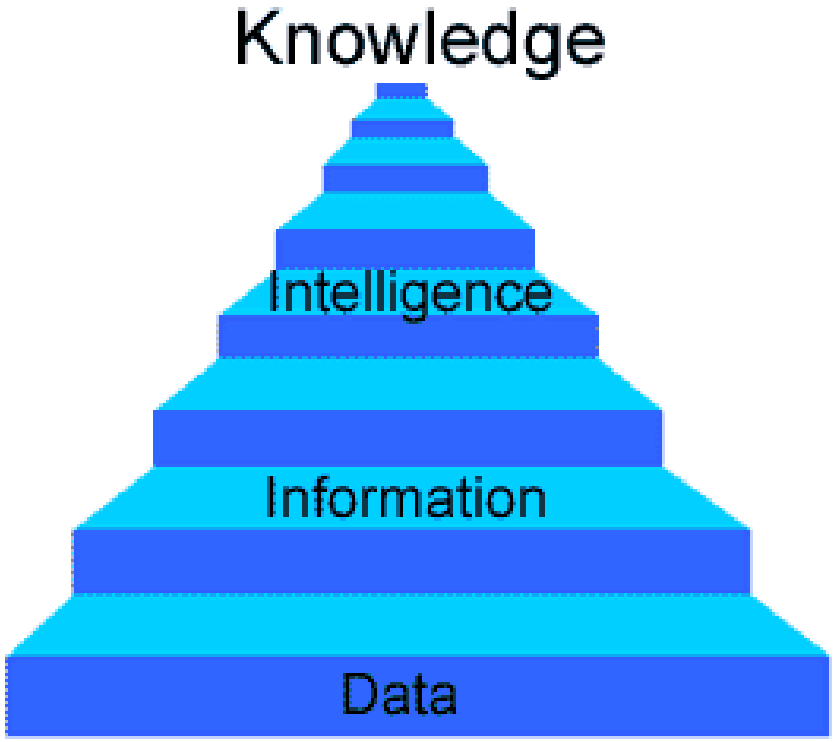
Mesleki Terminoloji II
Veri Madenciliği

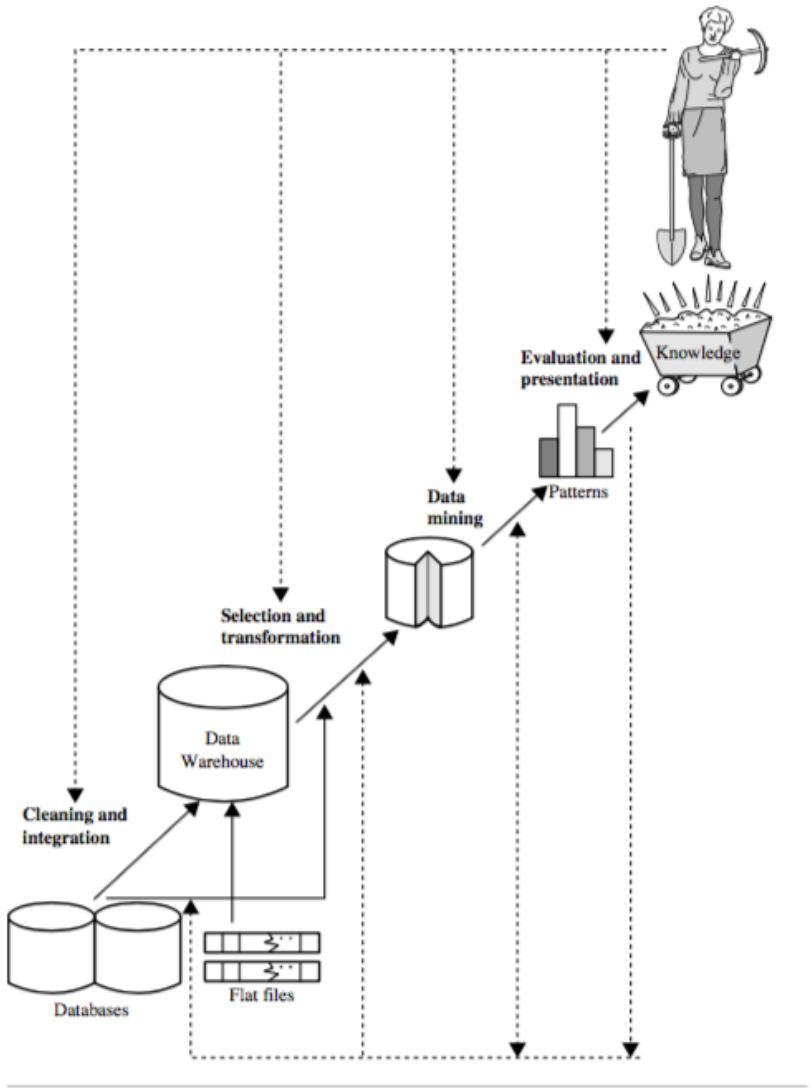
Akif Berkay Gürcan - 14011023
Burak Düşün - 14011055

