VERİ MADENCİLİĞİ

Data Mining

Prof. Dr. Ünal Halit ÖZDEN

Ders Bilgileri

- **Kaynaklar**
  - Veri Madenciliği Yöntemleri, Yalçın Özkan.
  - Veri Madenciliği: Kavram ve Algoritmaları, Gökhan Silahtaroğlu.
  - Veri Madenciliği(Kavram ve Teknikler), Aysan Şentürk.

- **Başarı Notu**
  - Vize (%40)
  - Final (%60)

Genel İçerik

- Veri Madenciliğine Giriş
- Veri Madenciliğinin Süreci
  - Veri Temizleme
  - Veri Bütünleştirme
  - Veri İndirgeme
  - Veri Dönüştürme
  - Veri Madenciliği Yöntemini Seçme ve Uygulama
  - Sonuçları Sunma ve Değerlendirme
- Veri Madenciliği Yöntemleri
  - Sınıflandırma
  - Kümeleme
  - Birliktelik (İlişkilendirme) kuralları
- Metin madenciliği
- WEB madenciliği
- Veri Madenciliği Uygulamaları

VERİ MADENCİLİĞİ

Veri Madenciliğine Giriş

Prof. Dr. Ünal Halit ÖZDEN

Veri Madenciliği Giriş

- İçinde yaşadığımız bilişim çağında elektronik ortamda mevcut verinin hızlı artışı ve bilginin fazlalaşması sebebiyle öncelikle, genelde **Veri Tabanlarında Bilgi Keşfi** olarak adlandırılan yeni bir alan ortaya çıkmıştır. Daha yaygın bir kullanımla bu alana Veri Madenciliği denilmektedir.

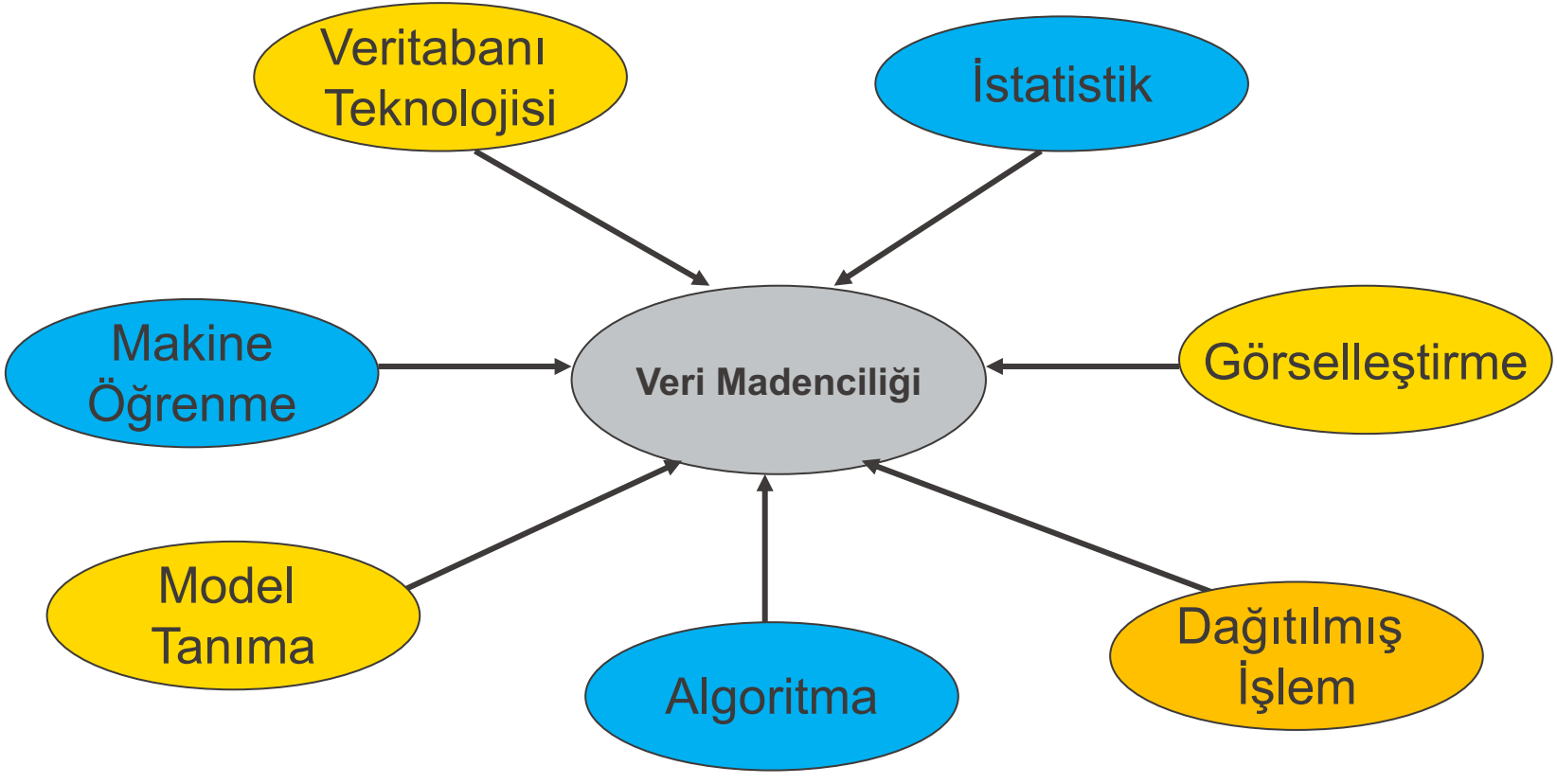
- Büyük veri kaynaklarından yararlı ve ilginç bilgiyi bulmak
- Bulunan bilgi
  - gizli,
  - önemli,
  - önceden bilinmeyen,
  - yararlı olmalı.

Veri Madenciliği Tanım

- Veri madenciliği, büyük ölçekli verileri kullanarak belirli yöntemlerle var olan ya da gelecekte ortaya çıkabilecek gizli bilgiyi, bağıntıyı, ilişkiyi, kalıbı ve kuralı ortaya çıkarma sürecidir. Bu açıdan bakıldığında veri madenciliği kurumların karar süreçlerinde önemli bir yere sahiptir.

- Veri madenciliği çalışmaları, *sınıflandırma, ilişki kurma, kümeleme, regresyon, veri özetleme, değişikliklerin analizi, sapmaların tespiti* gibi belirli sayıda teknik yaklaşımları içerir.

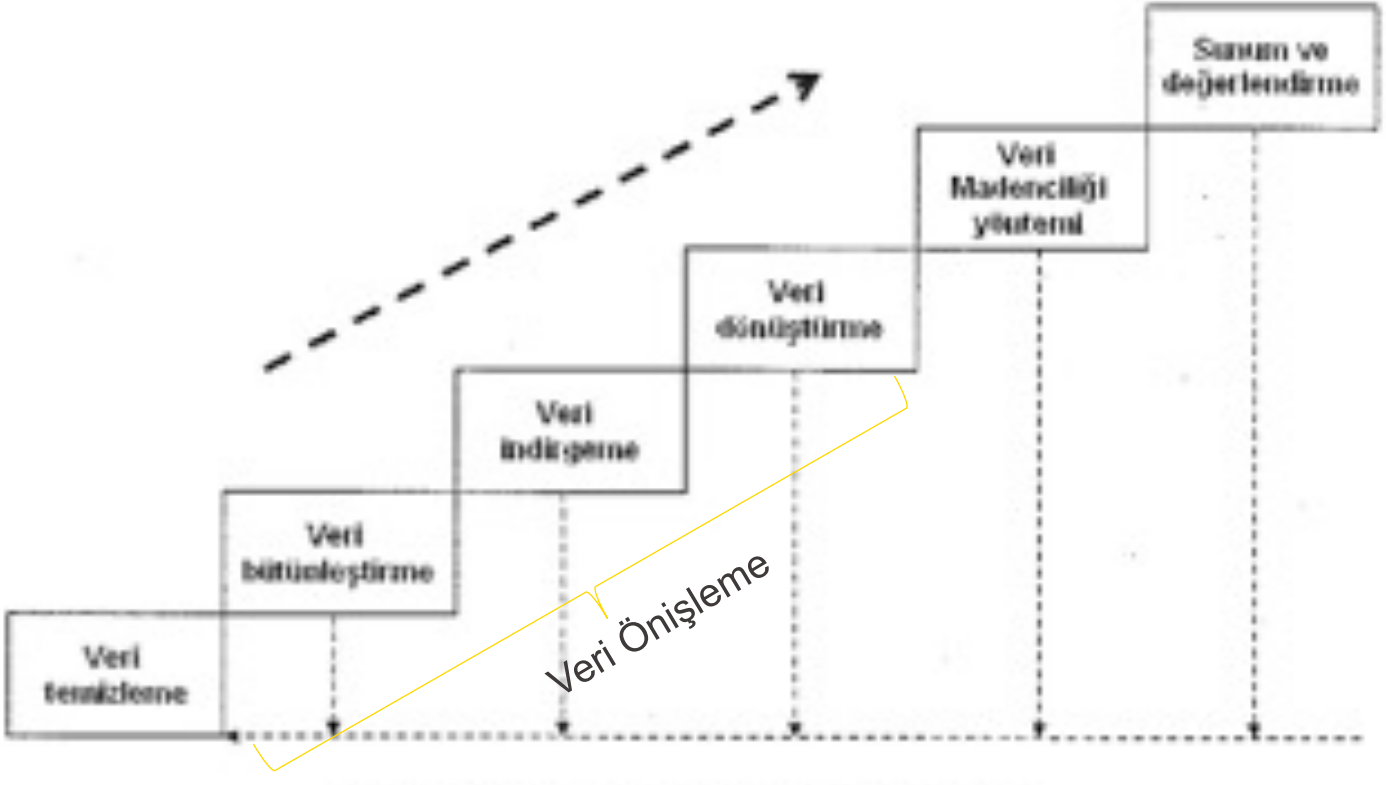
Veri Madenciliği ile İlişkili Diğer Disiplinler



İstatistik & Makine Öğrenmesi & Veri Madenciliği

- İstatistik
  - daha çok teoriye dayalı yaklaşımlar
  - bir varsayımın doğruluğunu araştırır
- Makine Öğrenmesi
  - daha çok sezgisel yaklaşımlar
  - öğrenme işleminin başarımını artırmaya çalışır
- Veri madenciliği ve bilgi keşfi
  - teori ve sezgisel yaklaşımları birleştirir
  - bilgi keşfinin tüm aşamalarını gerçekler: veri temizleme, öğrenme, sonucu sunma, yorumlama,...
- Aradaki ayrım net değil

Veri Madenciliği Süreci



Şekil-2.1. Veri madenciliği süreci [Han, 2000]

Bilgi Keşfinin Aşamaları

- Veri Temizleme : Gürültülü ve tutarsız verileri çıkarmak
- Veri Bütünleştirme: Birçok data kaynağını birleştirebilmek
- Veri Seçme : Yapılacak olan analiz ile ilgili olan verileri belirlemek
- Veri İndirgeme :
- Veri Dönüşümü : Verinin veri madenciliği yöntemine göre hale dönüşümünü gerçekleştirmek
- Veri Madenciliği : Verilerdeki örüntülerin belirlenmesi için veri madenciliği yöntemlerinin uygulanması
- Örüntü Değerlendirme: Bazı ölçütlere göre elde edilmiş ilginç örüntüleri bulmak ve değerlendirmek
- Bilgi Sunumu : Elde edilen bilgilerin kullanıcılara sunumunu

Veri Madenciliği Uygulama Alanları

- Pazarlama
  - Müşteri satın alma eğilimlerinin belirlenmesi
  - Müşteri kayıp analizleri
  - Pazar sepeti Analizleri
  - Satışlardaki anormal gelişmelerin saptanması
  - Müşteri segmentasyonu
  - Müşteri memnuniyet araştırması

Veri Madenciliği Uygulama Alanları

- Finans
  - Sigorta dolandırıcılıklarının tespiti
  - Kredi kartı dolandırıcılıklarının tespiti
  - Kredi kartı harcamalarına göre müşteri gruplarının belirlenmesi
  - Kredi taleplerinin analizleri
  - Personel kayıp analizleri
  - Finansal göstergeler arasındaki gizli ilişkilerin belirlenmesi
  - Riskli müşteri gruplarının belirlenmesi

Veri Madenciliği Uygulama Alanları

- Elektronik Ticaret
  - Sanal mağazalar için saldırıların tespiti
  - Web sayfalarına yapılan ziyaretlerin analizi
  - Web sayfalarına yapılan saldırıların belirlenmesi
  - Müşteri değerlendirmelerinin analizi
  - Kredi kartı dolandırıcılıklarının tespiti
  - Müşteri memnuniyet araştırması
- Sağlık
- Sosyal Medya

Veri Madenciliği Uygulama Alanları

Bilim	İş Hayatı	Web	Devlet
• Astronomi • Biyoinformatik • İlaç keşfi	• Reklam • CRM (Müşteri İlişkileri Yönetimi) ve Müşteri Modelleme • E-ticaret • Yatırım değerlendirme ve karşılaştırma • Sağlık • Üretim • Spor/eğlence • Telekom (telefon ve iletişim) • Hedef pazarlama	• Metin Madenciliği (haber grubu, email, dokümanlar) • Web analizi • Arama motorları	• Terörle Mücadele • Kanun Yaptırımı • Vergi Kaçakçılarının Profilinin Çıkarılması

Örnek Uygulamalar

- Hangi promosyonu ne zaman uygulamalıyım?
- Hangi müşteri aldığı krediyi geri ödemeyebilir?
- Bir müşteriye ne kadar kredi verilebilir?
- Sahtekarlık olabilecek davranışlar hangileridir?
- Hangi müşteriler yakın zamanda kaybedilebilir?
- Hangi müşterilere promosyon yapmalıyım?
- Hangi yatırım araçlarına yatırım yapmalıyım?

