Data- and intelligence-driven enterprises win*

Veri Madenciliği

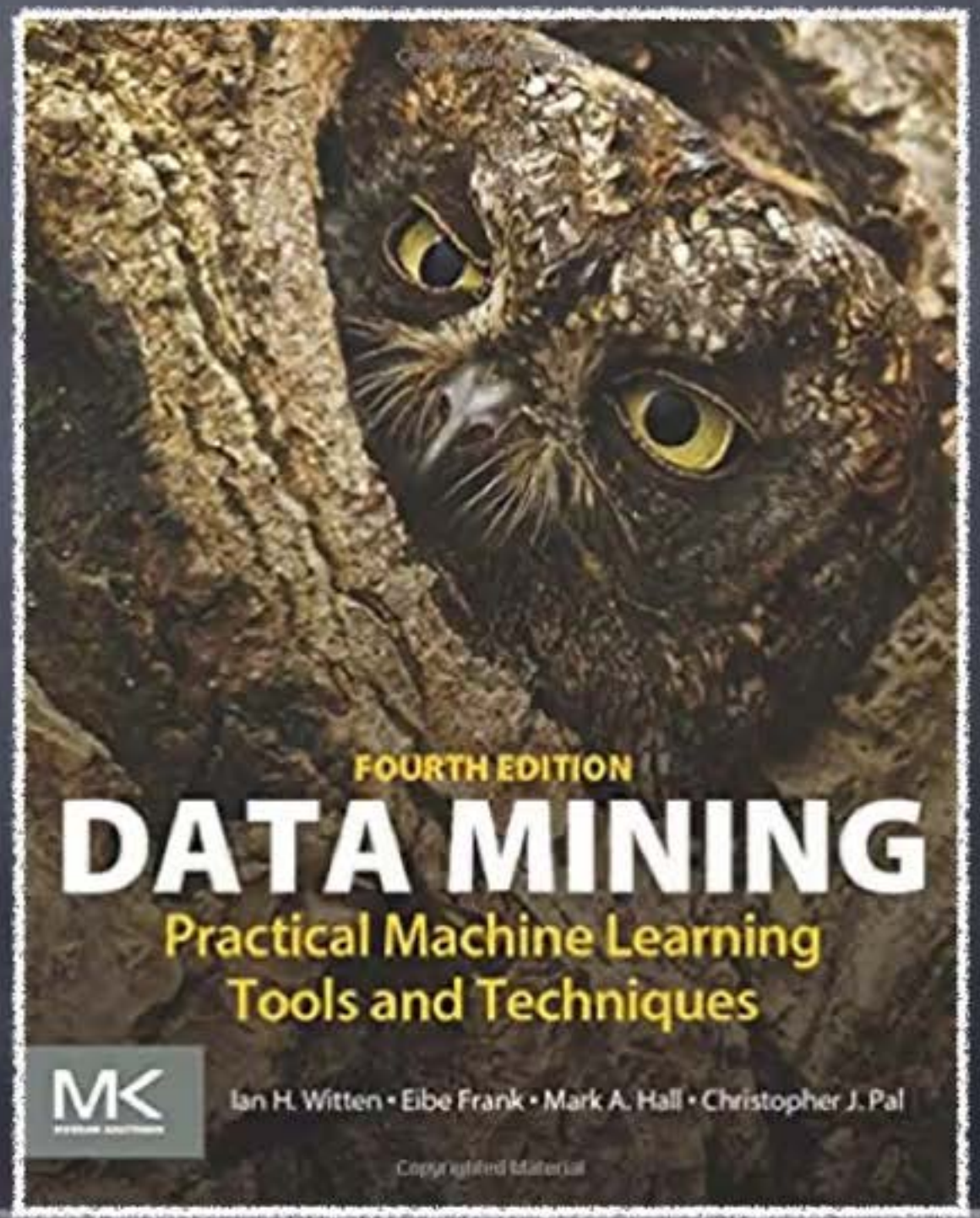
2023-2024 Güz

Ders 2

Desenleri Tanımlamak
Basit Örnekler

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)



Veriye karşı Bilgi

Toplum çok büyük miktarda veri üretir

Kaynaklar: işletme, bilim, tıp, ekonomi, coğrafya, çevre, spor, ...