Veri Madenciliği Nedir?

- Veri madenciliği, büyük miktarda verinin anlamlı örüntüler bulmak amacıyla otomatik veya yarı otomatik olarak işlenmesidir.



Veri Madenciliği Nedir?



Bilgi ortaya çıkarma adımları

- Veri temizleme (Data cleaning)

Gürültü (noise) ve tutarsız verinin silinmesi

- Veri birleştirme (Data integration)

Farklı veri kaynaklarının bir araya getirilmesi

- Veri seçme (Data selection)

Yapılacak analiz için anlamlı olan verinin ayıklanması

- Veri dönüştürme (Data transformation)

Verinin, madencilik işlemlerinin gerçekleştirilmesi için uygun formata getirilmesi

- Veri madenciliği (Data mining)

Akıllı yöntemlerle veriden örüntüler elde edilmesi

- Örüntü değerlendirmesi (Pattern evaluation)

Bulunan örüntüler içinden bilgiyi en iyi ifade edenlerin önem ölçütlerine göre belirlenmesi

- Bilgi sunumu (Knowledge presentation)

Veriden çıkartılan bilginin çeşitli görsel ve bilgi ifade etme yöntemleri kullanılarak sunulması

Veri Madenciliğine Neden İhtiyaç Var?

Çok büyük miktarda (onlarca petabyte) veri üretiliyor ve depolanıyor:

- Global telekomünikasyon ağları trafiği
- Tıp ve sağlık alanında üretilen tıbbi kayıtlar, hasta gözlemlenmesi ve tıbbi görüntüleme
- Şirketlerin ticari kayıtları, banka ve kredi kartı işlemleri
- Arama motorları aracılığıyla milyarlarca internet araması
- Topluluklar ve sosyal medya aracılığıyla oluşan resim, video, blog ve sosyal ağlar

Veriden bilgi elde edilerek veri kullanılabilir hale getirilmelidir.

Geleneksel yöntemler bilgi elde etmek için yetersiz. Analistlerin aynı işi yapması aylar sürebilir. Verinin büyük bir bölümü işlenmiyor.

Karar verilirken tam bilgi yerine sezgi veya eksik bilgi kullanılıyor.

Nasıl Veriler İşlenebilir?

- İlişkisel veritabanı verileri

Müşteri, ürün, satış gibi farklı ama birbiriyle ilişkili varlıkların ve ilişkilerinin saklandığı veritabanı türüdür.

Örnek: Girilen müşteri verilerinden müşterinin kredi uygunluk durumunu tespit etme

- Veri ambarı (data warehouse) verileri

Farklı fiziksel konumlardaki veritabanlarının birleştirilerek tek yerde saklanmasıdır. Veri ambarları oluşturulurken veri temizleme, birleştirme, dönüştürme adımlarından yararlanılır.