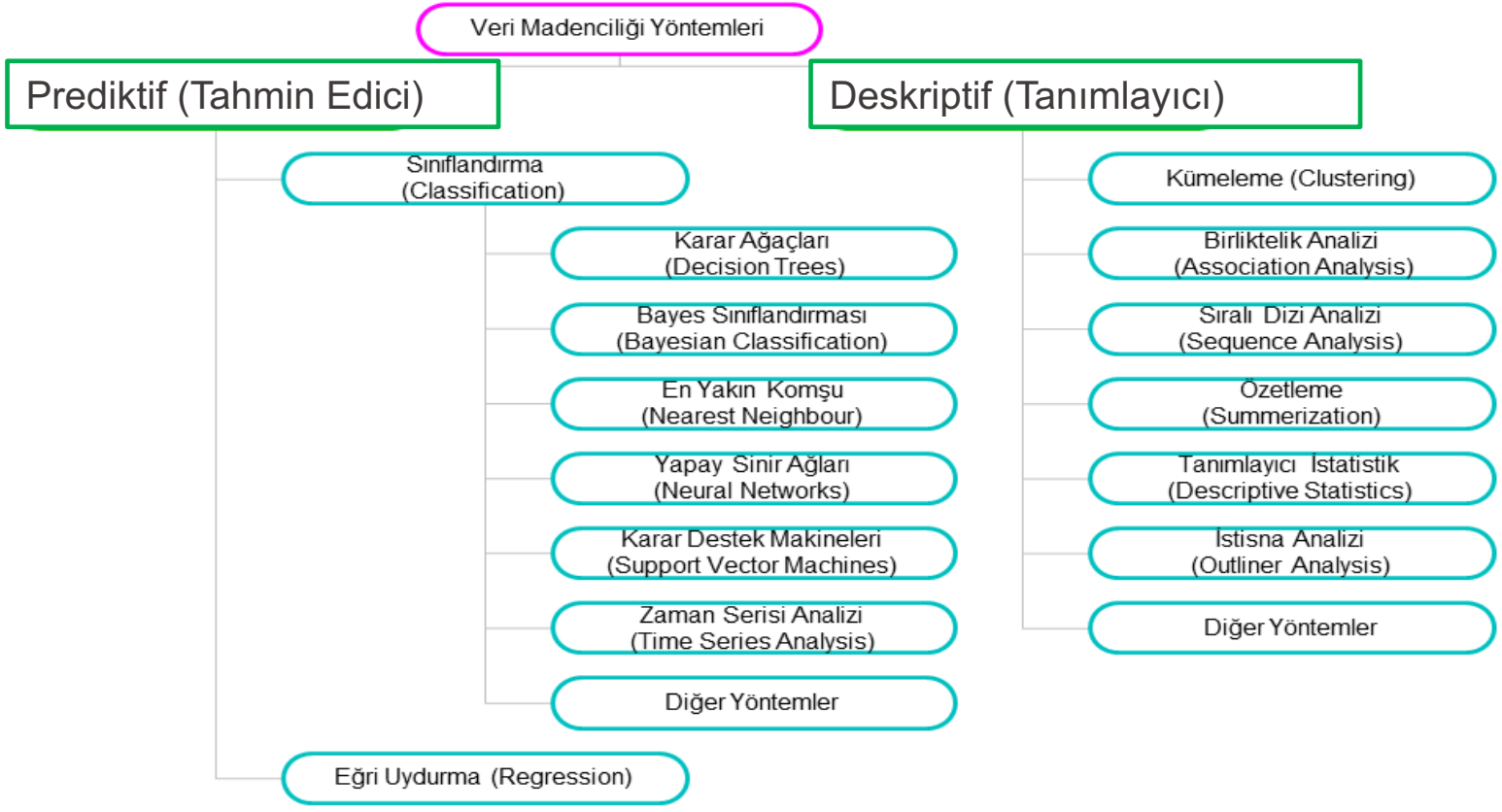
Veri Kaynakları

- Veri dosyaları
- Veritabanı kaynaklı veri kümeleri
  - ilişkisel veritabanları, veri ambarları
- Gelişmiş veri kümeleri
  - Veri akışı (data stream), algılayıcı verileri (sensor data)
  - zaman serileri, sıralı diziler (biyolojik veriler)
  - çizgeler, sosyal ağ (social networks) verileri
  - konumsal veriler (spatial data)
  - çoğul ortam veritabanları (multimedia databases)
  - nesneye dayalı veritabanları
  - www

Veri Madenciliği Algoritmaları

- amaç: veriyi belli bir modele uydurmak
  - tanımlayıcı
    - En iyi müşterilerim kimler?
    - Hangi ürünler birlikte satılıyor?
    - Hangi müşteri gruplarının alışveriş alışkanlıkları benzer?
  - kestirime dayalı
    - Kredi başvurularını risk gruplarına ayırma
    - Şirketle çalışmayı bırakacak müşterileri öngörme
    - Borsa tahmini
- seçim: veriye uyan en iyi modeli seçmek için kullanılan kriter
- arama: veri üzerinde arama yapmak için kullanılan teknik

Veri Madenciliği Modelleri (Yöntemleri)



Veri Madenciliği Yöntemleri (1/2)

- Sınıflandırma (Classification): Veriyi önceden belirlenmiş sınıflardan birine dahil eder.
  - Danışmanlı (Gözetimli) öğrenme
  - Örüntü tanıma
  - Kestirim
- Eğri uydurma (Regression): Veriyi gerçel değerli bir fonksiyona dönüştürür.
- Zaman serileri inceleme (Time Series Analysis): Zaman içinde değişen verinin değerini öngörür.
- İstisna Analizi (Outlier Analysis): Verinin geneline uymayan nesneleri belirleme

Veri Madenciliği Yöntemleri (2/2)

- Kümeleme (Clustering): Benzer verileri aynı grupta toplama
  - Danışmansız (Gözetimsiz) öğrenme
- Özetleme (Summarization): Veriyi alt gruplara ayırır. Her alt grubu temsil edecek özellikler bulur.
  - Genelleştirme (Generalization)
  - Nitelendirme (Characterization)
- İlişkilendirme kuralları (Association Rules)
  - Veriler arasındaki ilişkiyi belirler
- Sıralı dizileri bulma (Sequence Discovery): Veri içinde sıralı örüntüler bulmak için kullanılır.

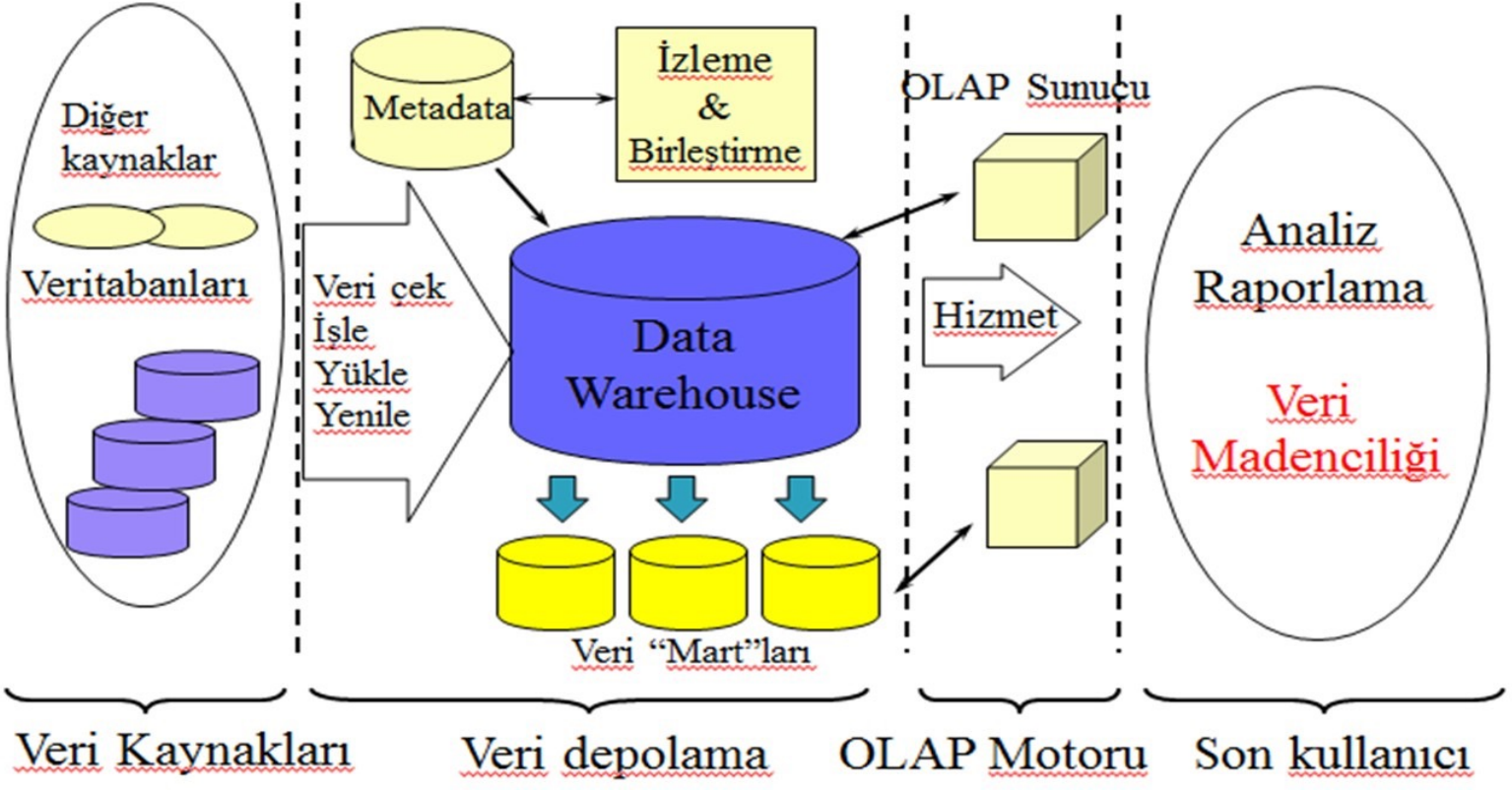
Araştırma Sonucunda Bulunan Bilgilerin Anlamlılığı

- Veri madenciliği büyük riski, anlamsız kalıpları "keşfetmeniz" dir..
- İstatistikçiler buna Bonferroni'nin ilkesi diyorlar: (kabaca) ilginç kalıplar için veri miktarınızın destekleyeceğinden daha fazla yere bakarsanız, saçmalık bulmak zorundasınız.
- Ren Paradoksu: Bilimsel araştırmanın nasıl yapılmayacağına dair harika bir örnek.
-

Veri Ambarı

- **Veritabanı:** birbirleriyle ilişkili bilgilerin depolandığı alanlardır.
- **Veri Ambarı**: ilişkili verilerin sorgulandığı ve analizlerinin yapılabildiği bir depodur. Veri ambarı veritabanını yormamak için oluşturulmuştur. Bir veri ambarı ilgili veriyi kolay, hızlı, ve doğru biçimde analiz etmek için gerekli işlemleri yerine getirir. Veri ambarı, işlemsel sistemlerdeki veriyi kopyalayıp, karar verme işlemi için uygun formda saklar.
- **Data Mart:** veri ambarlarının alt kümeleridir. Veri ambarları bir iş probleminin tamamına yönelik bir bakış sağlarken, data mart'lar sadece belli bir kısma bakış sağlarlar. Veri pazarları ile veriye hızlı erişim sağlayabiliriz. İkinci olarak, verinin gruplanmamış yapıda olması ve farklı iş birimlerinin farklı verileri görmesidir. Bu da bize gereksiz bir iş yükü ve güvenlik sorununa neden olmaktadır. İşte tam bu noktada, veri pazarları konuya, bölümlere uygun, veri ambarının küçük bir kopyası halinde çözüm sunmaktadır.

Veri Ambarı Mimarisi



Uyarı

- Veri madenciliği yöntemleri bilinçsiz olarak kullanılmamalı
  - Veri madenciliği yöntemleri geçmiş olaylara bakarak örüntüler bulur: Gelecekteki olaylar geçmiştekilerle aynı değildir
  - İlişkiler her zaman nedenleri açıklamaz

VERİ MADENCİLİĞİ

Veri Önişleme

Prof. Dr. Ünal Halit ÖZDEN

Veri Önişleme

- Veri
  - Veri Türleri/Değişken Türleri
  - Ölçekler
- Veri Önişleme
  - Veriyi Tanımlama
  - Veri Temizleme
  - Veri Bütünleştirme
  - Veri İndirgeme (azaltma)
  - Veri Dönüştürme
- Benzerlik ve Farklılık

Veri Önişleme/Veri

Veri Önişleme
-Veri

- Nesne (birim) ve niteliklere (değişken) ilişkin verilerin toplanması
- Nitelik bir nesnenin özelliği veya karekteristiğidir
  - Örnekler: Bir kişinin göz rengi, sıcaklık vb.
  - Nitelik, değişken, özellik veya karekter olarak da bilinir.
- Nitelikler (özellikler) kümesi, (nesneyi) birimi tanımlar
  - Nesne ayrıca kayıt, nokta, durum, vaka, varlık veya örnek olarak da bilinir
  -

Attributes

(Nitelik-Özellik-Değişken)

Nesneler (Birimler)

ID	Geri Ödeme	Medeni Durum	Gelir	Dolandırıcılık
1	Evet	Bekar	125.000	Hayır
2	Hayır	Evli	100.000	Hayır
3	Hayır	Bekar	70.000	Hayır
4	Evet	Evli	120.000	Hayır
5	Hayır	Boşanmış	95.000	Evet
6	Hayır	Evli	60.000	Hayır
7	Evet	Boşanmış	220.000	Hayır
8	Hayır	Bekar	85.000	Evet
9	Hayır	Evli	75.000	Hayır
10	Hayır	Bekar	90.000	Evet

Büyüklük (N): Birim (nesne) sayısı
Boyut: Değişkenlerin (nitelik) sayısı
Sparsity: Nesne-nitelik (birim-değişken) çiftlelerinin sayısı

Veri Önişleme

-Veri

--Veri/Değişken/Ölçek Türleri

- Nümerik Veriler/Nümerik Değişkenler
  - Sürekli
  - Kesikli
- Kategorik Veriler/Kategorik Değişkenler