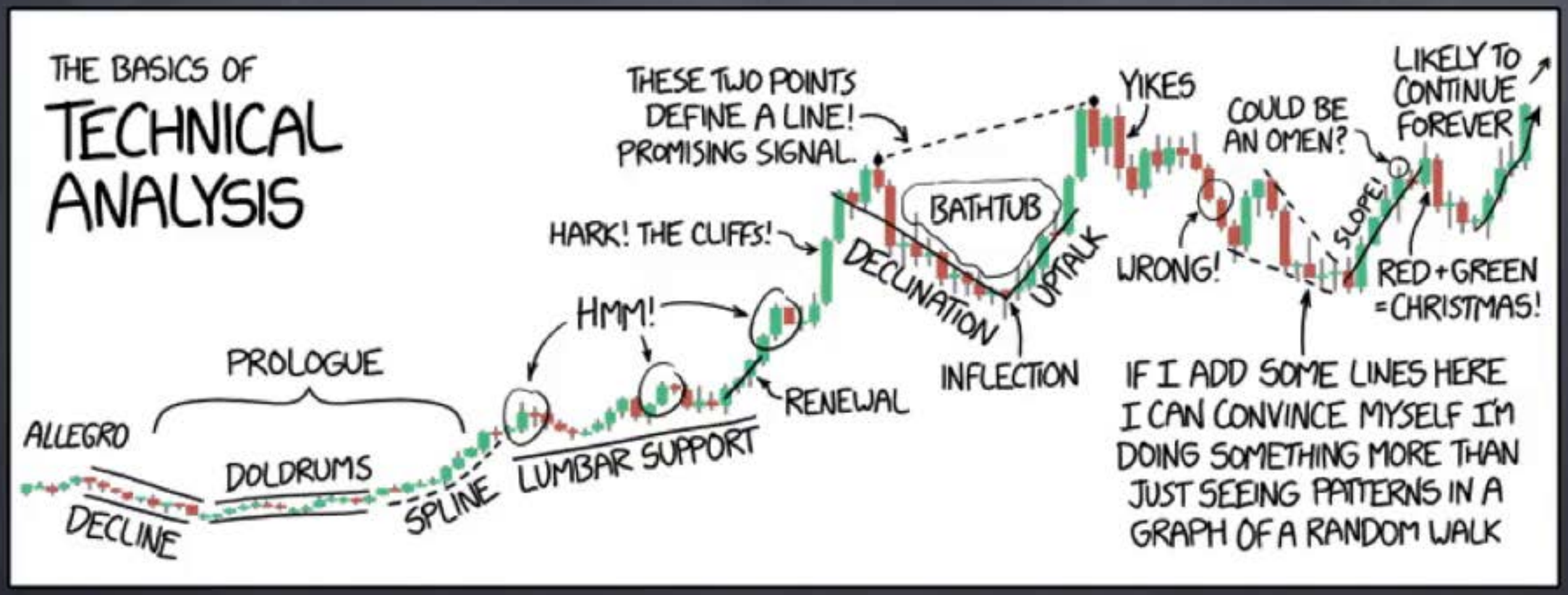
Potansiyel olarak değerli kaynak

Ham veriler işe yaramaz: ondan otomatik olarak bilgi çıkarmak için tekniklere ihtiyaç Vardır

- Veri: kaydedilen Doğrular
- Bilgi: verilerin altında yatan kalıplar

Rastgele gürültüden desen oluşturmamaya dikkat edin: https://xkcd.com/2101/

Veriye karşı Bilgi



Bilgi Çok Önemlidir

Örnek 1: suni dölleme

- Verilen: 60 özellik ile tanımlanan embriyolar
- Sorun: hayatta kalacak embriyoları seçin
- Veriler: embriyoların ve sonuçların tarihsel kayıtları

Örnek 2: inek itlaf

- Verilen: 700 özellik ile tanımlanan inekler
- Sorun: itlaf edilecek inekleri seçin
- Veriler: tarihsel kayıtlar ve çiftçilerin kararları

Veri madenciliği

Ayıklama
- örtük,
- önceden bilinmeyen,
- potansiyel olarak yararlı

verilerden bilgi

Gerekli: verilerdeki kalıpları ve düzenlilikleri algılayan programlar

Güçlü kalıplar ==> iyi tahminler
- Sorun 1: çoğu desen ilginç değil
- Sorun 2: desenler tam olmayabilir (veya sahte olabilir)
- Sorun 3: veriler bozuk veya eksik olabilir

Kara Kutu ve Yapısal Tanımlamalar

Makine öğrenimi farklı türde modeller üretebilir:

Kara Kutu tanımlaması:

- Yeni durumda sonucu tahmin etmek için kullanılabilir
- Tahminin nasıl yapıldığı konusu anlaşılmaz
- Nasıl tahminde bulunduklarını incelemek için kullanışlı değildir.

