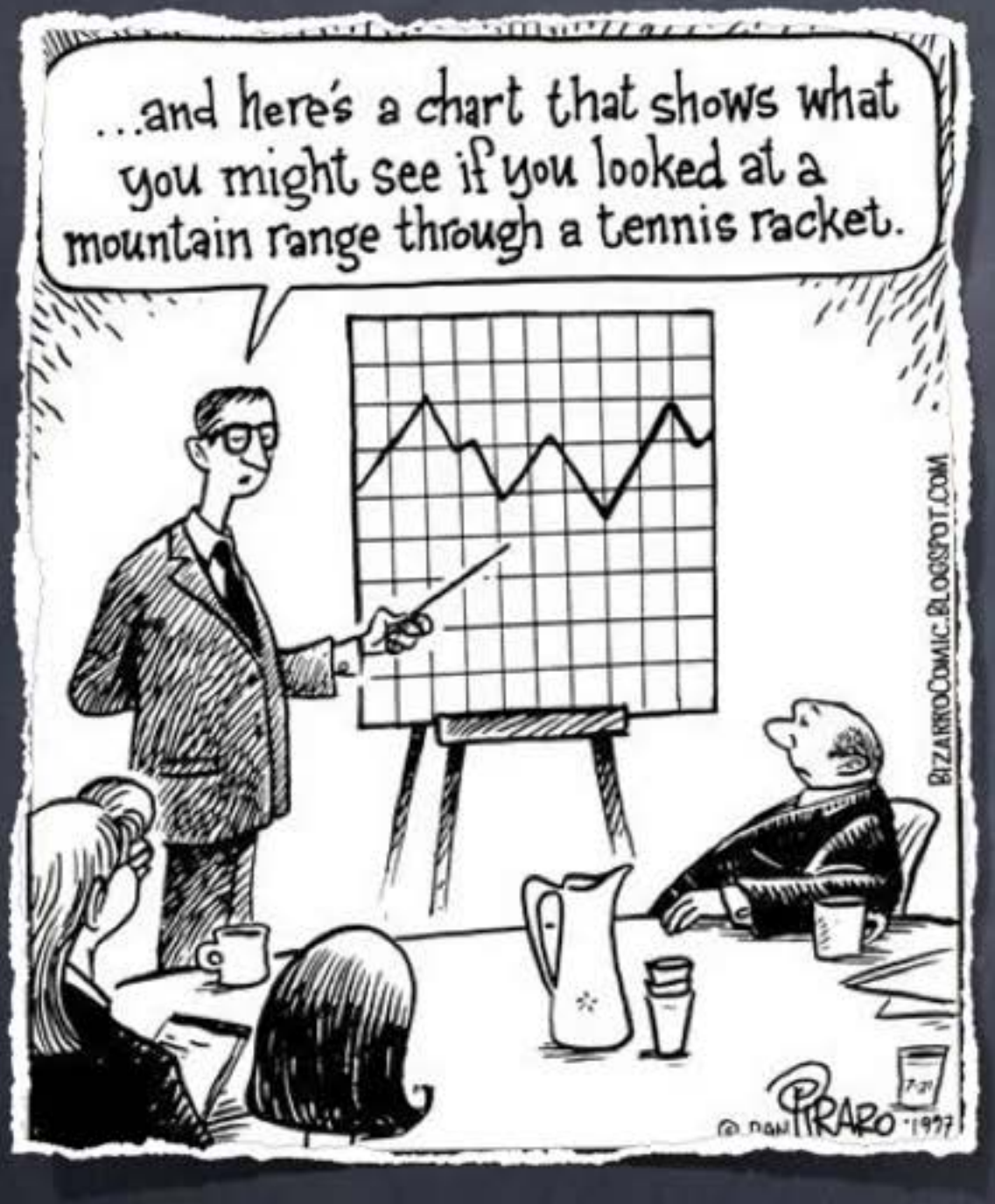


Veri Madenciliği

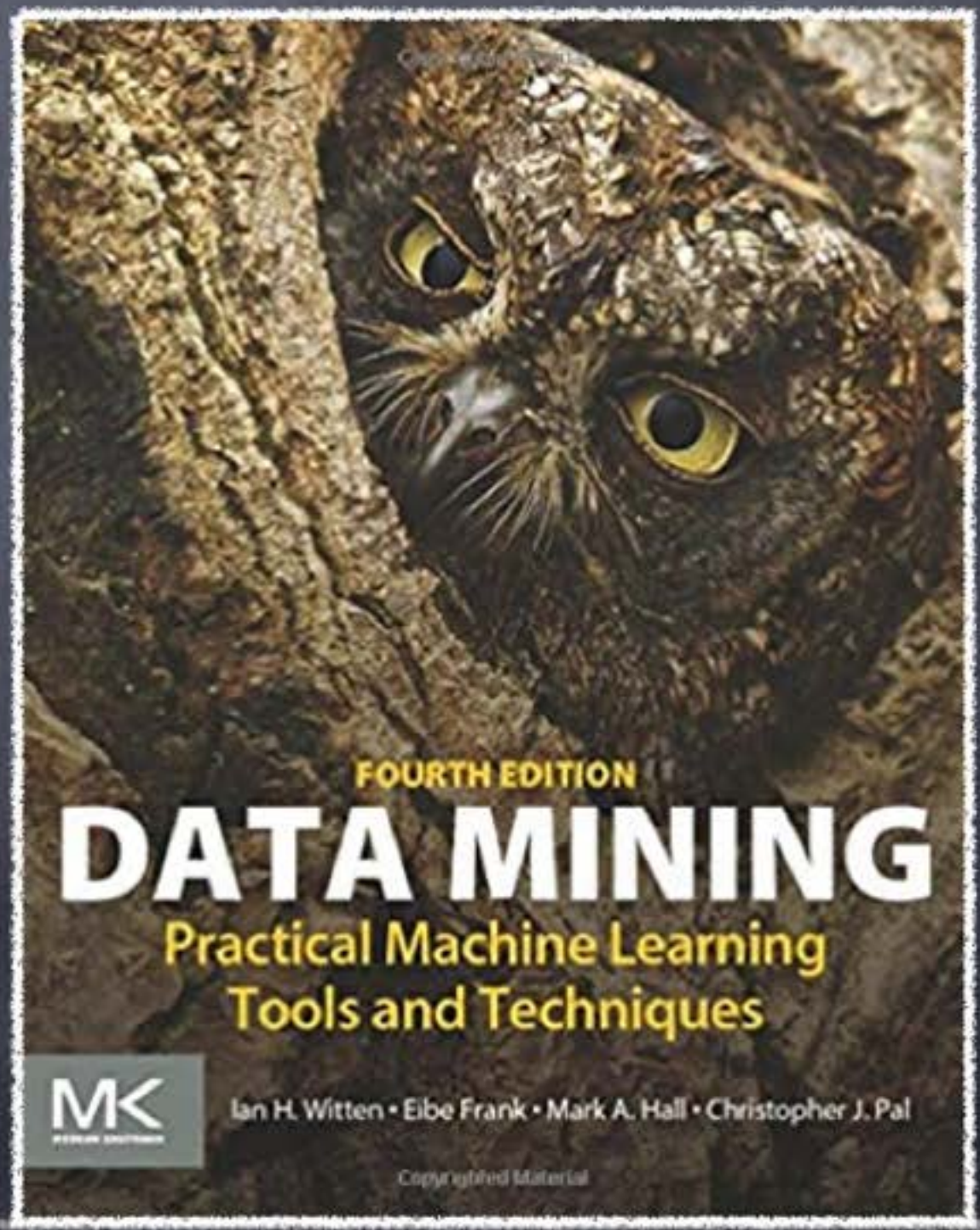
Güz 2023

Ders 4

- Girdi
- Kavramlar
- Örnekler ve
- Nitelikler

Dersin Kitabı

- Data Mining: Practical Machine Learning Tools and Techniques, 4th Ed., by Ian Witten, Eibe Frank, Mark Hall, and Christopher Pal (Morgan Kaufmann Publishers. 2017. ISBN: 978-0-12-804291-5)