Örnek: Mağazanın farklı şube verilerinin tek bir bölgede toplanması. Ardından, mağazanın, dünya genelinde her ürününün şubeler bazında satışlarının belirlenebilmesi

- İşlemsel (transactional) veriler

Her bir işlemin bir satır olarak saklandığı veritabanı türüdür.

Örnek: Satış kayıtlarının incelenerek hangi ürünlerin beraber daha çok satıldığının bulunması

Nasıl Veriler İşlenebilir?

Diğer veri çeşitleri

- Geçici veri (temporal data):

Örnek: Borsa verilerine göre yatırım planı yapılması

Bilgisayar ağı veri akışlarından ağ saldırılarının tespit edilmesi

- Mekansal veri (spatial data):

Şehirlerin anayollardan uzaklığı ile şehrin yoksulluk oranı arasında ilişki bulunması

- Metin verileri:

Veri madenciliğinin son 10 yıldaki literatürünün incelenmesi -> veri madenciliğinin gelişiminde ortaya çıkan popüler konuların belirlenmesi

Ürün yorumlarının incelenmesi -> ürün geliştirilmesine katkı

- Web verileri:

Arama verilerine bağlı olarak trendlerin belirlenmesi

Veri Madenciliğinin Yararlandığı Disiplinler

- İstatistik

İstatistik, bir veriyi ve veri sınıflarını modellemek için kullanılır. Bir sınıfın özellikleri, davranışları; rastgele değişkenler ve onların olasılık dağılımlarına göre tanımlanır.

- Makine Öğrenmesi

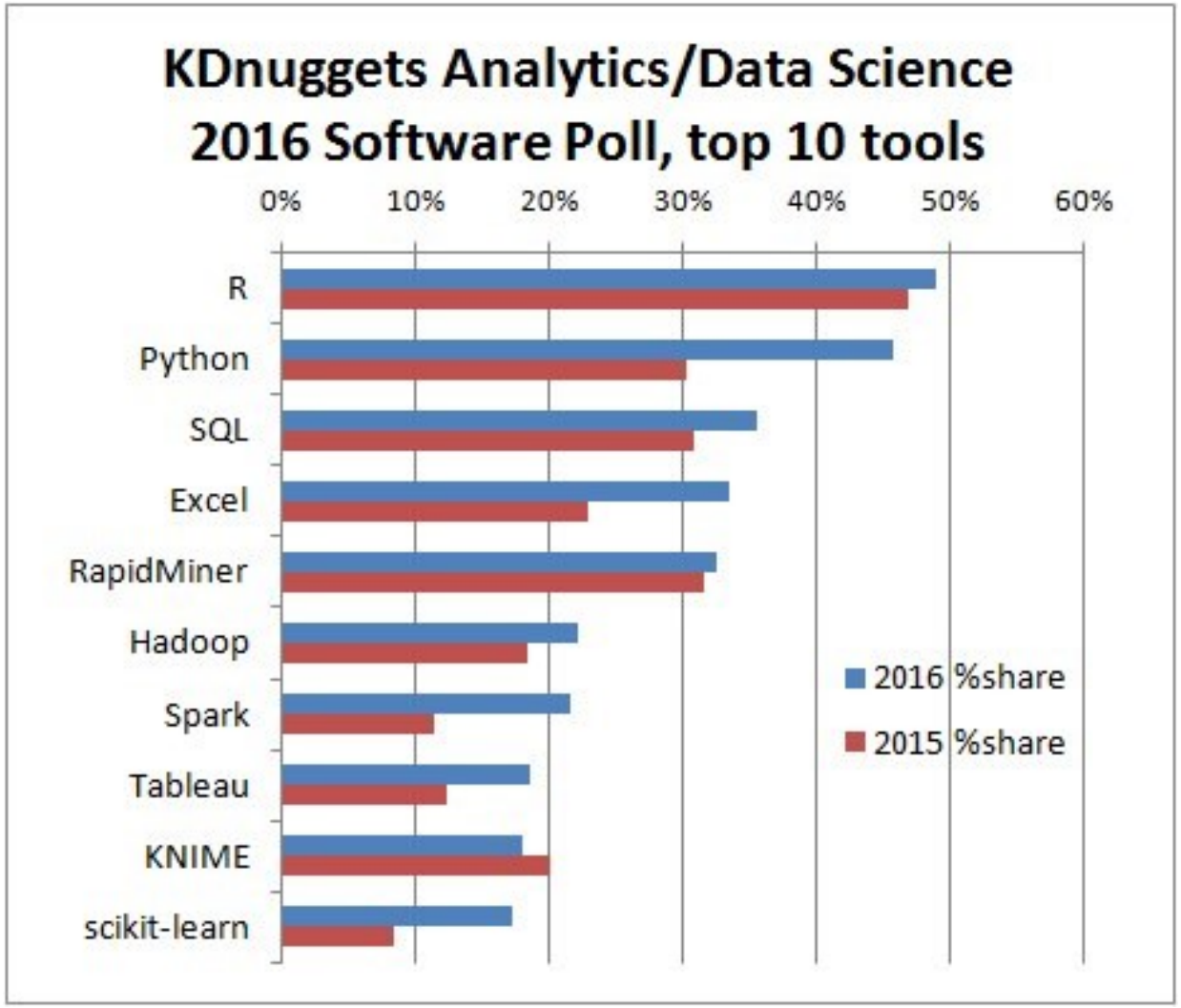
Bilgisayarların verilere göre otomatik olarak örüntüleri öğrenerek akıllı kararlar vermesidir.