Kara Kutu ve Yapısal Tanımlamalar

Yapısal Tanımlamalar:

- Kalıpları açıkça temsil edin (örneğin, bir dizi kural veya bir karar ağacı ile).
- Yeni durumda sonucu tahmin etmek için kullanılabilir
- Tahminin nasıl elde edildiğini anlamak ve açıklamak için kullanılabilir (daha da önemli olabilir)

Yöntemler yapay zekadan, istatistiklerden ve veri tabanlarındaki araştırmalardan kaynaklanır.

Yapısal Tanımlamalar

Örnek: if-then kuralları (kontakt lens verilerinden)

If gözyaşı üretim hızı = azalmış
  then öneri =Yok
Aksi takdirde, if yaş = genç and astigmat =Yok
  then öneri =Yumuşak

Yaş	Gözlük reçetesi	Astigmat	Gözyaşı üretim hızı	Önerilen lensler
Genç	Miyop	Yok	Azalmış	Yok
Genç	Hipermetrop	Yok	Normal	Yumuşak
Presbiyopik (40 - 60 arası)	Hipermetrop	Yok	Azalmış	Yok
Presbiyopik (40 - 60 arası)	Miyop	Var	Normal	Sert
...	...	...	...	...

Hava Durumu Problemi: Basit bir örnek

Belirli bir oyunu oynamak için koşullar

Görünüm	Sıcaklık	Nem	Rüzgarlı	Oyun
Güneşli	Sıcak	Yüksek	Yanlış	Hayır
Güneşli	Sıcak	Yüksek	Doğru	Hayır
Bulutlu	Sıcak	Yüksek	Yanlış	Evet
Yağmurlu	Hafif	Normal	Yanlış	Evet
...	...	...	...	...

If Görünüm = güneşli and nem = yüksek then oyun = hayır
If Görünüm = yağmurlu and rüzgarlı = Doğru then oyun = hayır
If Görünüm = bulutlu then oyun = evet
If Nem = normal then oyun = evet
If YukarıdakilerinHiçbiri değilse then oyun = evet

Bu kurallar sırayla kontrol edilmelidir, aksi takdirde yanlış sınıflandırılacaktır.

Sınıflandırmaya Karşı Birliktelik Kuralları

Sınıflandırma kuralı:

belirli bir özelliğin değerini tahmin eder (bir örneğin sınıflandırması)

```
If Görünüm = güneşli and nem = yüksek
   then oyun = hayır
```

Birliktelik kuralı:

keyfi özelliğin (veya kombinasyonun) değerini tahmin eder

```
If Sıcaklık = serin       then nem = normal
If Nem = normal           and rüzgarlı = yanlış
                          then oyun = evet
If Görünüm = güneşli      and oyun = hayır
                          then nem = yüksek
If Rüzgarlı = yanlış      and oyun = hayır
                          then görünüm = güneşli
                          and nem = yüksek
```

Karışık Özelliklere Sahip Hava Durumu Verileri

Bazı niteliklerin sayısal değerleri Vardır

Görünüm	Sıcaklık	Nem	Rüzgarlı	Oyun
Güneşli	85	85	Yanlış	Hayır
Güneşli	80	90	Doğru	Hayır
Bulutlu	83	86	Yanlış	Evet
Yağmurlu	75	80	Yanlış	Evet
...	...	...	...	...

```
If Görünüm = güneşli and nem > 83
      then oyun = hayır
If Görünüm = yağmurlu and rüzgarlı = doğru
      then oyun = hayır
If Görünüm = bulutlu
      then oyun = evet
If Nem < 85
      then oyun = evet
If YukarıdakilerdenHiçbiriYoksa
      then oyun = evet
```