Terminoloji

Girdi bileşenleri:

- **Konsept**: Öğrenilecek şey
- **Konsept Açıklama**: Öğrenme planı çıktısı
  Amaç: anlaşılır ve operasyonel konsept açıklaması
- **Örnekler (diğer adıyla demetler)**: bir kavramın bireysel, bağımsız örnekleri
  Not: daha karmaşık girdi biçimleri mümkündür
- **Nitelikler**: Bir örneğin özelliklerini ölçen özellikler

Not: Nominal ve sayısal olanlara odaklanacağız

Konsept nedir?

Konsept: öğrenilmesi gereken şey

Konsept Tanımı: öğrenme planının çıktısı

Öğrenme Stilleri:

- Sınıflandırma Öğrenimi: ayrı bir sınıfı tahmin etme
- İlişkilendirme Öğrenme: özellikler arasındaki ilişkileri algılama
- Kümeleme: benzer örnekleri kümeler halinde gruplama
- Sayısal Tahmin: sayısal bir miktarı tahmin etme

Sınıflandırma Öğrenme

Örnek problemler: hava durumu verileri, kontakt lensler, iris çiçekleri, iş görüşmeleri

Sınıflandırma öğrenimi denetlenir

- program, eğitim örnekleri için gerçek sonuçlarla sağlandığından, başarı değerlendirilebilir.

Sonuç, örneğin sınıfı olarak adlandırılır.

Sınıf etiketlerinin bilindiği yeni verilerdeki başarıyı ölçün (test verileri)

Uygulamada başarı genellikle öznel olarak ölçülür

Bir sınıfa ait örneklere bakıyoruz, çok etiketli sınıflandırma senaryoları vardır

Sayısal Tahmin

- "Sınıf"ın sayısal olduğu sınıflandırma öğrenimi çeşidi ("gerileme (regresyon)" olarak da adlandırılır).
- Öğrenme denetlenir.

Şema hedef değerle sağlanıyor

- Test verileri üzerindeki başarıyı ölçmeliyiz

Görünüm	Sıcaklık	Nem	Rüzgarlı	oyun zamanı
Güneşli	Sıcak	Yüksek	Yanlış	5
Güneşli	Sıcak	Yüksek	Doğru	0
Bulutlu	Sıcak	Yüksek	Yanlış	55
Yağmurlu	Hafif	Normal	Yanlış	40
...	...	...	...	...

Bağlantı Öğrenimi

Herhangi bir sınıf belirtilmemişse ve her türlü yapı "ilginç" kabul edilirse uygulanabilir.

Sınıflandırma öğreniminden farkı:

- Sadece sınıfın değil, herhangi bir özelliğin değerini ve aynı anda birden fazla özelliğin değerini tahmin edebilir
- Bu nedenle, sınıflandırma kurallarından çok daha fazla birliktelik kuralı.
- Bu nedenle, kısıtlamalar gereklidir.
  - Minimum kapsam (örn. %80).
  - Minimum doğruluk (ör. %95)
  - Yalnızca sayısal olmayan niteliklerle kullanılmalı.

Kümeleme

- Benzer öğe gruplarını bulma
- Kümeleme işi denetimsizdir
  Bir örneğin sınıfı bilinmiyordur.
- Başarı genellikle öznel olarak ölçülür
- Yeni örnekler atamasında sonuçları kurallar bulmak için ikinci şemada kullanabiliriz.

İris çiçeği örneği:
Verilen bir sınıf yoksa, 150 örneğin üç türe karşılık gelen doğal kümelere düşmesi muhtemeldir.
* Buradaki zorluk, bu kümelere yeni örnekler atamaktır.