- Veri Tabanı ve Ambarı

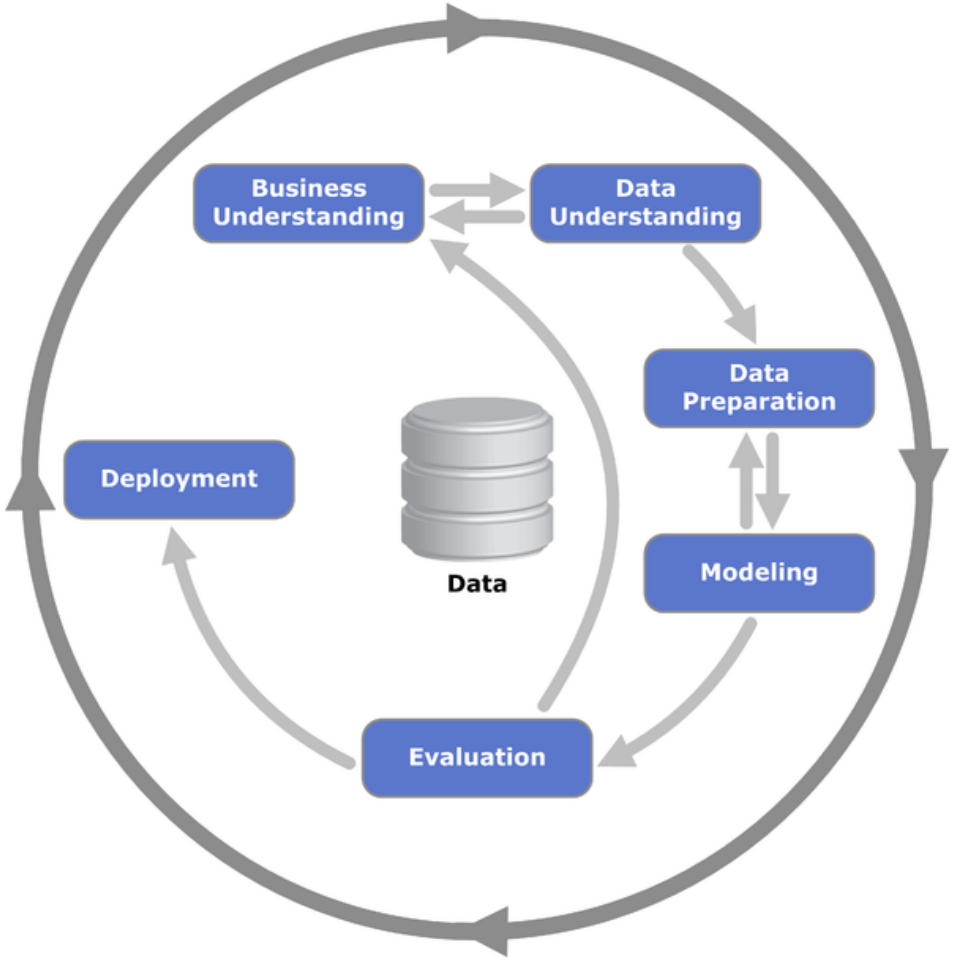
Bilgiler veri tabanından da alınabildiği için veri tabanı işlemleri bilinmelidir. Kimi sistemler anlık ve büyük miktarda veri işlediklerinden veri tabanı optimizasyonu ve ölçeklendirme, etkili sorgu yöntemlerine ihtiyaç duyulur. Veri ambarları sayesinde çok boyutlu veri madenciliği gerçekleştirilebilir.