Kontakt Lens Verileri

Yaş	Gözlük reçetesi	Astigmat	Gözyaşı üretim hızı	Önerilen lensler
Genç	miyop	Yok	Azalmış	Hiçbiri
Genç	miyop	Yok	Normal	Yumuşak
Genç	miyop	Var	Azalmış	Hiçbiri
Genç	miyop	Var	Normal	Sert
Genç	hipermetrop	Yok	Azalmış	Hiçbiri
Genç	hipermetrop	Yok	Normal	Yumuşak
Genç	hipermetrop	Var	Azalmış	Hiçbiri
Genç	hipermetrop	Var	Normal	Sert
Presbiyopik öncesi	miyop	Yok	Azalmış	Hiçbiri
Presbiyopik öncesi	miyop	Yok	Normal	Yumuşak
Presbiyopik öncesi	miyop	Var	Azalmış	Hiçbiri
Presbiyopik öncesi	miyop	Var	Normal	Sert
Presbiyopik öncesi	hipermetrop	Yok	Azalmış	Hiçbiri
Presbiyopik öncesi	hipermetrop	Yok	Normal	Yumuşak
Presbiyopik öncesi	hipermetrop	Var	Azalmış	Hiçbiri
Presbiyopik öncesi	hipermetrop	Var	Normal	Hiçbiri
presbiyopik	miyop	Yok	Azalmış	Hiçbiri
presbiyopik	miyop	Yok	Normal	Hiçbiri
presbiyopik	miyop	Var	Azalmış	Hiçbiri
presbiyopik	miyop	Var	Normal	Sert
presbiyopik	hipermetrop	Yok	Azalmış	Hiçbiri
presbiyopik	hipermetrop	Yok	Normal	Yumuşak
presbiyopik	hipermetrop	Var	Azalmış	Hiçbiri
presbiyopik	hipermetrop	Var	Normal	Hiçbiri

Kontakt Lens Verileri Tamamlandı

Özellikler:

- Yaş: genç, presbiyopik öncesi, presbiyopik
- Reçete: Miyop, Hipermetrop
- Astigmatizm: Evet veya Hayır
- Gözyaşı Üretimi: Azalmış, Normal

Özellik değerlerinin tüm olası kombinasyonları temsil edilir.

Soru: Bu kaç örnektir?

Not: Gerçek girdi kümeleri genellikle tamamlanmaz. Eksik değerlere sahip olabilirler veya tüm kombinasyonlar mevcut olmayabilir.