	Çanak yaprak uzunluğu	Sepal çanak yaprağı gelişliği	taç yaprağı uzunluğu	taç yaprağı genişliği	Türleri
1	5.1	3.5	1.4	0.2	tosa
2	4.9	3.0	1.4	0.2	I
… 51	7.0	3.2	4.7	1.4	Iris r
52	6.4	3.2	4.5	1.5	Iris lor
… 101	6.3	3.3	6.0	2.5	Iri
102	5.8	2.7	5.1	1.9	gin
…					

Bir Örnekte Neler Var?

Örnek: belirli bir örnek türü

- Sınıflandırılacak, ilişkilendirilecek veya kümelenecek şey
- Hedef kavramının bireysel, bağımsız örneği
- Önceden belirlenmiş bir dizi özellik ile karakterize edilir

Öğrenme şemasına giriş: örnekler/veri kümesi/ demet kümesi

- Tek bir ilişki/düz dosya olarak temsil edilir

Oldukça kısıtlı girdi biçimi

- Nesneler arasında ilişki yok

Pratik veri madenciliğinde en yaygın biçim

Aile ağacı


Mehmet
E
=
Ayşe
K
Fatma
K
=
Rahmi
E
Sinan
E
Gürkan
E
Pınar
K
=
İlhan
E
Elif
K
Batuhan
E
Aysu
K
Nihal
K

Tablo Olarak Temsil Edilen Aile Ağacı

İsim	Cinsiyet	Ebeveyn 1	Ebeveyn 2
Mehmet	Erkek	?	?
Ayşe	Kadın	?	?
Fatma	Kadın	?	?
Rahmi	Erkek	?	?
Sinan	Erkek	Mehmet	Ayşe
Gürkan	Erkek	Mehmet	Ayşe
Pınar	Kadın	Mehmet	Ayşe
İlhan	Erkek	Fatma	Rahmi
Elif	Kadın	Fatma	Rahmi
Batuhan	Erkek	Fatma	Rahmi
Aysu	Kadın	Pınar	İlhan
Nihal	Kadın	Pınar	İlhan

"Kardeş" İlişkisi

Bu iki tablo kardeşliği biraz farklı bir şekilde temsil ediyor.

144 çift insan

Birinci şahıs	İkinci şahıs	Kardeş?
Mehmet	Ayşe	Hayır
Mehmet	Sinan	Hayır
...	...	...
Sinan	Mehmet	Hayır
Sinan	Graham	Hayır
Sinan	Pınar	Evet
...	...	...
İlhan	Elif	Evet
...	...	...
Aysu	Nihal	Evet
...	...	...
Nihal	Aysu	Evet

Yalnızca pozitifler tanımlanmış

Birinci şahıs	İkinci şahıs	Kardeş?
Sinan	Pınar	Evet
Graham	Pınar	Evet
İlhan	Elif	Evet
Batuhan	Elif	Evet
Aysu	Nihal	Evet
Nihal	Aysu	Evet
Geriye kalan		Hayır

Kapalı dünya varsayımı

Her zaman gerçek dünyayla eşleşmez.

Tek Tabloda Tam Bir Gösterim

Düzleştirme diğer adıyla Denormalize etme:

Önceki iki tabloyu bire indirgediğimizde orijinal "ilişkileri" örnek forma dönüştürülmüş olur.

Birinci şahıs				İkinci şahıs				Kardeş?
İsim	Cinsiyet	Ebeveyn 1	Ebeveyn2	İsim	Cinsiyet	Ebeveyn 1	Ebeveyn 2	
Sinan	Erkek	Mehmet	Ayşe	Pınar	Kadın	Mehmet	Ayşe	Evet
Graham	Erkek	Mehmet	Ayşe	Pınar	Kadın	Mehmet	Ayşe	Evet
İlhan	Erkek	Fatma	Rahmi	Elif	Kadın	Fatma	Rahmi	Evet
Batuhan	Erkek	Fatma	Rahmi	Elif	Kadın	Fatma	Rahmi	Evet
Aysu	Kadın	Pınar	İlhan	Nihal	Kadın	Pınar	İlhan	Evet
Nihal	Kadın	Pınar	İlhan	Aysu	Kadın	Pınar	İlhan	Evet
Geriye kalanlar								Hayır