- Bilgiye Erişim (Information Retrieval)

Bilgiye erişim, bir iş için gerekli dökümanları arama ve gerekli bilgileri bulma bilimidir. Metine dayalı veri madenciliğinde kullanılır.


KDnuggets Analytics/Data Science
2016 Software Poll, top 10 tools
0%
10%
20%
30%
40%
50%
60%
R
Python
SQL
Excel
RapidMiner
Hadoop
Spark
Tableau
KNIME
scikit-learn
2016 %share
2015 %share

Veri Madenciliğinde Standart İş Akışı: **CRISP-DM**



Nitelik türleri

Nitelik (Attribute): Bir veri objesinin karakterini gösteren veri alanı.

- İsimsel (Nominal) Nitelikler

Her bir değer bir kategori, kod veya durumu belirtir. Kategorik olarak da bilinirler.

Örnek: saç_rengi, evlilik_durumu

- İkili (Binary) Nitelikler

İki durumdan birinin olduğunu ifade ederler.

Örnek: sarı_saçlı, mavi_gözlü

- Sırasal (Ordinal) Nitelikler

İsimsel niteliklerin sıralı olarak belirli bir anlama sahip oldukları nitelik türü.

Örnek: menü_boyutu: küçük, orta, büyük

- Sayısal Nitelikler

Nicel olarak ölçülebilen özelliklerdir.