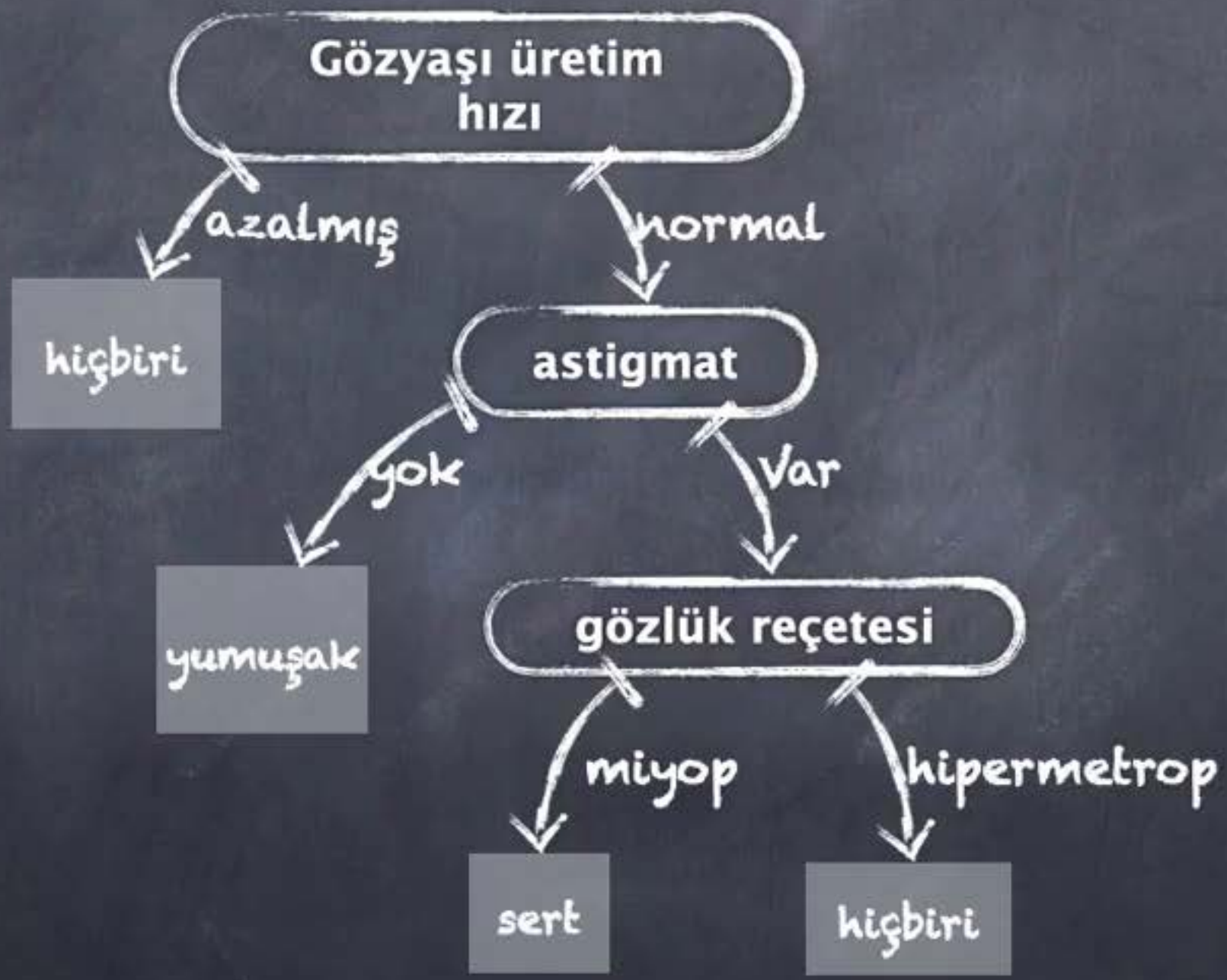
Eksiksiz ve Doğru Bir Kural Kümesi

If Gözyaşı üretim hızı = azalmış then öneri =Yok
If Yaş = genç and astigmat = hayır
  and gözyaşı üretim hızı = normal then öneri =Yumuşak
If Yaş = presbiyopik öncesi ve astigmat = hayır
  and gözyaşı üretim hızı = normal then öneri =Yumuşak
If Yaş = presbiyop and gözlük reçetesi = miyop
  and astigmat = hayır then öneri =Yok
If Gözlük reçetesi = hipermetrop and astigmat = hayır
  and gözyaşı üretim hızı = normal then öneri =Yumuşak
If Gözlük reçetesi = miyop and astigmat = evet
  and gözyaşı üretim hızı = normal then öneri = Sert
If Yaş genç and astigmat = evet
  and gözyaşı üretim hızı = normal then öneri = Sert
If Yaş = presbiyop öncesi
  and gözlük reçetesi = hipermetrop
  and astigmat = evet then öneri =Yok
If Yaş = presbiyopik ve gözlük reçetesi = hipermetrop
  and astigmat = evet then öneri =Yok

Gerçek hayatta, sınıflandırıcı her zaman doğru sınıfı üretmeyebilir.

Bu büyük bir kurallar dizisidir. Daha küçük bir set daha mı iyi olur?

Aynı Problem İçin Bir Karar Ağacı



Yaş	Reçete	Astigmat	Gözyaşı Hızı	Tavsiye
8 Genç	Hipermetrop	Var	Normal	Sert
18 Presbiyopik	Miyop	Yok	Normal	Yok

Burada hem nitelikler hem de sonuç nominaldir

İris Çiçeklerinin Sınıflandırılması

Bu ünlü veri setinin kuralları hantaldır ve sınıflandırmanın daha iyi bir yolu olabilir. Burada sayısal nitelikler olduğunu, ancak sonucun bir kategori olduğunu unutmayın.

	Çanak yaprağı uzunluğu	Çanak yaprağı -genişlik	Taç yaprağı uzunluğu	Taç yaprağı genişliği	Tip
1	5.1	3.5	1.4	0.2	Iris setosa
2	4.9	3.0	1.4	0.2	Iris setosa
...					
51	7.0	3.2	4.7	1.4	Iris versicolor
52	6.4	3.2	4.5	1.5	Iris versicolor
...					
101	6.3	3.3	6.0	2.5	Iris virginica
102	5.8	2.7	5.1	1.9	Iris virginica
...					

setosa

versicolor

virginica

If Taç yaprağı uzunluğu < 2,45 then İris-setosa
If Çanak yaprağı genişliği < 2.10 then İris-versicolor
If Çanak yaprağı genişliği < 2,45 and Taç yaprağı uzunluğu < 4,55
then Iris-versicolor

....

CPU Performansını Tahmin Etme

Örnek: 209 farklı bilgisayar konfigürasyonunun örneğidir.