Ölçekler

- İsimsel
- Sıralı
- Aralıklı
- Oransal

Veri Önişleme/Veri Tanımlama

Veri Önişleme
-Veriyi Tanımlama

- Amaç: Veriyi daha iyi anlamak
  - Merkezi eğilim (central tendency),
  - Değişkenlik ölçüleri,
  - Dağılım ölçüleri,
  - Grafikler
  - Frakans tabloları

Veri Önişleme
-Veri Tanımlama
--Merkezi Eğilimi Ölçme

Ortalama: $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ $\mu = \frac{\sum x}{N}$

- ağırlıklı ortalama $\bar{x} = \frac{\sum_{i=1}^{n} w_i x_i}{\sum_{i=1}^{n} w_i}$
- kırpılmış ortalama: Uç değerleri kullanmadan hesaplama

Ortanca (median): Verinin tümü kullanılarak hesaplanır

- veri sayısı tek ise ortadaki değer, çift sayı ise ortadaki iki değerin ortalaması

$$median = L_1 + (\frac{n/2 - (\sum f)_l}{f_{median}})c$$

Mod

- Veri içinde en sıklıkla görülen değer
- Unimodal, bimodal, trimodal

Veri Önişleme
-Veri Tanımlama
--Değişkenlik Ölçüleri

- Amaç: Verilerin ne kadar değişkenlik götsediğini anlamak (Homojenliği ölçmek)
  - Değişim aralığı,
  - Standar Sapma,
  - Varyans,
  - Değişim Katsayısı

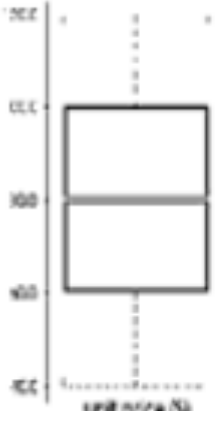
Veri Önişleme

-Veri Tanımlama

--Verinin Dağılımını Ölçme-1

Çeyrek, aykırılıklar, kutu grafiği çizimi

- Çeyrek (quartile) : nitelik değerleri küçükten büyüğe doğru sıralanır.
  - Q1: ilk %25, Q3: ilk %75
- Dörtlü aralık (Inter-quartile Range): IQR= Q3-Q1
- Five Number Summary: min, Q1, median, Q3, max
- Kutu Grafiği Çizimi:
  - Q1 ve Q3 aralığında bir kutu
  - kutu içinde ortanca noktayı gösteren bir çizgi
  - kutudan min ve max değerlere birer uzantı
- Aykırılıklar: 1,5xIQR değerinden küçük/büyük olan



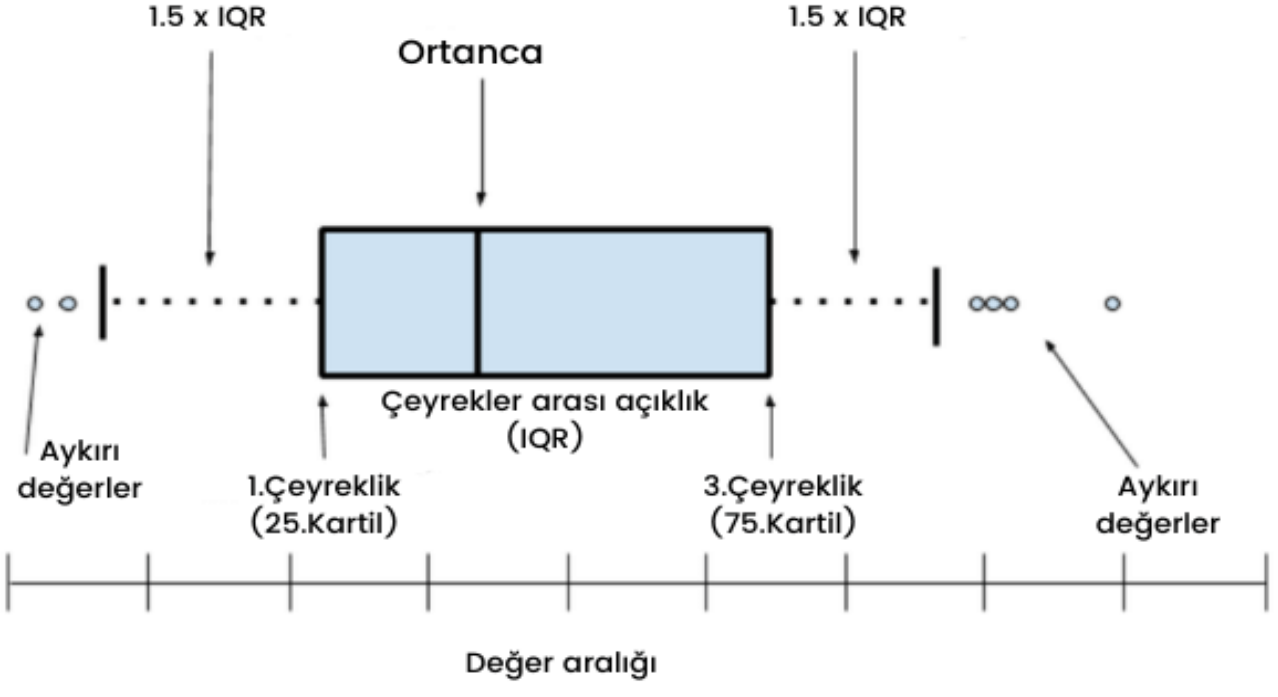
Varyans ve standart sapma

$$= \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 = \frac{1}{n-1}[\sum_{i=1}^{n}x_i^2 - \frac{1}{n}(\sum_{i=1}^{n}x_i)^2] \qquad \sigma^2 = \frac{1}{N}\sum_{i=1}^{n}(x_i - \mu)^2 = \frac{1}{N}\sum_{i=1}^{n}x_i^2 - \mu^2$$

Veri Önişleme

-Veri Tanımlama

--Verinin Dağılımını Ölçme-2



Veri Önişleme
--Genel Bilgi

- Veri temizleme
  - Eksik nitelik değerlerini tamamlama, hatalı veriyi düzeltme, aykırılıkları saptama ve temizleme, tutarsızlıkları giderme
- Veri birleştirme
  - Farklı veri kaynağındaki verileri birleştirme
- Veri indirgeme (azaltma)
  - Aynı veri madenciliği sonuçları elde edilecek şekilde veri miktarı ve/veya değişken sayısı azaltılabilirma
- Veri dönüştürme
  - Normalizasyon, standartlaştırma vs.

Veri Önişleme/Veri Temizleme

Veri Önişleme
-Veri Temizleme

- Gerçek uygulamalarda toplanan veri kirli olabilir.
  - Eksik Veriler: bazı nitelik değerleri bazı nesneler için girilmemiş, veri madenciliği uygulaması için gerekli bir nitelik kaydedilmemiş
    - meslek = " "
  - Gürültülü Veriler: Aykırı değerler ve hatalar var
    - maaş= "-10"
  - Tutarsız ve Tekrarlanan Veriler: nitelik isimleri, nitelik değerleri uyumsuz veya teklararlanan veriler söz konusu
    - yaş= "35", d.tarihi: "03/10/2004"
    - önceki oylama değerleri: "1,2,3", yeni oylama değerleri: "A,B,C"
    - bir kaynakta nitelik değeri 'ad', diğerinde 'isim'

Veri Önişleme
-Veri Temizleme

Bir hata mı yoksa milyoner mi?

Eksik değerler

Tutarsız yinelenen girişler

ID	Geri Ödeme	Medeni Durum	Gelir	Dolandırıcılık
1	Evet	Bekar	125K	Hayır
2	Hayır	Evli	100K	Hayır
3	Hayır	Bekar	70K	Hayır
4	Evet	Evli	120K	Hayır
5	Hayır	Boşanmış	10000K	Evet
6	Hayır	NULL	60K	Hayır
7	Evet	Boşanmış	220K	NULL
8	Hayır	Bekar	85K	Evet
9	Hayır	Evli	90K	Hayır
9	Hayır	Bekar	90K	Hayır

Veri Önişleme
-Veri Temizleme
--Eksik Veriler Tamamlama

- Elle doldurma
- Nesneyi (birimi) veri setinden çıkarma
- Eksik nitelik değerleri için global bir kavram kullanma (Null, bilinmiyor,...)
- Niteliğin ortalama değeri ile doldur
- Aynı sınıfa ait değerlerin ortalaması ile doldurma
- Olasılığı en fazla olan nitelik değeriyle doldurma (Mod değeri)

Veri Önişleme
-Veri Temizleme
--Gürültüyü yok etme yöntemleri-1

- Bölmeleme
  - veri sıralanır, eşit genişlik veyaeşit derinlik ile bölünür
- Kümeleme
  - aykırılıkları belirler
- Eğri uydurma
  - veriyi bir fonksiyona uydurarak gürültüyü düzeltir.

Veri Önişleme
-Veri Temizleme
--Gürültüyü yok etme yöntemleri-2
---Bölmeleme

- Veri sıralanır: 4, 8, 15, 21, 21, 24, 25, 28, 34
  - Eşit genişlik: Bölme sayısı belirlenir. Eşit aralıklarla bölünür
  - Eşit derinlik: Her bölmede eşit sayıda örnek kalacak şekilde bölünür.
    - her bölme ortalamayla ya da bölmenin en alt ve üst sınırlarıyla temsil edilir.

Bölme genişliği:3
1. Bölme: 4, 8, 15
2. Bölme: 21, 21, 24
3. Bölme: 25, 28, 34

Ortalamayla düzeltme:
1. Bölme: 9, 9, 9
2. Bölme: 22, 22, 22
3. Bölme: 29, 29, 29

Alt-üst sınırla düzeltme:
1. Bölme: 4, 4, 15
2. Bölme: 21, 21, 24
3. Bölme: 25, 25, 34

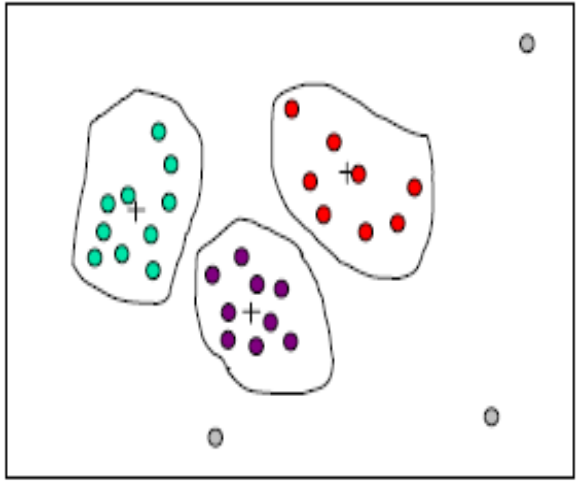
Veri Önişleme

-Veri Temizleme

--Gürültüyü yok etme yöntemleri-3

---Kümeleme

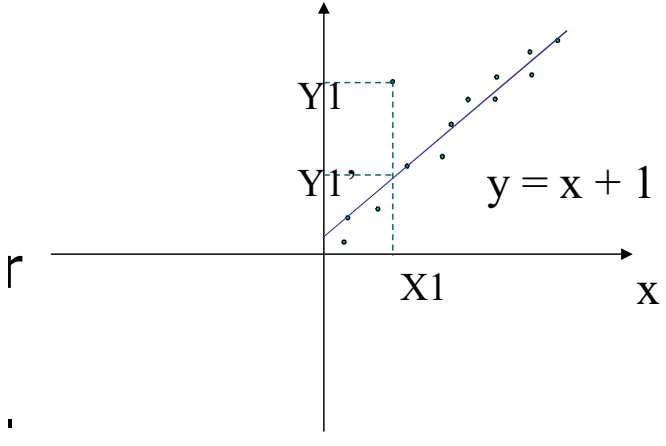
- Benzer veriler aynı kümede olacak şekilde gruplanır
- Bu kümelerin dışında kalan veriler aykırılık olarak belirlenir ve silinir.



Veri Önişleme
-Veri Temizleme
--Gürültüyü yok etme yöntemleri-4

---Eğri Uydurma

- Veri bir fonksiyona uydurulur. Doğrusal eğri uydurmada, bir değişkenin değeri diğer bir değişken kullanılarak bulunabilir.



--Sonuç

Veri kaliteli ise veri madenciliği uygulamaları ile yararlı bilgi bulma şansı daha fazla.