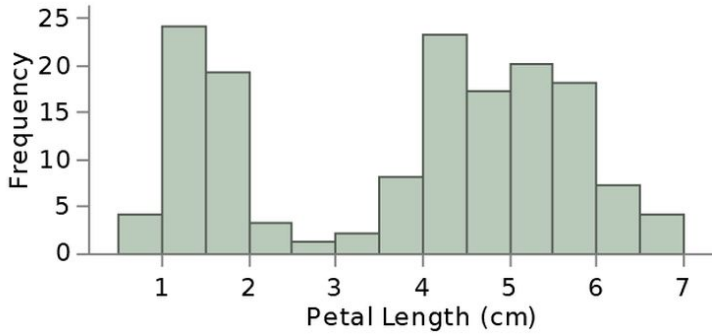
Örnek: Derece, boy, kilo

Ön işleme

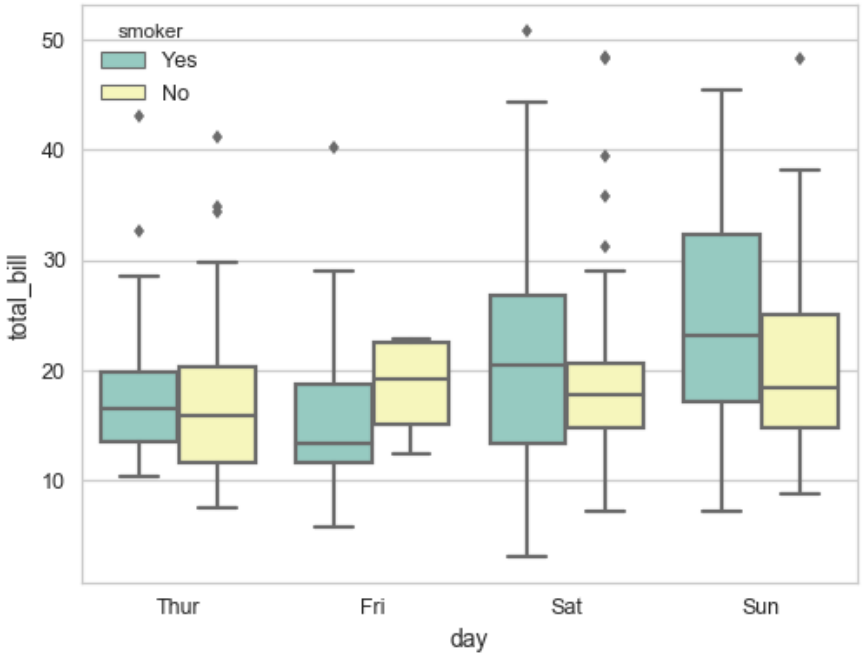
- Örnekleme (sampling)
- Boyut İndirgeme
- Veri Temizleme
- Eksik Veri Sorunu
- Yanlış Sınıflandırmaların Tespiti
- Uçdeğer (outlier) Sorunu
- Özellik Dönüşümü

Veri Görselleştirme

Histogram



Kutu Grafiği



Chernoff Yüzleri

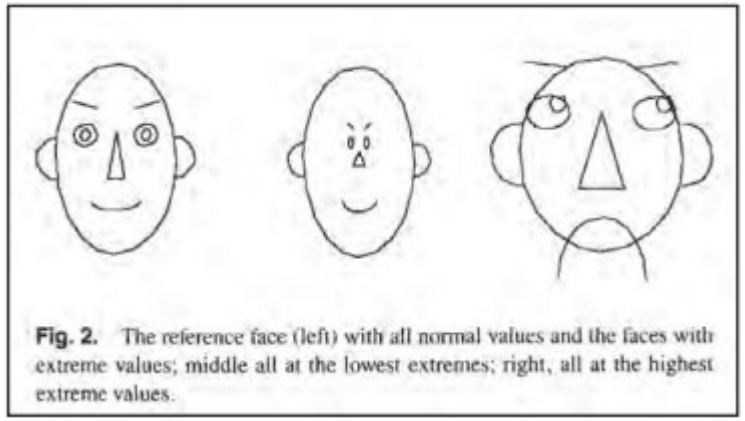


Fig. 2. The reference face (left) with all normal values and the faces with extreme values; middle all at the lowest extremes; right, all at the highest extreme values.

Veri Madenciliğinin Temelleri

- Model Oluşturma
  - İlişki Kuralları ve Ardışıl Örüntüler (Association Rules and Sequential Patterns)
  - Gözetimli Öğrenme
  - Gözetimsiz Öğrenme
  - Yarı-Gözetimli Öğrenme
- Performans Değerlendirmesi
- Topluluk Yöntemleri (Ensemble methods)

İlişki Kuralları ve Ardışıl Örüntüler

İlişki Kuralları (Association Rules)

Veri örnekleri (data instance) arasında birlikte bulunurluk (co-occurrence) durumu aranır.

TID	Items
1	Bread, Coke, Milk
2	Beer, Bread
3	Beer, Coke, Diaper, Milk
4	Beer, Bread, Diaper, Milk
5	Coke, Diaper, Milk

Rules Discovered:
{Milk} --> {Coke}
{Diaper, Milk} --> {Beer}

Ardışıl Örüntüler (Sequential Patterns)

Veri örneklerinin görüldüğü sıra hesaba katılarak n elemanlı altdiziler oluşturulur.

