaretlerin çözümlenmesi.**

2.1.3. Veri Madenciliği Süreci

Veri madenciliğini bir süreç olarak değerlendirmek gerekiyor. Söz konusu süreç aşağıda belirtilen adımları içermektedir:

1. Veri temizleme
2. Veri bütünleştirme
3. Veri indirgeme
4. Veri dönüştürme
5. Veri madenciliği algoritmasını uygulama
6. Sonuçları sunum ve değerlendirme



Şekil 2.1: Veri madenciliği süreci

2.1.3.1. Veri Temizleme

Bazı uygulamalarda, üzerinde çözümleme yapılacak verilerin istenen özelliklere sahip olmadığı görülebilir.

Örneğin eksik verilerle ve uygun olmayan verilerin oluşturduğu tutarsız verilerle karşılaşılabilir. Veri tabanında yer alan tutarsız ve hatalı verilere **gürültü** olarak değerlendirilmektedir. Bu gibi durumlarda verinin söz konusu sorunlardan temizlenmesi gerekecektir. Eksik verilerin yerine yenileri belirlenerek konulmalıdır. Bunun için aşağıda belirtilen yöntemlerden biri kullanılabilir.

- Eksik değer içeren kayıtlar veri kümesinden atılabilir.
- Kayıp değerlerin yerine bir genel sabit kullanılabilir. Bütün kayıp değerler için aynı sabit kullanılabilir. Örneğin "bilinmiyor" değeri bu eksik veri yerine kullanılabilir. Ancak bütün değişkenlerde kayıp değerler yerine aynı sabit değerin kullanımı sorun yaratacaktır (*HAN, 2000*).

AD	KAYIT_TAR
Hakan Kırık	Bilinmiyor
Deniz Bilmiş	07.08.04

- Değişkenin tüm verileri kullanılarak ortalaması hesaplanır ve eksik değer yerine bu değer kullanılabilir.
- Değişkenin tüm verileri yerine, sadece bir sınıfa ait örneklerin değişken ortalaması hesaplanarak eksik değer yerine kullanılabilir.
- Verilere uygun bir tahmin yapılarak, örneğin regresyon yada karar ağacı modeli kurularak eksik değer tahmin edilebilir ve eksik değer yerine kullanılabilir.

2.1.3.2. Veri Bütünleştirme

Veri bütünleştirmesinin ne anlama geldiği üzerinde **"Veri Ambarı"** ile ilgili bölümde durmuştuk. Farklı veri tabanlarından yada veri kaynaklarından elde edilen verilerin birlikte değerlendirmeye alınabilmesi için farklı türdeki verilerin tek türe dönüştürülmesi yani bütünleştirilmesi söz konusu olacaktır.

Eğer veri madenciliği uygulaması için bir veri ambarı altyapısı hazırlanmış ise söz konusu veri bütünleştirme işleminin yapılmış olması gerekmektedir. Ancak böyle bir yapı yoksa söz konusu veri bütünleştirme işleminin doğrudan veri madenciliğine esas oluşturacak veriler üzerine uygulanması gerekecektir.

2.1.3.3. Veri İndirgeme

Veri madenciliği uygulamalarında bazen çözümleme işlemi uzun süre alabilir. Eğer çözümlemeden elde edilecek sonucun değişmeyeceğine inanılıyorsa veri sayısı yada değişkenlerin sayısı azaltılabilir. Veri indirgeme çeşitli biçimlerde yapılabilir (*HAN, 2000*):

Veriyi indirgeme aşamasında verileri, çok boyutlu **veri küpleri** biçimine dönüştürmek söz konusu olabilir. Böylece çözümlemeler sadece belirlenen boyutlara göre yapılır. Veriler arasında bir seçme işlemi yapılarak, gereksiz veriler veri tabanından çıkarılır ve **boyut azaltılması** sağlanabilir.

Veri sıkıştırma aşamasında, büyük veri kümelerinin sıkıştırılarak daha az yer işgal etmeleri sağlanır. **Örnekleme** aşamasında ise, büyük veri topluluğu yerine onu temsil eden daha küçük veri kümelerinin oluşturulması amaçlanır.**Genelleme** verilerin tek tek değil, genel kavramlarla ifade edilmesini sağlar.

2.1.3.4. Veri Dönüştürme

Veriyi bazı durumlarda veri madenciliği çözümlemelerine aynen katmak uygun olmayabilir. Değişkenlerin ortalama ve varyansları birbirlerinden önemli ölçüde farklı olduğu takdirde büyük**ortalama** ve **varyansa** sahip değişkenlerin diğerleri üzerindeki baskısı daha fazla olur ve onların rollerini önemli ölçüde azaltır.

Ayıca 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.

2.1.3.4.1. Min-Max Normalleştirilmesi

Verileri 0 ile 1 arasındaki sayısal değerlere dönüştürmek için **min-max normalleştirme** yöntemi uygulanır. Bu yöntem, veri içindeki en büyük ve en küçük sayısal değerin belirlenerek diğerlerini buna uygun biçimde dönüştürme esasına dayanmaktadır. Söz konusu dönüştürme bağıntısı şu şekilde ifade edilmektedir:

$$X^* = \frac{X - X_{min}}{X_{max} - X_{min}}$$

Burada **X*** dönüştürülmüş değerleri, **X** gözlem değerini, $\mathbf{X_{min}}$ en küçük gözlem değerini ve $\mathbf{X_{max}}$ en büyük gözlem değerini ifade etmektedir.