Veri Önişleme/Veri Bütünleştirme

Veri Önişleme

Veri Bütünleştirme

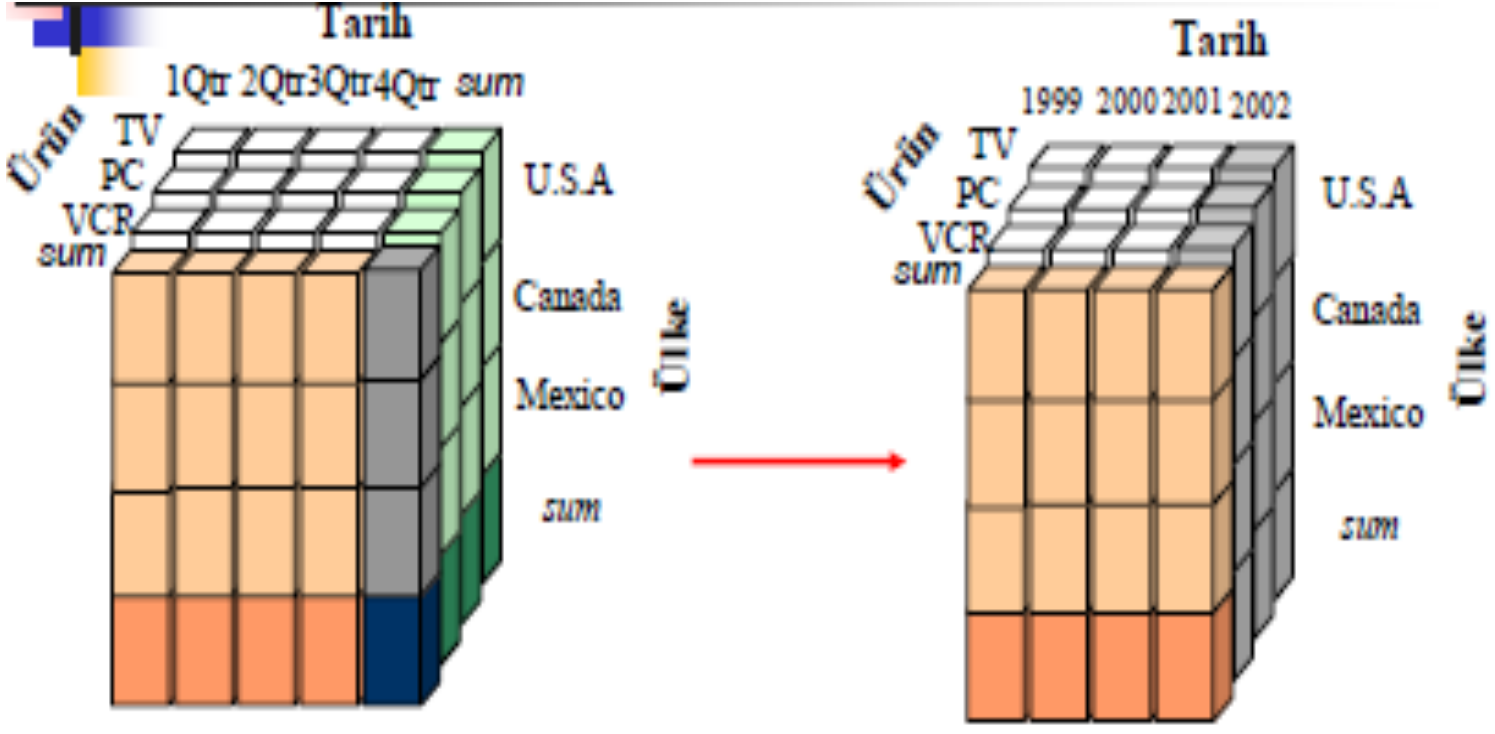
- Farklı kaynaklardaki verilerin tutarlı olarak bir araya getirilerek birleştirilmesi
- Nitelik değerlerinin tutarsızlığının saptanması
  - Aynı nitelik için farklı kaynaklarda farklı değerlerin ve/veya isimlerin olması
  - Farklı ölçü birimlerinin kullanılması (kg ve gr. gibi)
- Bütünleştirme sonucu gereksiz (fazla) veri ortaya çıkabilir. Bunların saptanarak düzeltilmesi gerekir

Veri Önişleme/Veri İndirgeme

Veri Önişleme
-Veri İndirgeme

- Nitelik ve veri miktarı çok fazla olduğu zaman veri madenciliği algoritmalarının çalışması ve sonuç üretmesi çok uzun sürebilir
  - veriyi azaltma başarımı artırabilir
  - Ancak veri indirgemesi yapıldığında sonucun (nerdeyse) hiç değişmemesi gerekir
- Veri İndirgeme
  - nitelik birleştirme
  - nitelik azaltma
  - veri ayrıştırma
  - kavram oluşturma
  - veri küçültme
    - eğri uydurma
    - kümeleme
    - histogram
    - örnekleme

Veri Önişleme
-Veri İndirgeme
--Nitelik Birleştirme



Veri Önişleme
-Veri İndirgeme
--Nitelik Seçme/Nitelik Azaltma-1

- Nitelik Seçme
  - Nitelikler kümesinin bir alt kümesi seçilerek veri madenciliği işlemi yapılır.
- Nitelik azaltma
  - d boyutlu veri kümesi k<d olacak şekilde k boyuta taşınır.

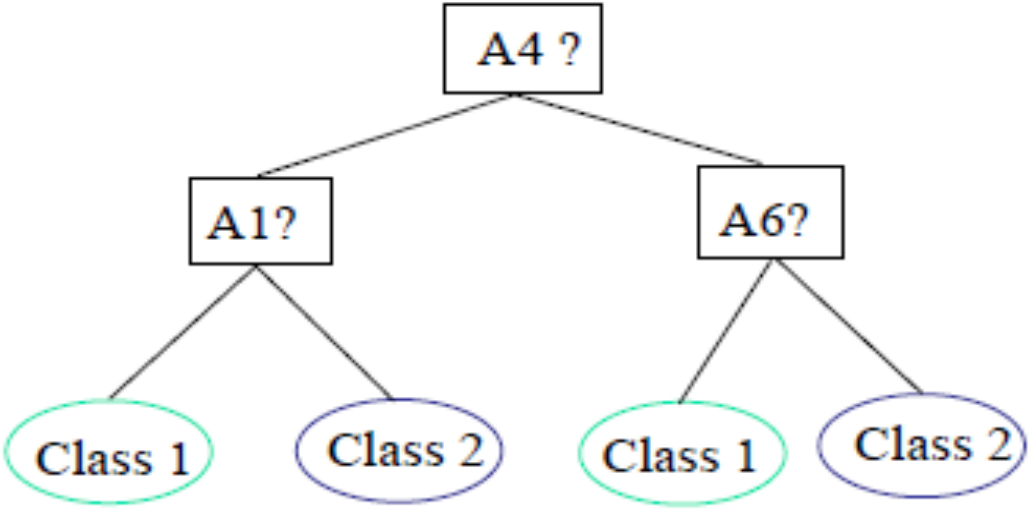
Veri Önişleme
-Veri İndirgeme
--Nitelik Seçme/Nitelik Azaltma-2

- Nitelik seçme
  - Veri madenciliği uygulaması için gerekli olan niteliklerin seçilmesi
  - Nitelikler altkümesi kullanılarak elde edilen sınıfların dağılımları gerçek dağılıma eşit ya da çok yakın olmalı
  - Veri madenciliği işlemi yer ve zaman karmaşıklığını azaltma
- Sistemin başarımını artırma
  - Sezgisel yöntemler kullanılarak nitelikler seçilebilir.
  - istatistiksel anlamlılık testi (statistical significance)
  - bilgi kazancı (information gain)
  - karar ağaçları

Veri Önişleme
-Veri İndirgeme
--Nitelik Seçme/Nitelik Azaltma-3

Başlangıç nitelikler kümesi:

{A1, A2, A3, A4, A5, A6}



Seçilen nitelik kümesi: {A1, A4, A6}

Veri Önişleme
-Veri İndirgeme
--Veri Ayrıştırma

- Bazı veri madenciliği algoritmaları sadece ayrık veriler ile çalışır.
- Sürekli bir nitelik değerini bölerek her aralığı etiketler.
- Verinin değeri, bulunduğu aralığın etiketi ile değişir.
- Veri boyutu küçülür.
- Kavram oluşturmak için kullanılır.
- Örn. Verilerde yuvarlama yapma vs.

Veri Önişleme
-Veri İndirgeme
--Kavram Oluşturma

- Yaşlara göre; çocuk, genç, yaşlı
- Kiloya göre; zayıf, normal, kilolu
- Gelire göre; fakir, orta direk, zengin
- Kodlama yapılabilir….

Veri Önişleme
-Veri İndirgeme
--Veri Küçültme

- Veriyi farklı şekillerde gösterme
  - parametrik
    - eğri uydurma (regresyon)
  - parametrik olmayan
    - histogram
    - kümeleme
    - örnekleme

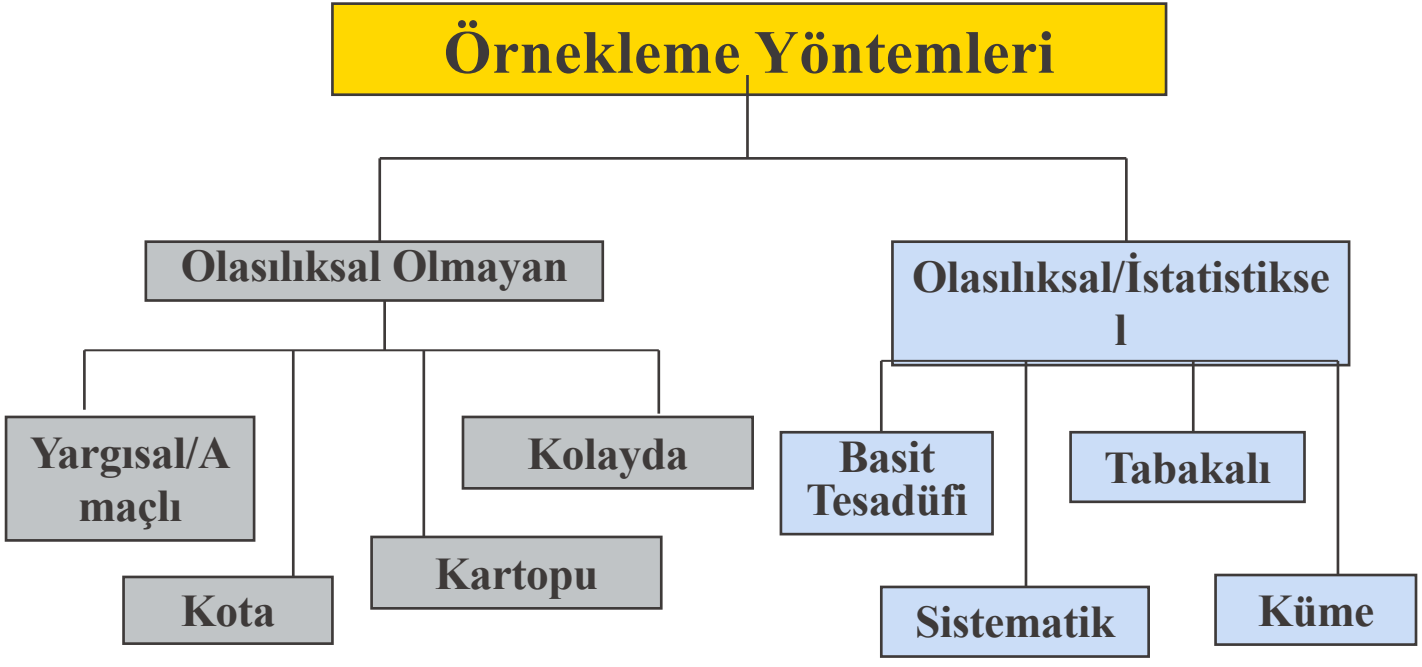
Veri Önişleme
-Veri İndirgeme
--Veri Küçültme
---Kümeleme

- Veri kümelere ayrılır
- Veri kümeleri temsil eden örnekler (küme merkezleri) ve aykırılıklar ile temsil edilir
- Etkisi verinin dağılımına bağlı.
- Hiyerarşik kümeleme yöntemleri kullanılabilir.

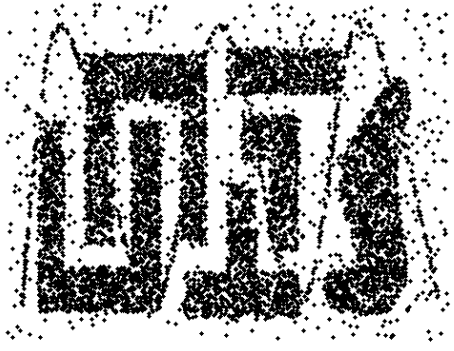
Veri Önişleme
-Veri İndirgeme
--Veri Küçültme
---Örnekleme-1

- Büyük veri kümesini daha küçük bir alt küme ile temsil etme
- Bir örneklem kullanmak, örneklem temsiliyse neredeyse tüm veri kümelerini kullanmayla aynı sonucu verir.

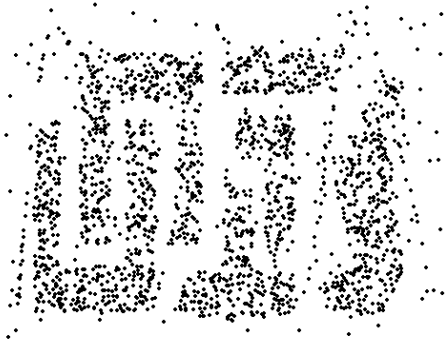
Veri Önişleme
-Veri İndirgeme
--Veri Küçültme
---Örnekleme-2



Veri Önişleme
-Veri İndirgeme
--Veri Küçültme
---Örnekleme-3



8000 nokta



2000 nokta

500 nokta

Örneklem Büyüklüğü

Veri Önişleme
-Veri Dönüşümü

- Veri, veri madenciliği uygulamaları için uygun olmayabilir.
- Seçilen algoritmaya uygun olmayabilir.

  Böyle bir durumda var olan veriler dönüştürülerek kullanılabilir hale getirilir.

  - Veri Dönüştürme Yöntemleri:
    - Genelleme
    - Normalizasyon
    - Yeni nitelik oluşturma

Veri Önişleme
-Veri Dönüştürme
---Genelleme

Böyle bir durumda niteliklerin her bir değeri yerine karekteristik değerleri kullanılır.

- Aritmetik Ort.
- Mod
- Medyan
- Vs.

Veri Önişleme
-Veri Dönüştürme
--Normalizasyon

- min-max normalizasyon

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

$$X_{min} = 30$$

$$X_{max} = 62$$

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

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

- z-score normalizasyon

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

$$X^* = \frac{X - \bar{X}}{\sigma_X} = \frac{30 - 44.6}{12.44} = -1.1735$$

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

- ondalık normalizasyon

$X^* = X/10^j = 30/10^2 = 0{,}30$ vs.

Veri Önişleme
-Veri Dönüştürme
--Yeni Nitelik Oluşturma

- Yeni nitelikler yarat
  - orjinal niteliklerden daha önemli bilgi içersin
    - alan=boy x en
  - veri madenciliği algoritmalarının başarımı daha iyi olsun

VERİ MADENCİLİĞİ

Veri Madenciliği Algoritmasını Uygulama

Prof. Dr. Ünal Halit ÖZDEN

Veri Madenciliği Algoritmasını Uygulama

- Veri madenciliği yöntemlerini uygulayabilmek için yukarıda açıklanan uygun olan işlemlerin gerçekleştirilmesi gerekir. Bu işlemler gerçekleştirilip veriler uygun hale getirildikten sonra, veri madenciliği yöntemlerinden sınıflama, kümeleme ve birliktelik kuralı algoritmalarından birisi verilere uygulanır.