Sequential Patterns with Support ≥ 25%
⟨{30}⟩, ⟨{40}⟩, ⟨{70}⟩, ⟨{80}⟩, ⟨{90}⟩
⟨{30} {40}⟩, ⟨{30} {70}⟩, ⟨{30}, {90}⟩, ⟨{30, 70}⟩, ⟨{30, 80}⟩, ⟨{40, 70}⟩, ⟨{70, 80}⟩
⟨{30} {40, 70}⟩, ⟨{30, 70, 80}⟩

Öğrenme

- Öğrenme, bilgisayarların bu iş için özel olarak programlanmadan sonuç tahmin etmesinde daha başarılı olmasını sağlayan bir yöntemdir.
- Öğrenme için kullanılan veri kümesine "eğitim kümesi" denir.
- Bir model öğrenildikten veya eğitim kümesine dayanılarak bir öğrenme algoritması tarafından oluşturulduktan sonra "test kümesi" kullanılarak modelin başarısı değerlendirilir.

Gözetimli Öğrenme

- Bu tip öğrenme eylemi, insanın geçmiş tecrübelerinden edindiği bilgileri kullanarak iş yapma becerisini geliştirmesine benzetilebilir. Öte yandan, bilgisayarların "tecrübeleri" olmayacağından dolayı, saf veriden öğrenme gerçekleştirilir.
- Gözetimli öğrenmede eğtim aşamasında veri kümesindeki sınıflar bellidir.

Gözetimsiz Öğrenme

- Verinin altında yatan birtakım ilişkilerin tespit edilmesi gerekebilir. Bu ilişkilerin tespiti için kullanılan tipik yöntem kümeleme (clustering) olarak isimlendirilir. Kümeleme, örnek uzayını benzerlik gruplarına böler. Buradaki anafikir aynı gruptaki veri örneklerinin (data instance) birbirlerine benzer olanlardan seçilmesi ve farklı gruplardakilerin birbirlerinden alakasız olmasıdır.
- Kümelemeden sıklıkla gözetimsiz öğrenme şeklinde bahsedilir, zira gözetimli öğrenmenin aksine a priori bir gruplama/sınıflandırma söz konusu değildir.

Yarı-gözetimli Öğrenme

- Eğitimde, veri kümesinin çoğunluğunun etiketsiz verilerden oluştuğu eğitim türüdür. Örneğin "normal" sınıfına ait veriler eğitilir ve test edilecek verilerin normal veya anormal olduğunun tespit edilmesi beklenir.

Performans Değerlendirmesi

- Temel Metrikler

Doğru pozitif, yanlış pozitif, doğru negatif, yanlış negatif, doğruluk, kesinlik, Duyarlılık, F - ölçütü (Accuracy , Precision, Recall , F - Measure)

- Karmaşıklık Matrisi

Bir öğrenme algoritmasının başarısını gösteren bir tablodur. Her satırda sınıfların gerçekte kaç tane oldukları ve sütunlarda bu sınıfın ha ngi sınıf olarak tahmin edildiği bilgisi bulunur.

- Çapraz Doğrulama (Cross-Validation)

K-katlama çapraz doğrulamasında, başlangıç verileri rastgele birbirine yakın alt gruplara ayrılır. İ. adımda, i. kesit test, kalan veriler eğitim verisi olur.

Kaynaklar

- http://blog.euromsg.com/data-mining-veri-madenciligi-nedir/
- https://www-users.cs.umn.edu/~kumar/dmbook/dmslides/chap1_intro.pdf
- Jiawei Han, Micheline Kamber, Jian Pei - Data Mining and Concepts, 3rd Edition
- Bing Liu – Web Data Mining: Exploring Hyperlinks, Contents, and Usage Data, 2nd Edition

Teşekkürler