	Döngü süresi (ns)	Ana bellek (Kb)		Önbellek (Kb)	Kanallar		Performans
	MYCT	MMIN	MMAX	CACH	CHMIN	CHMAX	PRP
1	125	256	6000	256	16	128	198
2	29	8000	32000	32	8	32	269
...							
208	480	512	8000	32	0	0	67
209	480	1000	4000	0	0	0	45

Bu durumda hem nitelikler hem de sonuç sayısaldır.

Lineer regresyon fonksiyonu

$$PRP = -55.9 + 0.0489\,MYCT + 0.0153\,MMIN + 0.0056\,MMAX + 0.6410\,CACH - 0.2700\,CHMIN + 1.480\,CHMAX$$

İş Görüşmelerinden elde edilen veriler

Burada özellikler normal sütunlar yerine satırlardadır.
Bu durumda örnekler sütunlardadır.

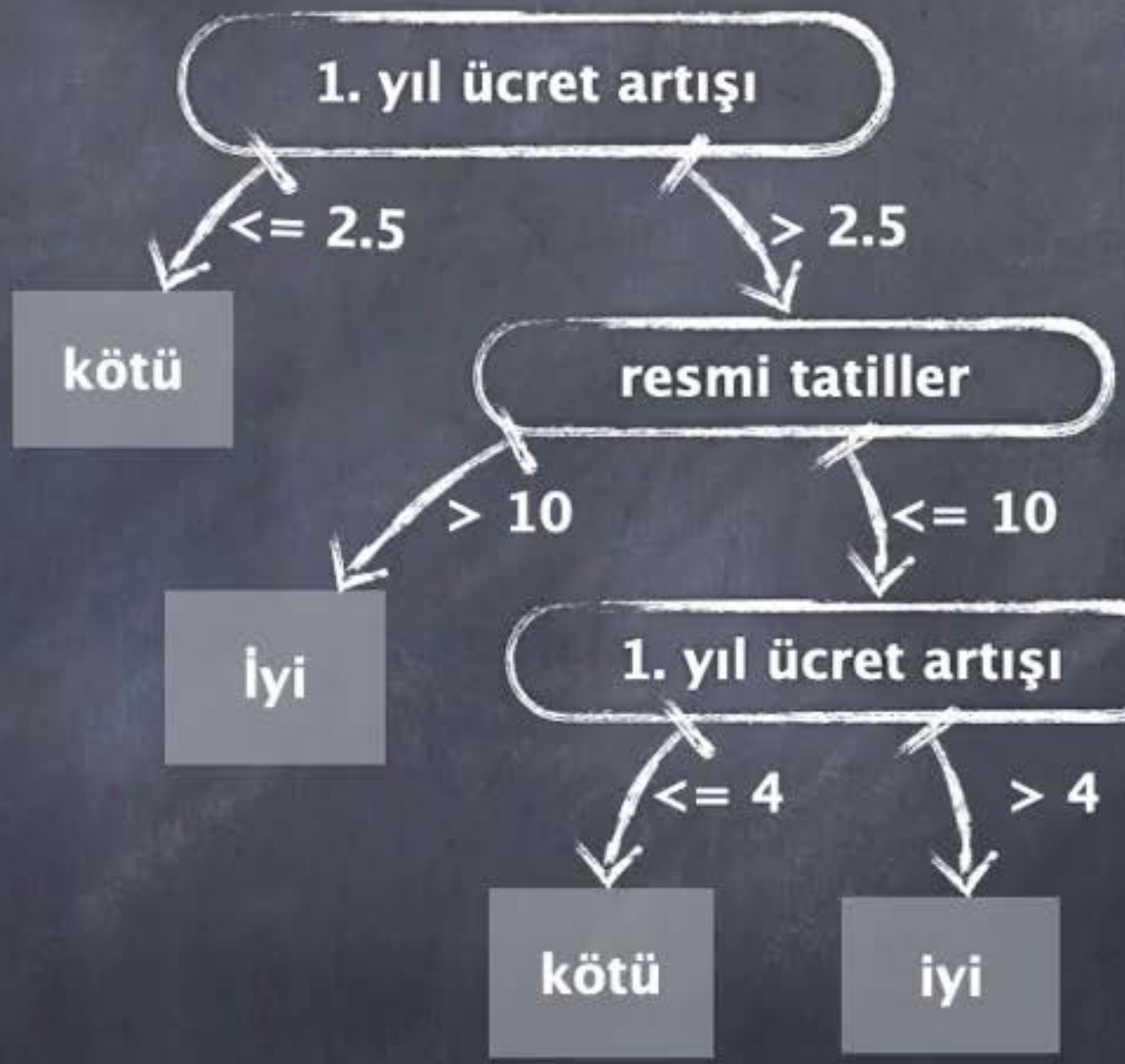
Özellik	Tip	1	2	3	...	40
Süre	(Yılların sayısı)	1	2	3		2
İlk yıl ücret artışı	Yüzde	%2	%4	%4.3		4.5
İkinci yıl ücret artışı	Yüzde	?	%5	%4.4		4.0
Üçüncü yıl ücret artışı	Yüzde	?	?	?		?
Yaşam maliyeti ayarlaması	{yok, tcf, tc}	Yok	tcf	?		Yok
Haftalık çalışma saatleri	(Saat sayısı)	28	35	38		40
Emeklilik	{hiçbiri, izin verilen, işçi	Yok	?	?		?
bekleme ödemesi	Yüzde	?	%13	?		?
Vardiyalı çalışma eki	Yüzde	?	%5	%4		4
Eğitim ödeneği	{Evet Hayır}	Evet	?	?		?
resmi tatiller	(Gün sayısı)	11	15	12		12
Tatil	{ortalamanın	ortala	gen	gen		ortalama
Uzun süreli sakatlık yardımı	{Evet Hayır}	hayır	?	?		Evet
Diş planı katkısı	{hiçbiri, yarım, tam}	Yok	?	tam		tam
yas yardımı	{Evet Hayır}	hayır	?	?		Evet
Sağlık planı katkısı	{hiçbiri, yarım, tam}	Yok	?	tam		yarım
Sözleşmenin kabul	{İyi kötü}	kötü	iyi	iyi		iyi