VERİ MADENCİLİĞİ

Sonuçların Sunumu ve Değerlendirme

Prof. Dr. Ünal Halit ÖZDEN

Sonuçların Sunumu ve Değerlendirme

- Veri madenciliği algoritması verile üzerine uygulandıktan sonra, sonuçlar değerlendirilip düzenlenerek ilgili mercilere sunulur. Bu aşama raporlama aşaması olarak da adlandırılır. Raporlamada talolalar grafikler vs. gibi unsurlardan yararlanılır.

VERİ MADENCİLİĞİ

KNIME Analitik Platformuna Giriş

Prof. Dr. Ünal Halit ÖZDEN

İndirme ve yükleme

KNIME Analitik Platformu

- Açık kaynaklı modüler veri bilim platformu
- Tüm veri bilim ihtiyaçlarını kapsar

Veri Erişimi	Veri Hazırlama	Veri Görselleştirme
Makine Öğrenim	Test Etme	Dağıtım

- Görsel programlama paradigmasına dayalı
- Çok çeşitli uzantılar sağlar:
  - Metin Madenciliği
  - Ağ Madenciliği
  - Derin Öğrenme
  - Java, R, Python, Weka, Keras, Plotly, H2O gibi birçok entegrasyon
  - ... Ve daha fazlası

KNIME Sunucusu

KNIME Analitik Platformu

- Veri bilimi çözümleri geliştirmek
  - Yapılandırılmış veri
  - Yapılandırılmamış veriler
  - Makine öğrenme
  - İstatistik
- Açık kaynak
- Özgür

KNIME Sunucusu

- Çözümleri BT ortamına entegre etmek
  - Planlama
  - ML Operasyonları
  - Kolay dağıtım
  - REST (web servisi oluşturma) mimarisi
  - Denetim araçları
- Kapalı kaynak
- Yıllık lisans

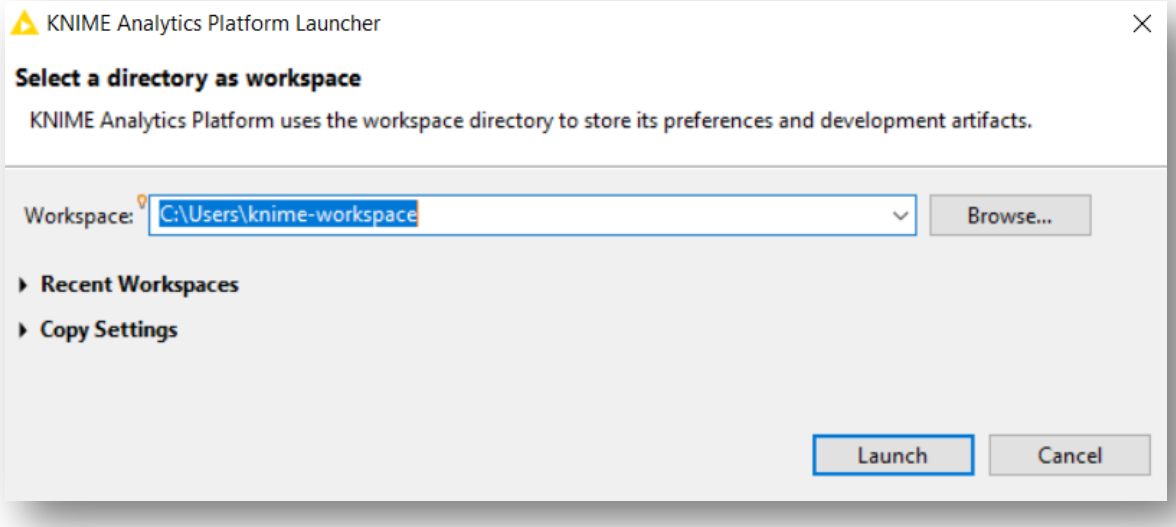
Kurulum

https://www.knime.com/downloads

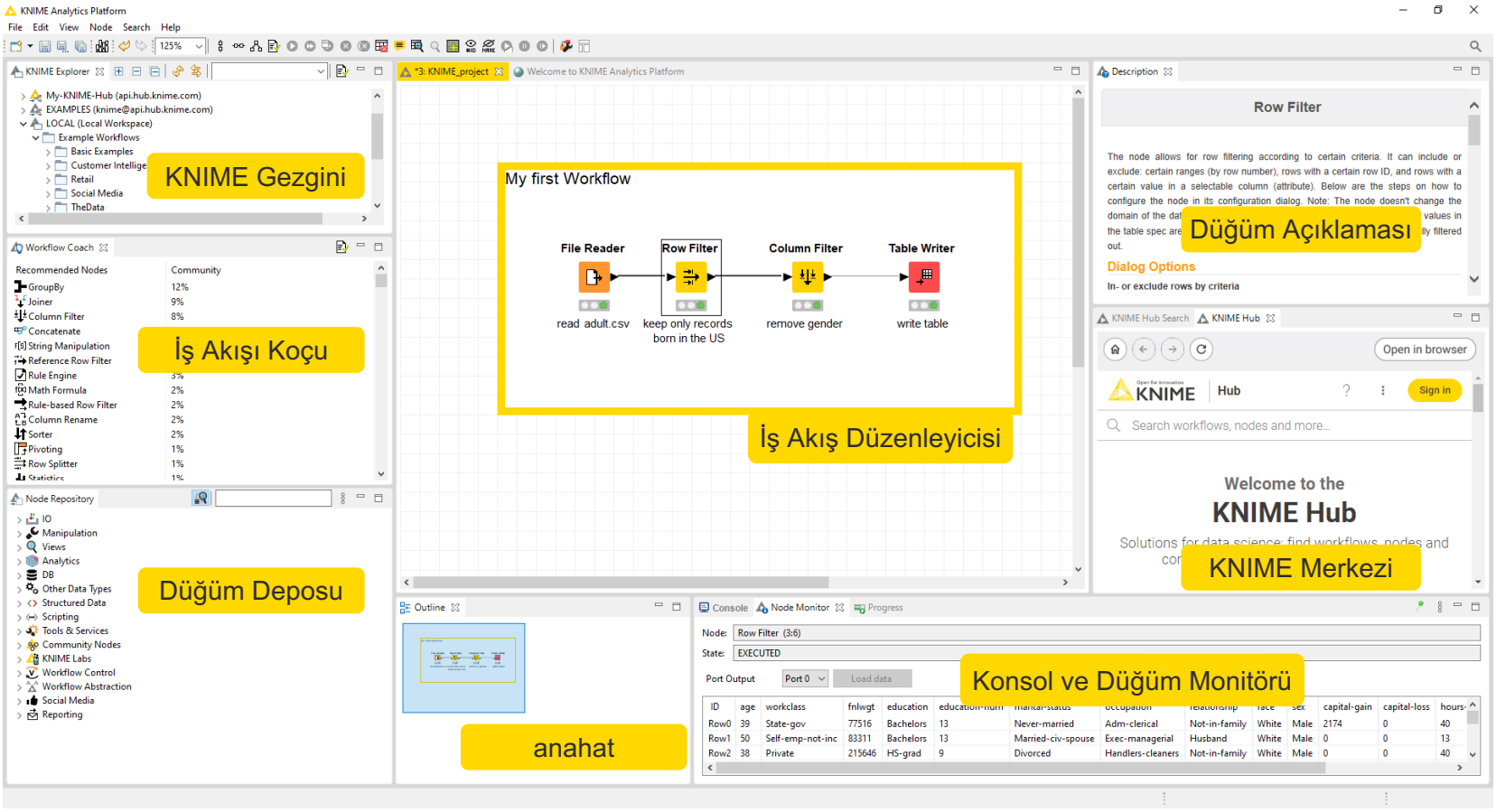
- Bilgisayarınız için KNIME Analytics Platform sürümünü seçin
  - Mac
  - Windows - 32 veya 64 bit
  - Linux
- İşletim sisteminize uygun kurulum dosyasını indirin ve bilgisayarınıza yükleyin
- «Knime»ı çalıştırın

KNIME Çalışma Alanı

- Çalışma alanı, geçerli oturum için iş akışlarının (ve potansiyel olarak veri dosyalarının) depolandığı **klasör/dizindir .**
- Çalışma alanları taşınabilirdir (tıpkı KNIME Analytics Platformu gibi)



KNIME Analitik Platformu



KNIME Analitik Platformu

KNIME Çalışma Alanı

Üst Menü : Dosya (File), Düzenle (Edit), Görünüm (View), Düğüm (Node), Yardım (Help)

Araç Çubuk (Tool Bar): Yeni, Kaydet (Farklı Kaydet, Tümünü Kaydet), Geri Al/Yinele, Raporu Aç (raporlama yüklüyse), Seçili düğümleri dikey/yatay olarak hizala, yakınlaştır (% cinsinden), Otomatik düzen, Yapılandır, Yürütme seçenekleri, Yürütme seçeneklerini iptal et, Sıfırla, Düğüm adını ve açıklamasını düzenle, Düğümün ilk çıkış bağlantı noktası tablosunu aç, Düğümün ilk görünümünü aç, "Meta düğüm Ekle" Sihirbazını aç, , Düğüm adlarına Kimlikler ekle, Tüm düğüm adlarını gizle, Döngü yürütme seçenekleri, İş Akışı Düzenleyicisi Ayarlarını Değiştir, Sarmalanmış Metadüğümlerde Düzeni Düzenle, iş yöneticisini yapılandırın.

KNIME Gezgini (Explorer)

Bu panel, seçilen çalışma alanında (YEREL-LOCAL) veya ÖRNEKLER (EXAMPLES) sunucusunda ya da diğer bağlı KNIME sunucularında kullanılabilen iş akışı projelerinin listesini gösterir.

iş akışı (Workflow) Koçu

Bu bir düğüm öneri motorudur. Şu anda seçili olan düğümü takip etmek için en olası düğümlerin listesini sağlar.

iş Akış Düzenleyicisi Editör (Workflow Editör)

Merkezi alan "İş Akışı Düzenleyicisi" nin kendisinden oluşur.

"Düğüm Deposu" panelinden bir düğüm seçilebilir ve "İş Akışı Düzenleyici" panelinde sürüklenip bırakılabilir.

Düğümler, bir düğümün çıkış bağlantı noktasına tıklanarak ve fareyi bir sonraki düğümün giriş bağlantı noktasında veya bir sonraki düğümün kendisinde serbest bırakarak bağlanabilir.

Düğüm Açıklaması (Node Description)

"İş Akışı Düzenleyicisi"nde veya "Düğüm Deposu"nda bir düğüm seçiliyse, bu panelde seçili düğümün işlevlerinin özet bir açıklaması görüntülenir.

Knime Merkezi (Hub)

Knime Merkezi web sitesine bağlanılır.

Düğüm deposu (Node Repisotory)

Bu panel, KNIME kurulumunuzda bulunan tüm düğümleri içerir. Bir raporda veya bir web tasarımcısı yazılımıyla çalışırken bir araç paletine benzer bir şeydir. Orada grafik araçları kullanıyoruz, KNIME'de ise veri analizi araçlarını kullanıyoruz.

Anahat (outline)

"Anahat" paneli, "İş Akışı Düzenleyicisi" nin içeriğine küçük bir genel bakış içerir. "Anahat" paneli, küçük iş akışları için çok fazla ilgi çekici olmayabilir. Ancak, iş akışları kayda değer bir boyuta ulaşır ulaşmaz, iş akışının tüm düğümleri artık kaydırılmadan "İş Akışı Düzenleyicisi" nde görünmeyebilir. Örneğin, "Anahat" paneli, yeni oluşturulan düğümleri bulmanıza yardımcı olabilir.

Konsol (Console) ve Düğüm (Node) Monitörü

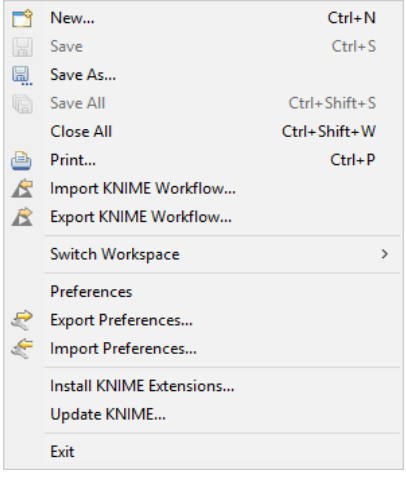
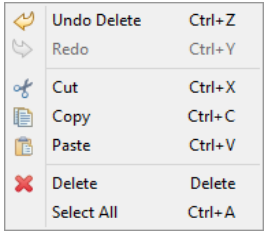
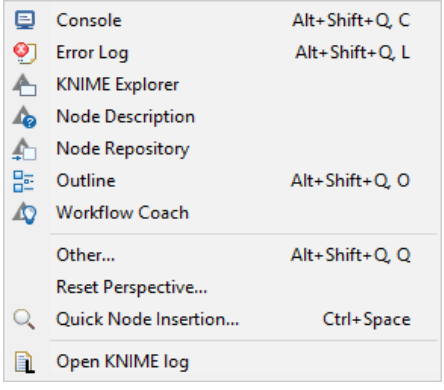
"Konsol" panelinde kullanıcıya hata ve uyarı mesajları görüntülenir.

Bu panel ayrıca günlük dosyasının konumunu da gösterir ve konsol tüm iletileri göstermediğinde ilginizi çekebilir.

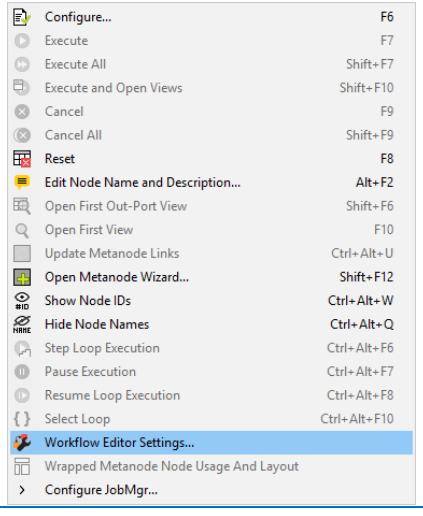
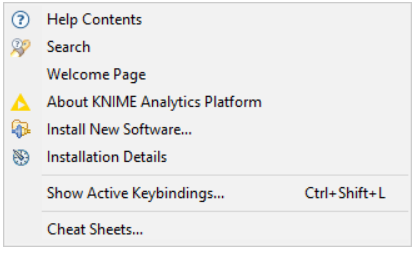
Araç çubuğunda, bu KNIME örneğiyle ilişkili günlük dosyasını göstermek için bir düğme de vardır.