2.1.3.4.1.1. Örnek

Aşağıdaki tabloda yer alan **X** değişkeni değerlerine **min-max normalleştirme bağıntısını** uygulayarak dönüştürmek istiyoruz. Bunun için, veriler için önce aşağıdaki değerler belirlenir:

$X_{min} = 30$

$X_{max} = 62$

Bu değerlere dayanarak **X** örneğini birinci elemanı için şu şekilde bir hesaplama yapılır:

$$x^* = \frac{X - X_{min}}{X_{max} - X_{min}} = \frac{30-30}{62-30} = 0$$

Benzer biçimde diğer gözlemler için aynı hesaplamalar yapılır.

X	X^*
30	0,0000
36	0,1875
45	0,4688
50	0,6250
62	1,0000

Tablo 2.1: Min-max normalleştirme dönüşümü sonucu elde edilen değerler

2.1.3.4.2. Z-score Standartlaştırma

İstatistik çözümlemelerde sıkça kullanılan bir dönüşüm biçimi **Z-score** adıyla anılmaktadır. Bu yöntem, verilerin ortalaması ve standart hatası göz önüne alınarak yeni değerlere dönüştürülmesi esasına dayanmaktadır .Söz konusu dönüşümlerde şu şekilde bir bağıntıya yer verilir:

$$X^* = \frac{X - \overline{X}}{\sigma_X}$$

Burada X^* dönüştürülmüş değerleri, X gözlem değerleriini, $\overline{X}$ verilen aritmetik ortalamasını ve σ_X gözlem değerlerinin sapmasını ifade etmektedir.

2.1.3.4.2.1. Örnek

Önceki örnekte ele alınan verileri bu kez **z-score standartlaştırılmasını** uygulayarak dönüştüreceğiz. Bu amaçla önce aşağıdaki hesaplamaların yapılması gerekmektedir.

Bunların birincisi $\overline{X}$ aritmetik ortalamanın bulunmasıdır.

$$\overline{X} = \frac{1}{n} \sum_{i=1}^{n} X_i = 44.6$$

Z-score standartlaştırma işlemi için X serisinin standart hatasının bunması gerekmektedir. Söz konusu hata şu şekilde hesaplanır:

$$\sigma_x = \sqrt{\frac{\sum_{i=1}^{n} (X_i - \overline{X})^2}{n-1}} = 12.44$$

Bu durumda birinci satır için Z-score dönüşümü şu şekilde olabilir:

$$X^* = \frac{X - \overline{X}}{\sigma_x} = \frac{30-44.6}{12.44} = -1.1735$$

Benzer biçimde diğer gözlemler içinde hesaplamalar yapılarak aşağıdaki tablo elde edilir

X	X^*
30	-1,1735
36	-0,6912
45	0,0321
50	0,4340
62	1,3985

Tablo 2.2: Z-score dönüşümü sonucu elde edilen değerler

2.1.3.5. Veri Madenciliği Algoritmasını Uygulama

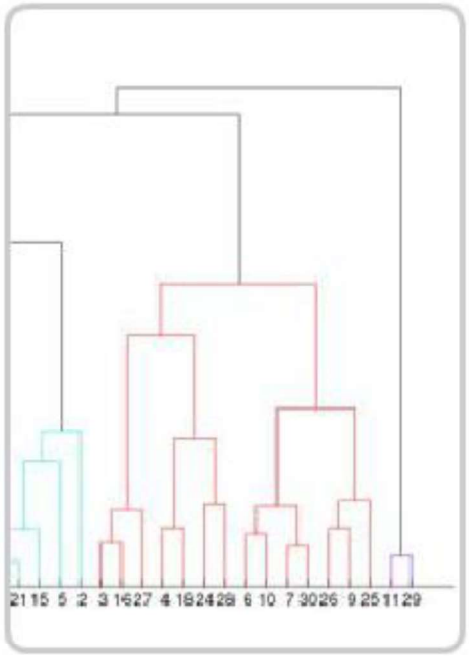
Veri madenciliği yöntemlerini uygulayabilmek için burada sıralanan işlemlerin uygun görülenleri yapılır. Veri hazır hale getirildikten sonra konuyla ilgili **veri madenciliği algoritmaları**uygulanır.

Bu algoritmaların bir kısmını bu ders kapsamında ele alarak inceleyeceğiz. Söz konusu algoritmalar sınıflandırma, kümeleme ve birliktelik kuralları konusunda olacaktır.

2.1.3.6. Sonuçları Sunum ve Değerlendirme

Veri madenciliği algoritması veriler üzerinde uygulandıktan sonra, sonuçlar düzenlenerek ilgili yerlere sunulur. Sonuçlar çoğu kez grafiklerle desteklenir.

Örneğin bir **hiyerarşik kümeleme modeli** uygulanmış ise sonuçlar **dendrogram** adı verilen özel grafiklerle sunulur.



2.1.4. Veri Madenciliği Yöntemleri

Veri madenciliği konusunda çok sayıda yöntem ve algoritma geliştirilmiştir. Bu yöntemlerin bir çoğu istatistiksel tabanlıdır. Biz bu ders kapsamında üç ayrı konuda veri madenciliği modellerini tanıtacağız.

Söz konusu veri madenciliği modellerini temel olarak şu şekilde gruplandırılabiliriz:

2.1.4.1. Sınıflandırma

Sınıflama, veri madenciliğinde sıkça kullanılan bir yöntem olup, veri tabanlarındaki **gizli**