If ikinci kişinin cinsiyeti == kadın
ve birinci kişinin ebeveyni == ikinci kişinin ebeveyni ise
then Kardeş = evet

Düz Dosya Oluşturma

- "Denormalizasyon" adı verilen düzleştirme süreci
  Bir tane yapmak için birkaç ilişki birleştirilir
- Herhangi bir sonlu - sonlu ilişki kümesiyle mümkün
- Problem: önceden belirlenmiş sayıda nesne olmayan ilişkiler
  Örnek: çekirdek aile kavramı
- Denormalizasyon, veritabanının yapısını yansıtan sahte düzenlilikler üretebilir.
  Örnek: "tedarikçi", "tedarikçi adresini" tahmin eder

Müşteriler ürünleri satın alır, DB'yi düzleştirerek her bir örneği üretir: müşteri, ürün, tedarikçi, tedarikçi adresi.

Süpermarket yöneticisi, her müşterinin satın aldığı ürünlerin kombinasyonlarını önemseyebilir, ancak tedarikçilerin adresinin "keşfi" ile ilgilenmeyebilir.

"Ata" İlişkisi

Birinci şahıs				İkinci şahıs				"Ata" İlişkisi
İsim	Cinsiyet	Ebeveyn 1	Ebeveyn 2	İsim	Cinsiyet	Ebeveyn 1	Ebeveyn 2	
Mehmet	Erkek	?	?	Sinan	Erkek	Mehmet	Ayşe	Evet
Mehmet	Erkek	?	?	Pınar	Kadın	Mehmet	Ayşe	Evet
Mehmet	Erkek	?	?	Aysu	Kadın	Pınar	İlhan	Evet
Mehmet	Erkek	?	?	Nihal	Kadın	Pınar	İlhan	Evet
Pınar	Kadın	Mehmet	Ayşe	Nihal	Kadın	Pınar	İlhan	Evet
Fatma	Kadın	?	?	İlhan	Erkek	Fatma	Rahmi	Evet
Fatma	Kadın	?	?	Nihal	Kadın	Pınar	İlhan	Evet
"Buradaki diğer olumlu örnekler Tüm kalan"								Evet
								Hayır

Recursion/özyineleme

Bu genel ilişkiler ders kitabının ve bu dersin kapsamı dışındadır.

Sonsuz ilişkiler özyineleme gerektirir

Bu tanım, iki insan ne kadar uzaktan akraba olursa olsun işe yarar.

```
If şahıs1, şahıs2'nin ebeveyni ise
   then şahıs1, şahıs2'nin atasıdır

If şahıs1, şahıs2'nin ebeveyni
   ve şahıs2, şahıs3'ün atası ise
   then şahıs1, şahıs3'ün atasıdır
```

Uygun teknikler "endüktif mantık programlaması" olarak bilinir.

(örneğin Quinlan'ın Birinci Dereceli Tümevarımlı Öğrenicisi, (FOIL), kural tabanlı bir öğrenme algoritmasıdır)

Problemler: (a) gürültü ile iyi ilgilenmiyor ve (b) hesaplama karmaşıklığı, yani büyük veri kümeleri yavaş.

Çoklu-Örnek Kavramları

Her bireysel örnek, bir dizi örnek içerir

- Aynı nitelikler tüm örnekleri tanımlar

- Bir örnek içindeki bir veya daha fazla örnek, sınıflandırmasından sorumlu olabilir.

Öğrenmenin amacı hala bir kavram tanımı üretmektir.

Önemli gerçek saha uygulamaları var

- örneğin, farklı formlar alan ilaç molekülü şekilleri, pozitif veya negatif bağlanma aktivitesini öngören bir kümedir.

Tüm set, pozitif veya negatif olarak sınıflandırılır.

Bir Nitelikte Neler Vardır?

Her bir örnek, önceden tanımlanmış sabit bir dizi özellik, "nitelikleri" ile tanımlanır.