KNIME Üst Menüler-1

Üst Menü

File	Edit	View
		
File, aşağıdakiler gibi bazı KNIME'a özgü komutlara ek olarak "Yeni" ve "Kaydet" gibi geleneksel Dosya komutlarını içerir: örneğin: - KNIME İş Akışını İçe Aktar/Dışa Aktar … - Çalışma Alanını Değiştir - Tercihler - Dışa Aktarma/İçe Aktarma Tercihleri - KNIME uzantılarını yükle - KNIME'ı güncelleme	**Edit** düzenleme komutlarını içerir. **Kes, Kopyala, Yapıştır** ve **Sil** , seçilen düğümlere başvurur. iş akışı. **Select All**, İiş akışı düzenleyicisindeki iş akı düğümlerinin tümünü seçer.	**View** , KNIME platformunda açılabilecek tüm panellerin (pencerelerin) listesini içerir. Kapalı bir panel burada yeniden açılabilir. Ayrıca, panel düzensizliği bozulduğunda, "**Reset Perspective**" seçeneği, ilk kez başlatıldığında KNIME'nin orijinal panel düzenini yeniden oluşturur. "**Other**" seçeneği, özelleştirmek için yararlı olan ek görünümleri açar. tezgah.

KNIME Üst Menüler-2

Node (Düğüm)	Help
	
Node menüsü içinde , bir düğüm üzerinde gerçekleştirilebilecek tüm olası işlemleri gerçekleştirecek seçenekler yer alır. Bunlar şunlar olabilir - :Yapılandırma (Configured) - Uygulama (Execute) - İptal etme (Cancel) ("Execute" sırasında durdurma) - Sıfırla (Reset) (Sson "Execute" işlemi sonuçlarını sıfırla) - Bir isim ve açıklama girme - Görünümü ayarlama (eğer hiç) Seçenekler yalnızca mümkünse etkindir. Örneğin, başarıyla yürütülmüş bir düğüm, önce sıfırlanmadıkça veya yapılandırması değiştirilmedikçe yeniden yürütülemez. "**Cancel**" ve "**Execute**" seçenekleri etkin değildir. "**Open Meta Node Wizard**" seçeneği, iş akışı düzenleyicisinde yeni bir meta düğüm oluşturmak için sihirbazı başlatır.	**Help**, KNIME hakkında genel Yardım sağlar **Search** , Yardım istenen konuları veya düğümleri arayın **Install New Software**, KNIME Güncelleme sitelerinden KNIME Uzantılarını kurmanın kapısıdır. **Cheat Sheets** , belirli konularında eğitimler sunar: raporlama aracı, cvs, eklentiler vs. **Show Active Keybindings,** iş akışı editörü için tüm klavye komutlarını özetler

1.9. KNIME Uzantıları İndirme-1

KNIME Analytics Platform açık kaynaklı bir üründür. Her açık kaynaklı ürün gibi, açık kaynak topluluğunun geliştirdiği geri bildirimlerden ve işlevlerden yararlanır. KNIME Analytics Platform için bir dizi uzantı mevcuttur. Tüm ücretsiz uzantıları dahil olmak üzere KNIME Analytics Platform'u indirip yüklediyseniz, Düğüm Deposu panelinde KNIME Labs, Metin İşleme, R Entegrasyonu ve diğerleri gibi ilgili kategorileri göreceksiniz.

Ancak, yükleme sırasında ücretsiz uzantılar olmadan çıplak KNIME Analytics Platformunu yüklemeyi seçtiyseniz, bunları çalışan bir KNIME'ye bir noktada ayrı olarak yüklemeniz gerekebilir.

KNIME Uzantılarını Yükleme

To install a new KNIME extension, there are two options.

1. Üst Menüden "Dosya" -> "KNIME Uzantılarını Yükle" yi seçin, istediğiniz uzantıyı seçin, "İleri" düğmesini tıklayın ve sihirbaz talimatlarını izleyin.

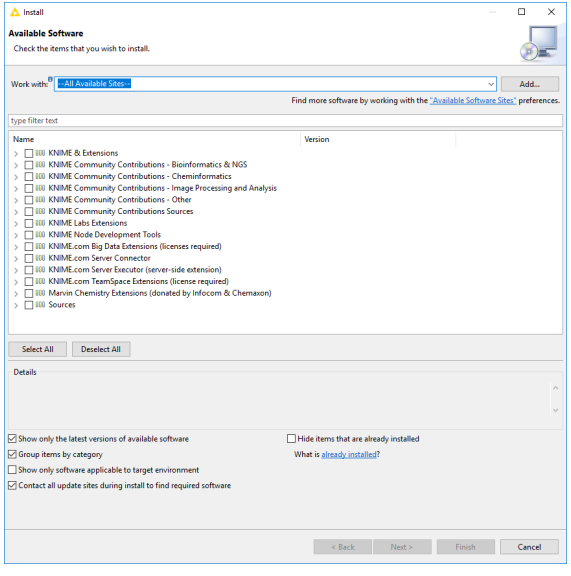
Veya

2. Üst Menüden "**Help**" -> "**Install New Software**"i seçin. "**Available Software**" penceresinde, "**Work With**" metin kutusunda, KNIME güncelleme sitesinin bulunduğu URL'yi seçin (genellikle "KNIME Güncelleme Sitesi" - http://www.knime.org/update/3.x olarak adlandırılır). Ardından uzantıyı seçin, "Next" düğmesini tıklayın ve sihirbaz talimatlarını izleyin.

Seçilen KNIME uzantıları yüklendikten ve KNIME yeniden başlatıldıktan sonra, KNIME platformunun"Düğüm Deposunda" yüklü uzantıya karşılık gelen yeni kategoriyi görmelisiniz.

Örneğin, "KNIME Report Designer uzantısını yükledikten sonra,

"Node Repisotory" panelinde "Reporting" kategorisini görmelidir.



1.11. Alıştırmalar

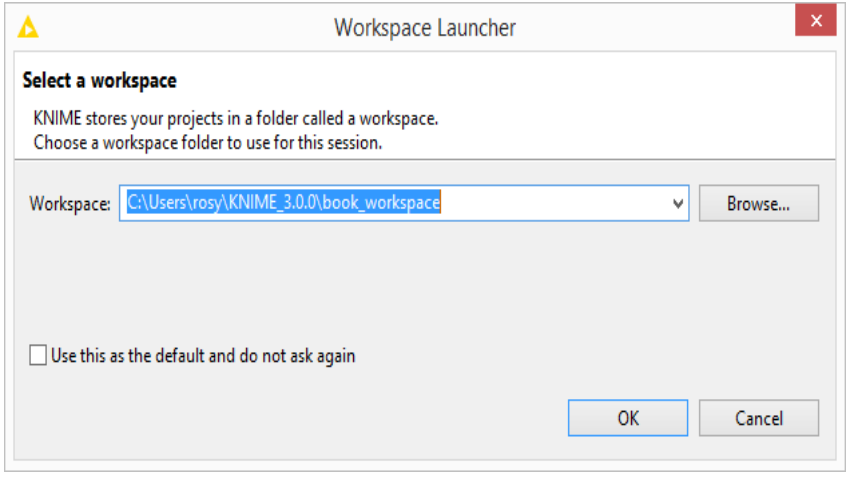
Alıştırma 1

Kendi çalışma alanınızı oluşturun ve "book_workspace" olarak adlandırın. Bu çalışma alanını sonraki iş akışları ve alıştırmalar için kullanabilirsiniz.

Bunu varsayılan çalışma alanınız olarak tutmak için sol alt köşedeki seçeneği etkinleştirin.

- KNIME'yi başlatın
- Workspace Launcher penceresinde "Browse"ı tıklayın
- Yeni çalışma alanınızın yolunu seçin
- ”OK"a tıklayın

Alıştırma 1 Cevap: "book_workspace" çalışma alanı oluşturma



1.11. Alıştırmalar

Alıştırma 2

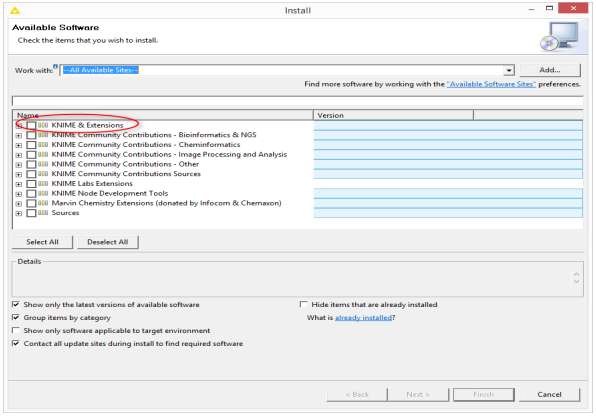
Aşağıdaki uzantıları yükleyin:

- KNIME Math Expression Extension (JEP)
- KNIME External Tool Node
- KNIME Report Designer

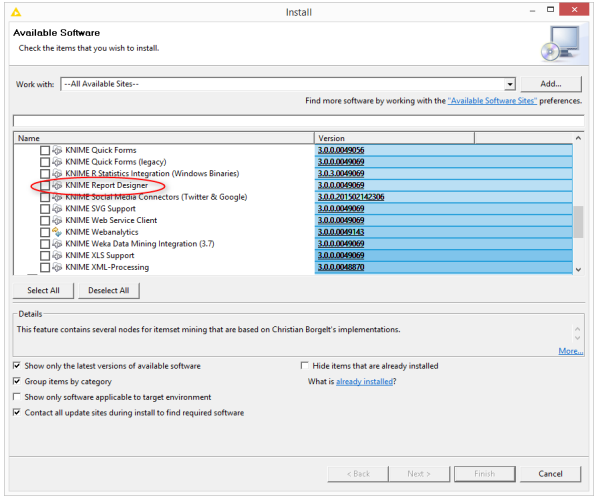
Alıştırma 2 Çözümü

Üst Menüden
“File“ -> “Install KNIME Extensions” ı seçin
Gerekli Uzantıları seçin
”Next"e tıklayın ve talimatları izleyin

List of KNIME Extensions

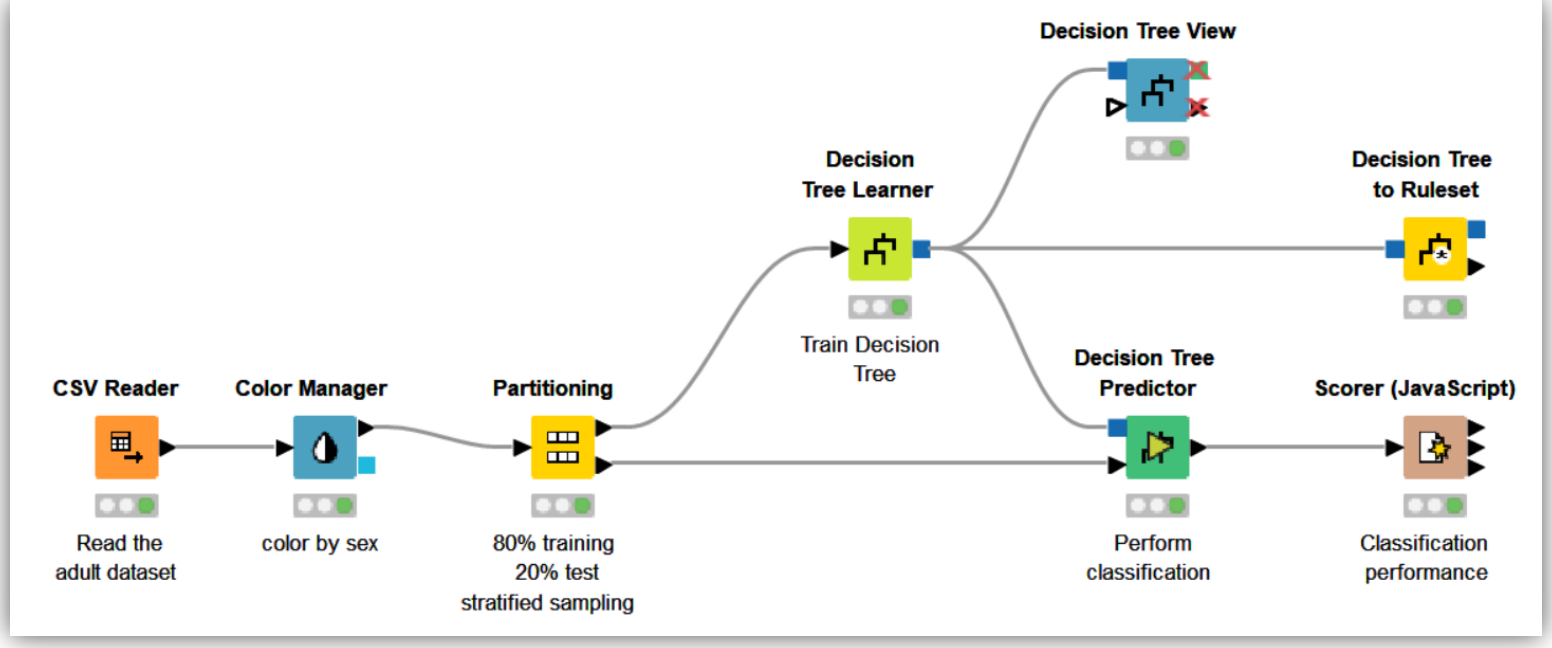


Reporting Extension

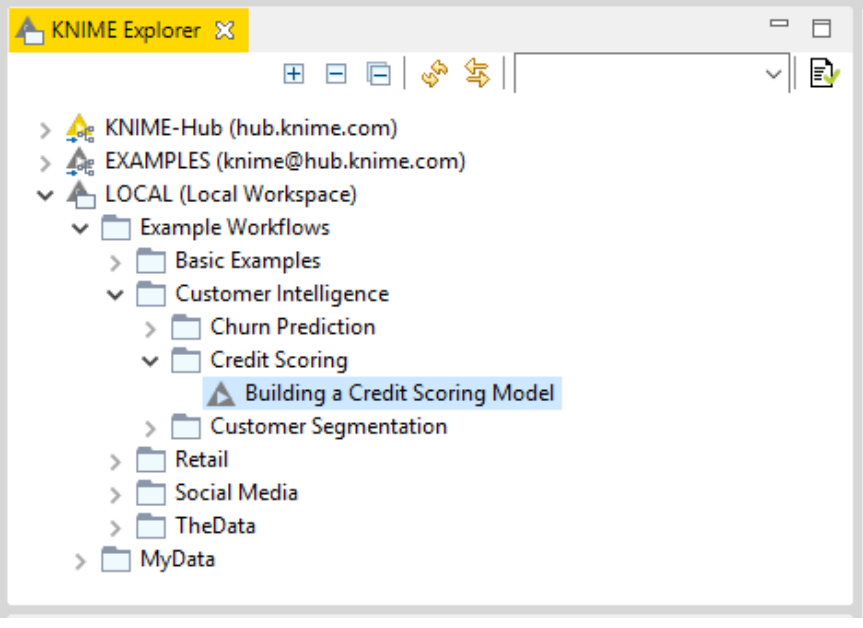


WorkFlow (İş Akışı) Düzenleyicisi

İş akışı, her biri belirli bir görevi gerçekleştirmek için yapılandırılabilen bir düğümler hattıdır.
Veriler düğümlerden soldan sağa doğru akar.