sınıflar

İşgücü Verileri için Karar Ağaçları

basittir,

Ağaç sezgisel bir anlam ifade ediyor



İşgücü Verileri için Karar Ağaçları

İşgücü Verileri için Karar Ağaçları

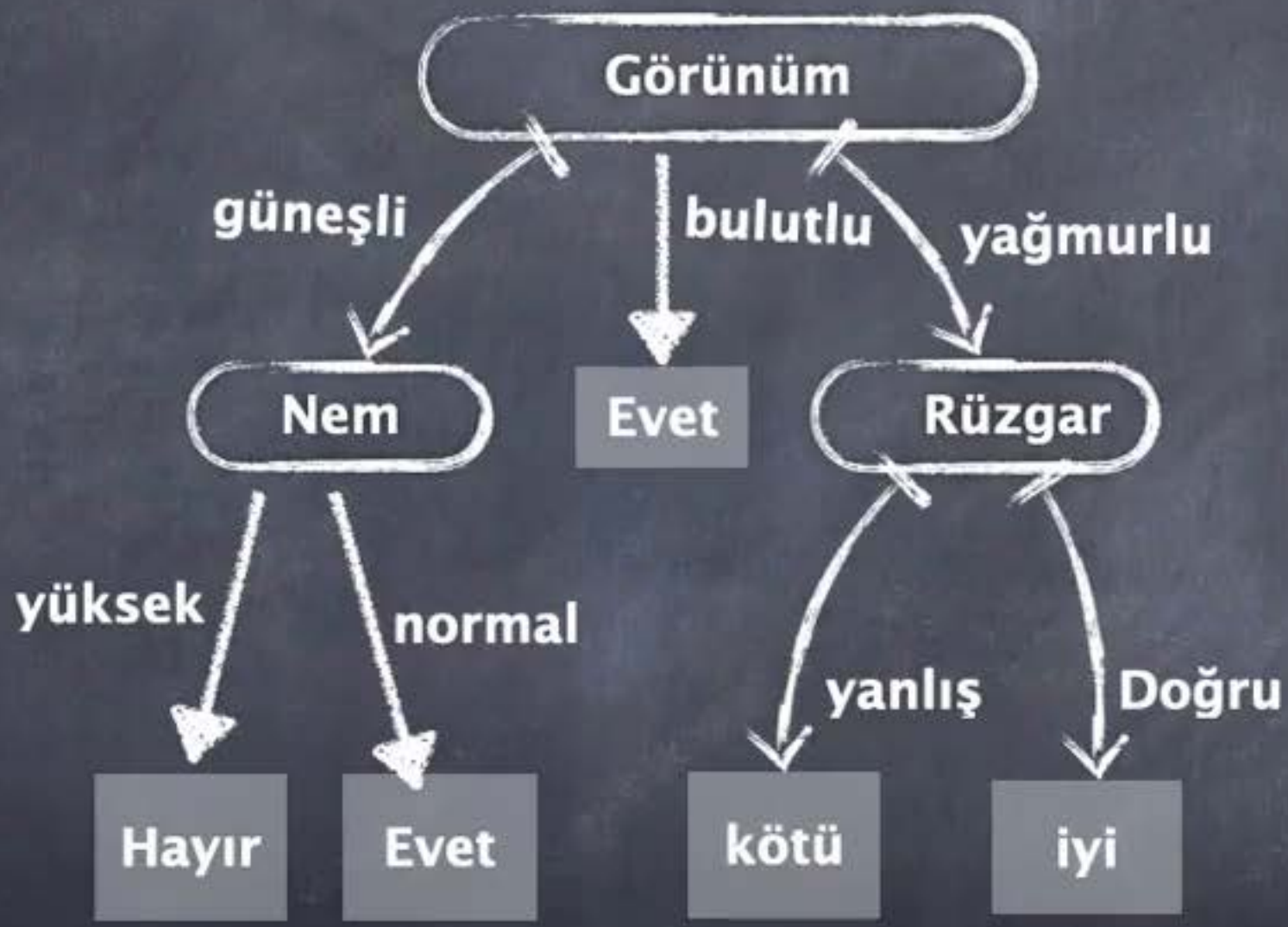
Bu ağaç basit ve yaklaşıktır,
tam olarak sınıflandırmaz.