Ancak: nitelik sayısı pratikte değişebilir
Olası çözüm: "alakasız değer" işareti

İlgili problem: Bir özelliğin varlığı, diğerinin değerine bağlı olabilir.

Olası nitelik türleri ("ölçüm seviyeleri"):
İstatistikçiler genellikle nominal, sıralı, aralık ve oran kullanır.

Nominal diğer adıyla kategorik;
Sayısal diğer adıyla sürekli

Nominal Nicelikler

- Değerler farklı sembollerdir
  - Değerlerin kendileri yalnızca etiket veya ad işlevi görür
  - Nominal Latince isim kelimesinden gelir
- Örnek: hava durumu verilerinden görünüm özelliği
  - Değerler: güneşli, bulutlu ve yağmurlu
- Nominal değerler arasında hiçbir ilişki ima edilmez (sıralama veya mesafe ölçümü yok)
- Sadece eşitlik testleri yapılabilir

Sıralı Nicelikler

Değerlere düzen empoze edin
Amaç: Hayır tanımlanan değerler arasındaki uzaklık

Örnek: hava durumu verilerinde öznitelik sıcaklığı
Değerler: sıcak > ılık > serin
Not: toplama ve çıkarma mantıklı değildir

Örnek kural:
sıcaklık < çok yüksek ==> oyun = Evet

Nominal ve sıralı arasındaki ayrım gözlemle her zaman net değildir (Örneğin, öznitelik görünümü güneşli ve yağmurlu arasında bulutlu mu?)

Aralıklı Nicelikler

Aralıklı nicelikler sadece sıralanmakla kalmaz, aynı zamanda sabit ve eşit birimlerde ölçülür.

Örnek 1: Fahrenheit derece olarak ifade edilen öznitelik sıcaklığı
Örnek 2: nitelik yılı

- İki değerin (aynı özelliğe sahip) farkı anlamlıdır
- Toplama veya çarpım mantıklı değildir
- Sıfır noktası tanımlı değil!

Oran Nicelikleri

- Oran miktarları, ölçüm şemasının bir sıfır noktası tanımladığı miktarlardır.
  Örnek: nitelik mesafesi
  Bir nesne ile kendisi arasındaki mesafe sıfırdır

- Oran miktarları gerçek sayılar olarak kabul edilir
  Tüm matematiksel işlemlere izin verilir

- Ancak: "doğal olarak" tanımlanmış bir sıfır noktası var mı?
  Çevap bilimsel bilgiye bağlıdır
  Örneğin. Daniel Fahrenheit, sıcaklık için daha düşük bir sınır bilmiyordu, ancak bugün ölçek mutlak sıfıra dayanıyor.
  Örneğin. M.Ö. 0'da kültürel olarak tanımlanan sıfırdan bu yana geçen zamanın ölçümü bir oran değil, Big Bang'den bu yana geçen yıllar 0'dır.

Uygulamada Kullanılan Nitelik Türleri

- Çoğu şema sadece iki ölçüm seviyesi barındırır: nominal ve sıralı
- Nominal nitelikler ayrıca kategorik, numaralandırılmış veya ayrık olarak adlandırılır.
  - Ama ne yazık ki, numaralandırılmış ve ayrık -> sıralı (düzenli) anlamına gelir
- Özel durum: dikotomi/ikiye ayrılma (boole niteliği)
- Sıralı nitelikler sayısal veya sürekli olarak adlandırılır.
  - Ama ne yazık ki, sürekli -> matematiksel süreklilik anlamına gelir

Metaveri *"Verilerle ilgili veriler"*

Meta veriler, arka plan bilgisini kodlayan veriler hakkındaki bilgilerdir.

Arama alanını kısıtlamak için kullanılabilir
Örnekler:

- Boyutsal hususlar (yani aramayı, boyutsal olarak doğru olan ifade veya karşılaştırmalarla sınırlandırın)
- Dairesel sıralamalar test türlerini etkileyebilir, örn. pusulada dereceler; örn. gün özelliği, sonraki gün, önceki gün, sonraki haftanın günü vb. kullanabilir.
- Kısmi sıralamalar