KNIME Explorer (Gezgini)



- Local (Yerel)'de kendi iş akışı projelerinize erişebilirsiniz.
- Diğer bağlantı noktalarını, kullanarak
  - ÖRNEK Sunucusuna
  - KNIME Merkezine
  - KNIME Sunucusuna

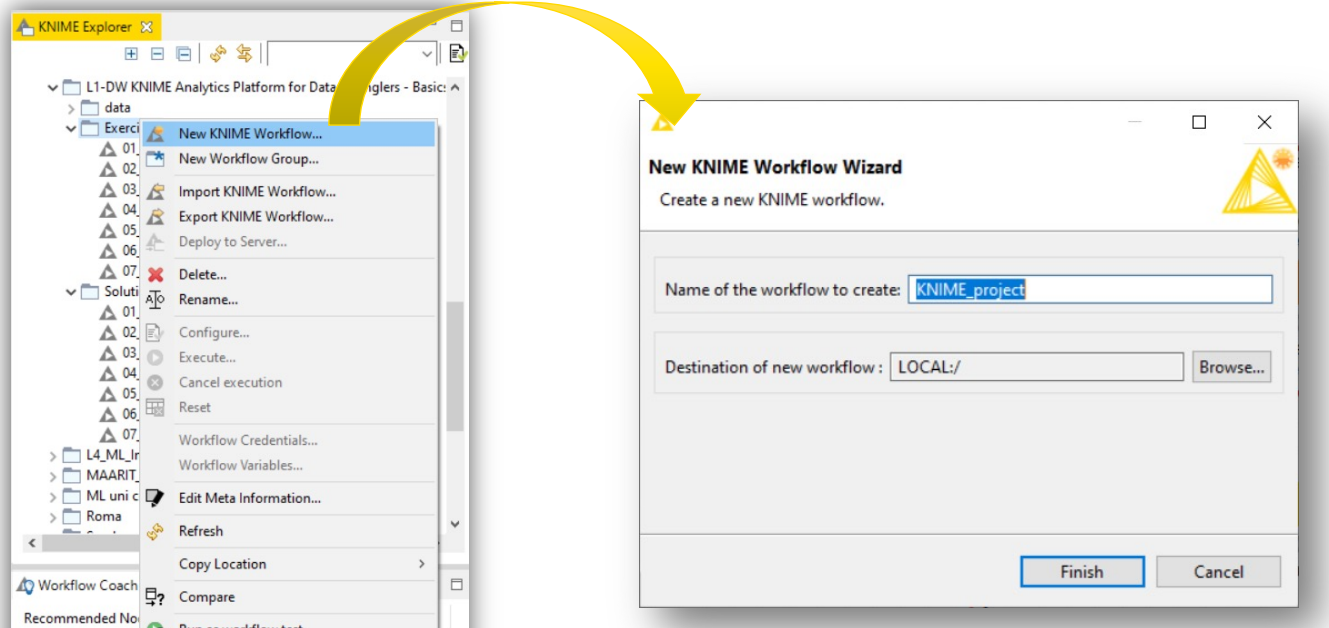
  Erişebilirsiniz.
- Üstteki Knime Explorer araç çubuğunda bir arama kutusu ve arama yapmak için düğmeler bulunur.

  aktif düzenleyicide görüntülenen iş akışını seçin

  görünümü yenile
- KNIME Explorer'da 4 tür içerik yer alabilir:
  - iş akışları
  - İş akışı grupları
  - Veri dosyaları
  - Paylaşılan bileşenler

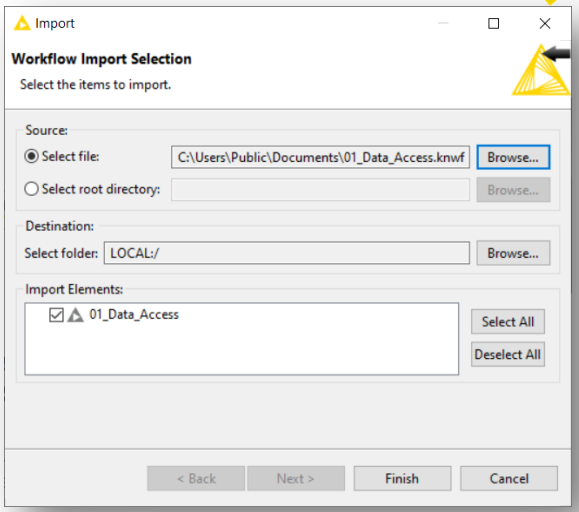
Yeni bir iş akışı oluşturma

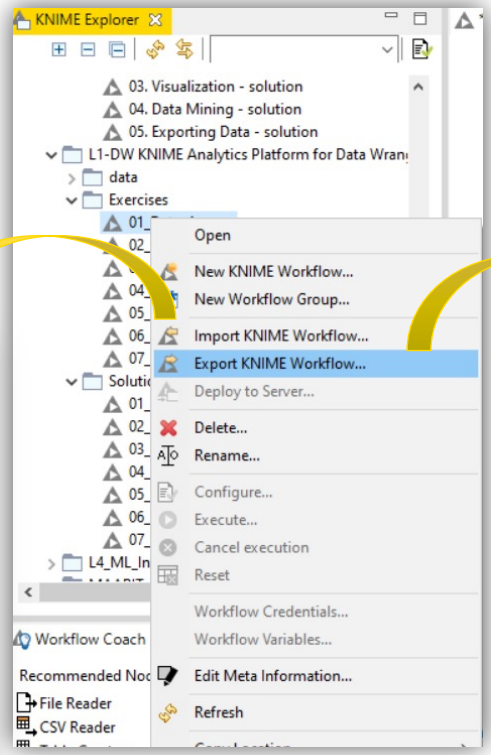
Yeni bir iş akışı veya iş akışı grubu oluşturmak için KNIME Explorer'da herhangi bir yere tıklayın



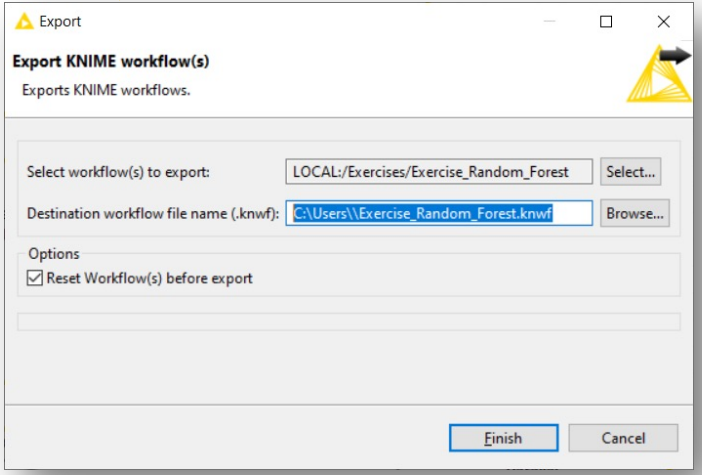
İş Akışlarını İçe ve Dışa Aktarma

Bir iş akışını içe aktarmak için KNIME Explorer'da herhangi bir yere sağ tıklayın

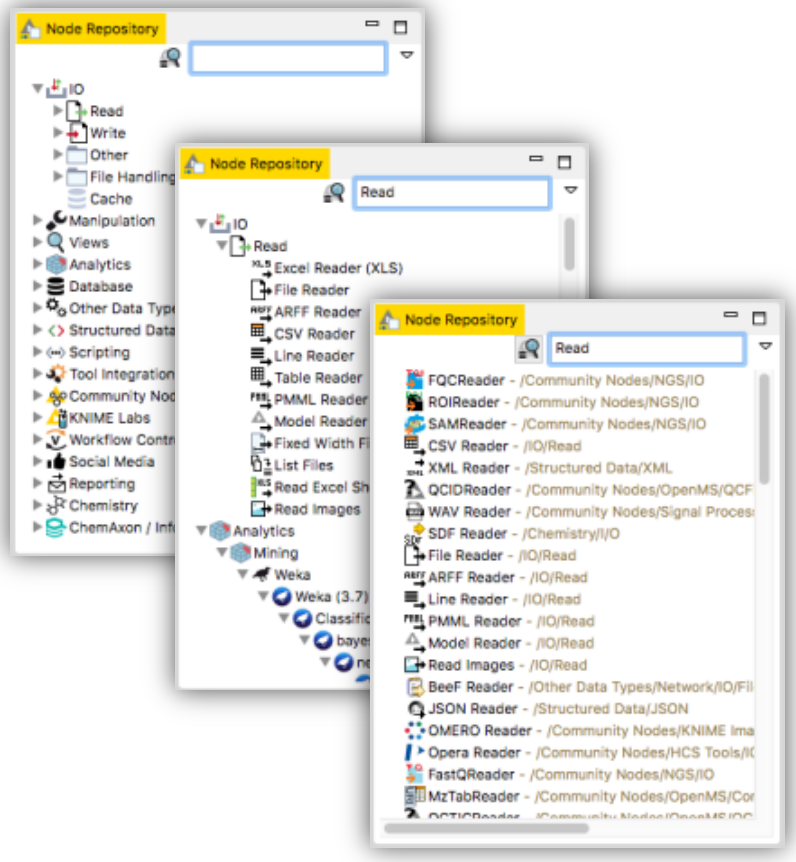




Seçilen iş akışını dışa aktarmak için bir iş akışına veya iş akışı grubuna sağ tıklayın

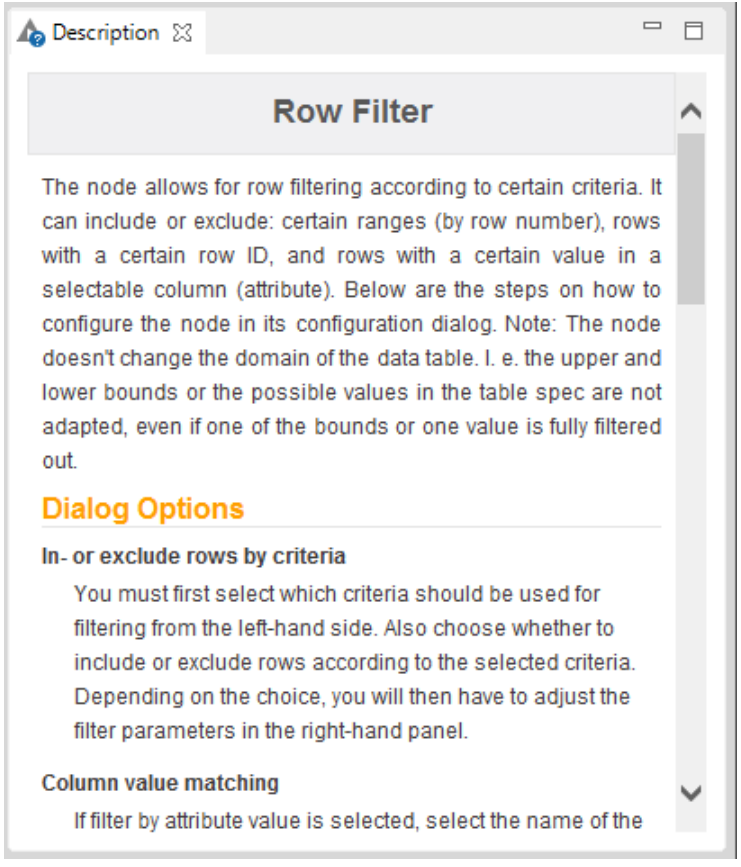


Düğüm Deposu (Node Repository)