İşgücü Verileri için Karar Ağaçları

Bu ağaç basit ve yaklaşıktır, tam olarak sınıflandırmaz.

Yukarıdaki basit ağaç, sağdakinin budanmış halidir.

Bütün ağaç, eğitim verilerinde daha doğrudur, AMA gerçek hayatta aslında daha iyi çalışmayabilir. "overfitted" olabilir.


Görünüm
güneşli
bulutlu
yağmurlu
Nem
Evet
Rüzgar
yüksek
normal
yanlış
Doğru
Hayır
Evet
kötü
iyi

Soya'nın Sınıflandırması

Bir Makine Öğrenimi başarı öyküsü!

	Özellik	Değer sayısı	Örnek değer
Çevre	Oluşma zamanı	7	Temmuz
	Yağış	3	Normalin üstü
Kök	Durum	2	Normal
	Küf gelişimi	2	Mevcut değil
Meyve	Meyve kabuklarının durumu	4	Normal
	Meyve lekeleri	5	?
Yaprak	Durum Yaprağı	2	Anormal
	nokta boyutu	3	?
Kök	Durum	2	Anormal
	kök yeri	2	Var
Kök Teşhisi	Durum	3	Normal
		19	Diaporthe kök kanseri

Alanından bir uzman, bilgisayar tarafından oluşturulan kurallar (%97.5 doğru) kadar iyi performans göstermeyen kurallar (%72 doğru) üretti.

Alan Bilgisinin Rolü

If Yaprak durumu normal
and gövde durumu anormal
and kök kanserleri toprak hattının altında
and kanser lezyon rengi kahverengi
then
teşhis rhizoctonia kök çürüklüğüdür

If Yaprak malformasyonu (kusuru) yoksa
and gövde durumu anormal
and kök kanserleri toprak hattının altında
and kanser lezyon rengi kahverengi
then
teşhis rhizoctonia kök çürüklüğüdür

Bu alanda, "kusurluluk yok", "yaprak durumu normal"in özel bir durumudur. Sadece yaprak durumu normal olmadığında devreye girer.

"Yaprak durumu normal" ile "yaprak kusuru yok" aynı şey midir?

Buraya kadar.... örnek problemler çocuk oyuncağıydı...

Küçük araştırma problemlerine örnekler.
Algoritmaları ve teknikleri anlamayı kolaylaştırdığı için bunları çok kullanacağız.

Peki ya gerçek uygulamalar?

Veri madenciliğini şu amaçlarla için kullanın:

- Karar vermek
- Bir işi bir uzmandan daha hızlı yapmak
- Uzmanın planı daha iyi hale getirmesine izin vermek
- vesaire.