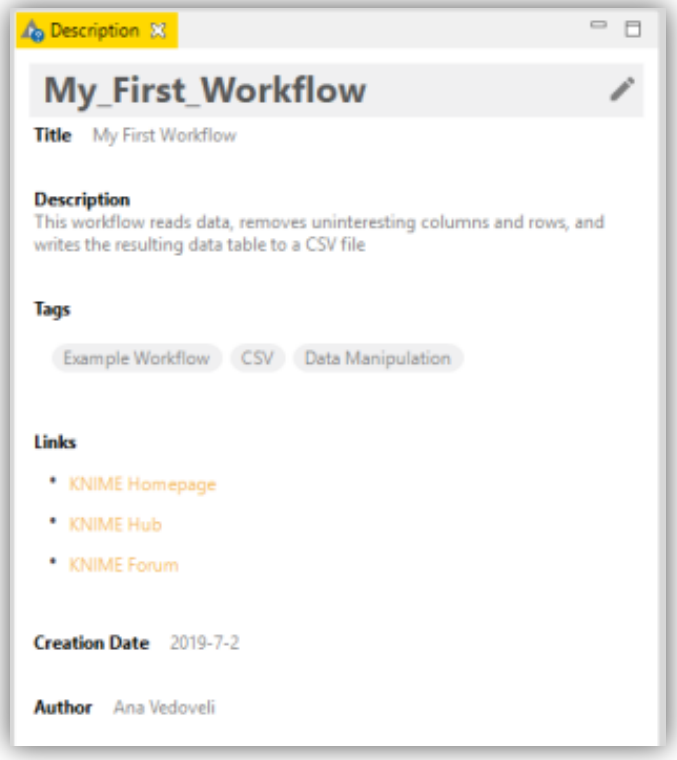
- Düğüm Deposu, diğer alt kategorilerle birlikte kategoriye göre sıralanmış tüm KNIME düğümlerini içerir.
- Uzantı kurulumu, düğüm sayısını makul ölçüde artırabilir
- İki arama yöntemi:
  - Net Arama
  - Bulanık Arama
- Düğümler, Düğüm Havuzundan İş Akışı Düzenleyicisine sürükle ve bırak yöntemiyle eklenebilir

Description (Açılama) Penceresi



- Açıklama penceresi aşağıdakiler hakkında bilgi verir:
  - Düğüm İşlevleri
  - Giriş-çıkış
  - Düğüm Ayarları
  - Portlar
  - Referanslar (Kaynaklar)

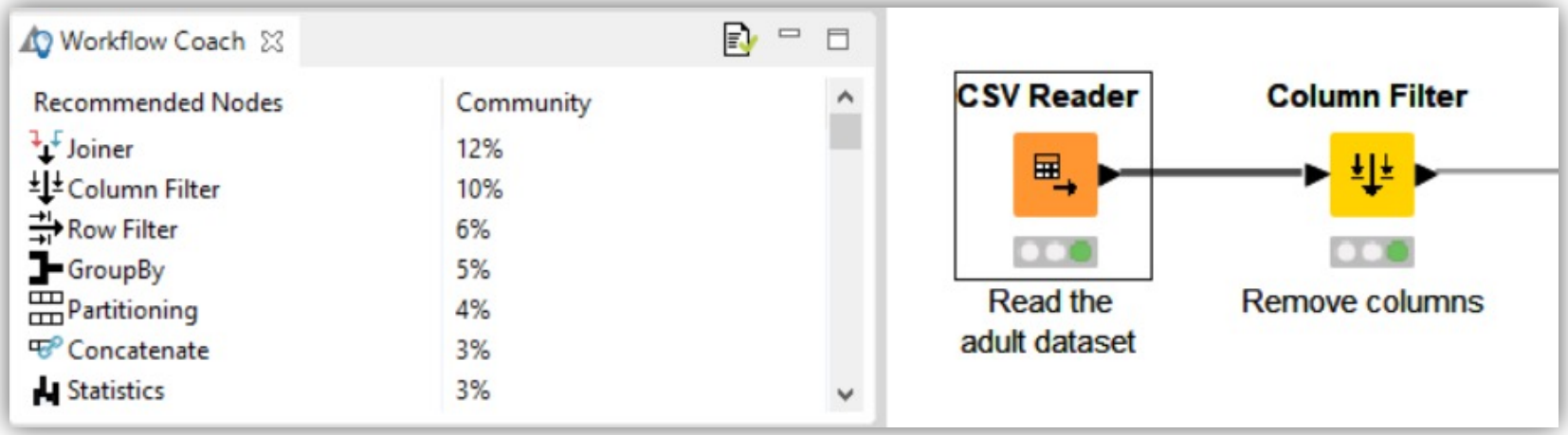
İş Akışı Açıklaması



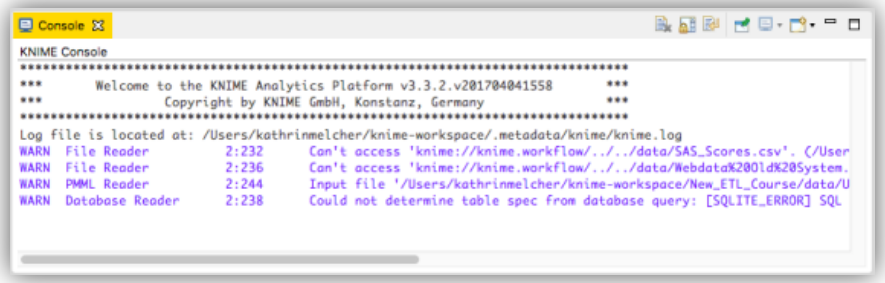
- Bir iş akışı, seçildiğinde, «Açıklama penceresi»nde aşağıdakiler hakkında bilgiler yer a verir:
  - Başlık
  - Tanım
  - İlişkili Etiketler ve Bağlantılar
  - Oluşturulma tarihi
  - Yazar

İş Akışı Koçu

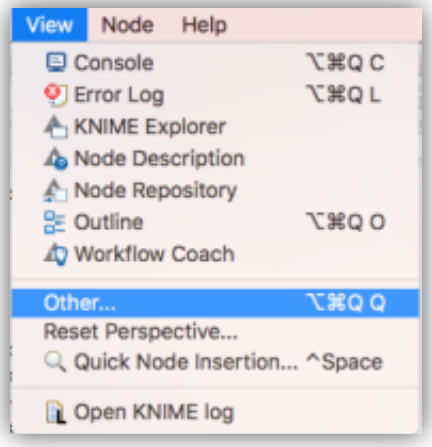
- Düğüm Öneri motoru
- İş akışında bir sonraki adımda hangi düğümün kullanılacağına dair ipuçları verir.
- Dünya çapındaki KNIME topluluk kullanım istatistiklerine dayanmaktadır.
- Kişisel ve yerel grup kullanım istatistiklerini kullanmak için de ayarlanabilir.

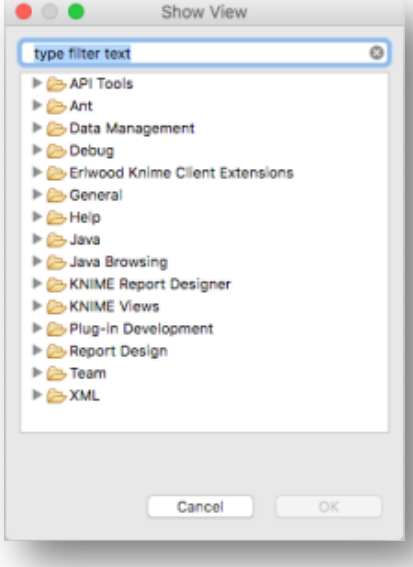


Konsol ve Diğer görünümler



- Konsol penceresi, arka planda neler olduğu hakkında hata ve uyarı mesajları yazdırır

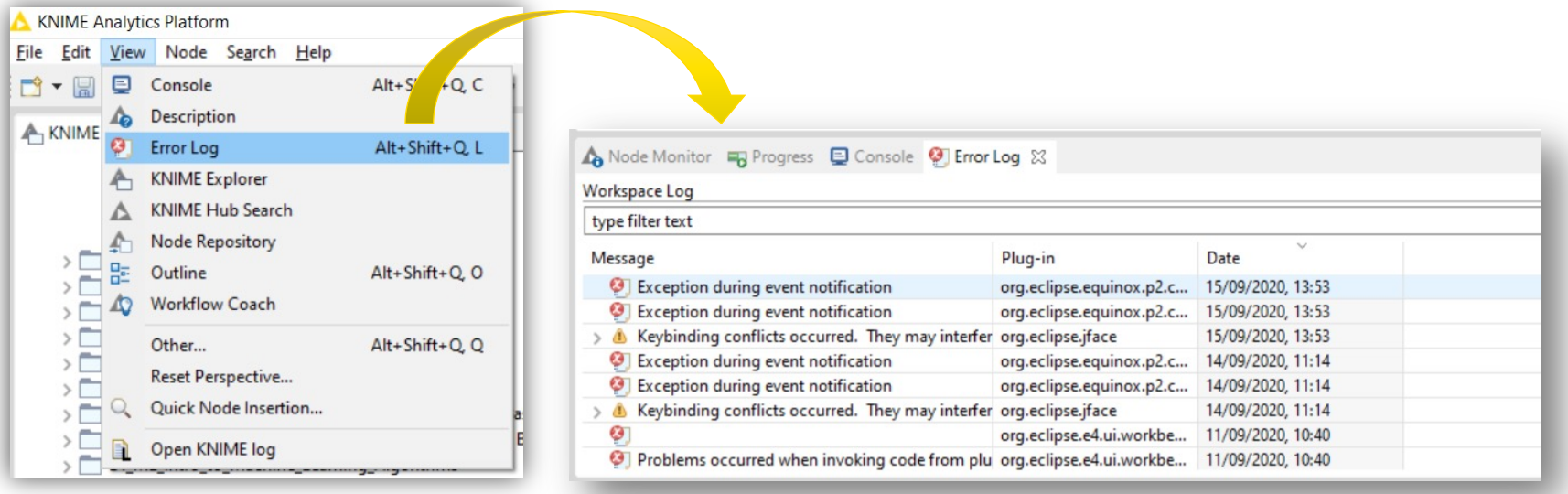




- Ek görünümler eklemek için «View»'e tıklayın ve *Diğer… 'i seçin*

Hata Günlüğü Görünümü

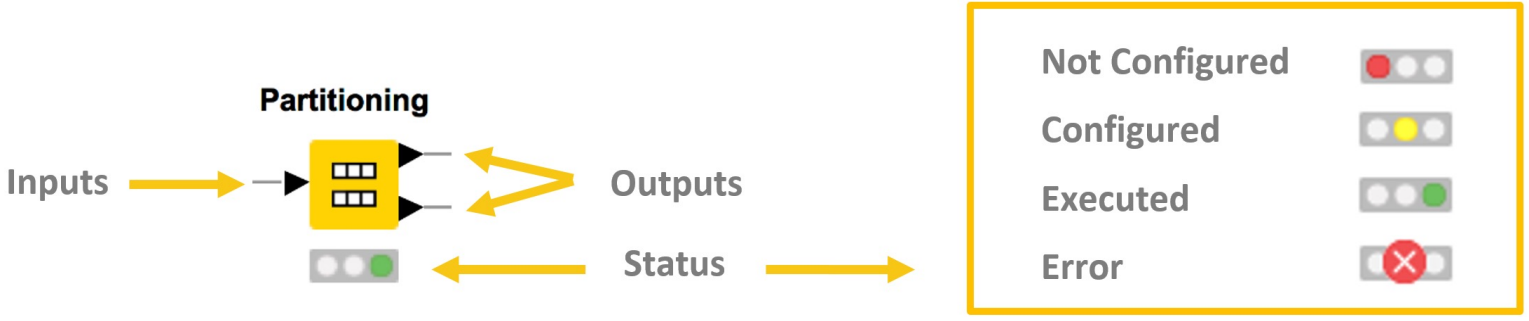
İpucu: Hata Günlüğü (Error Log) görünümünü etkinleştirmek ve kontrol etmek, projenizde hata ayıklarken yardımcı olabilir



Düğümler hakkında daha fazlası

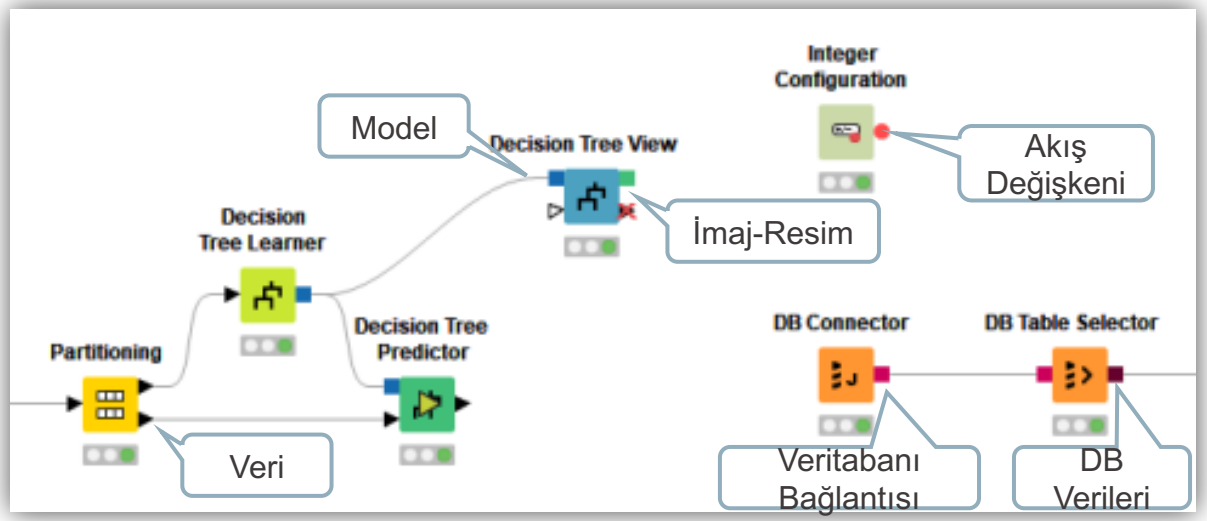
Düğümler hakkında daha fazlası …

- Düğümler, bir iş akışının temel işlem birimleridir.
- Her düğümün bir dizi giriş ve/veya çıkış bağlantı noktası vardır.
- Veriler, bir düğüm üzerindeki çıktı bağlantı noktasından diğer düğümlerin giriş bağlantı noktalarına aktarılır
- Her düğümün altında bir ışık vardır ve düğümün durumunu gösterir



Veri Bağlantı Noktası Türleri

- Bu tür düğümlerden oluşan bir boru hattı, bir **iş akışı oluşturur.**
- Düğümün veriler üzerindeki çalışmasının sonucu, ardıl düğümlere çıkış noktasında sağlanır.
- Sadece aynı tip port bağlanabilir

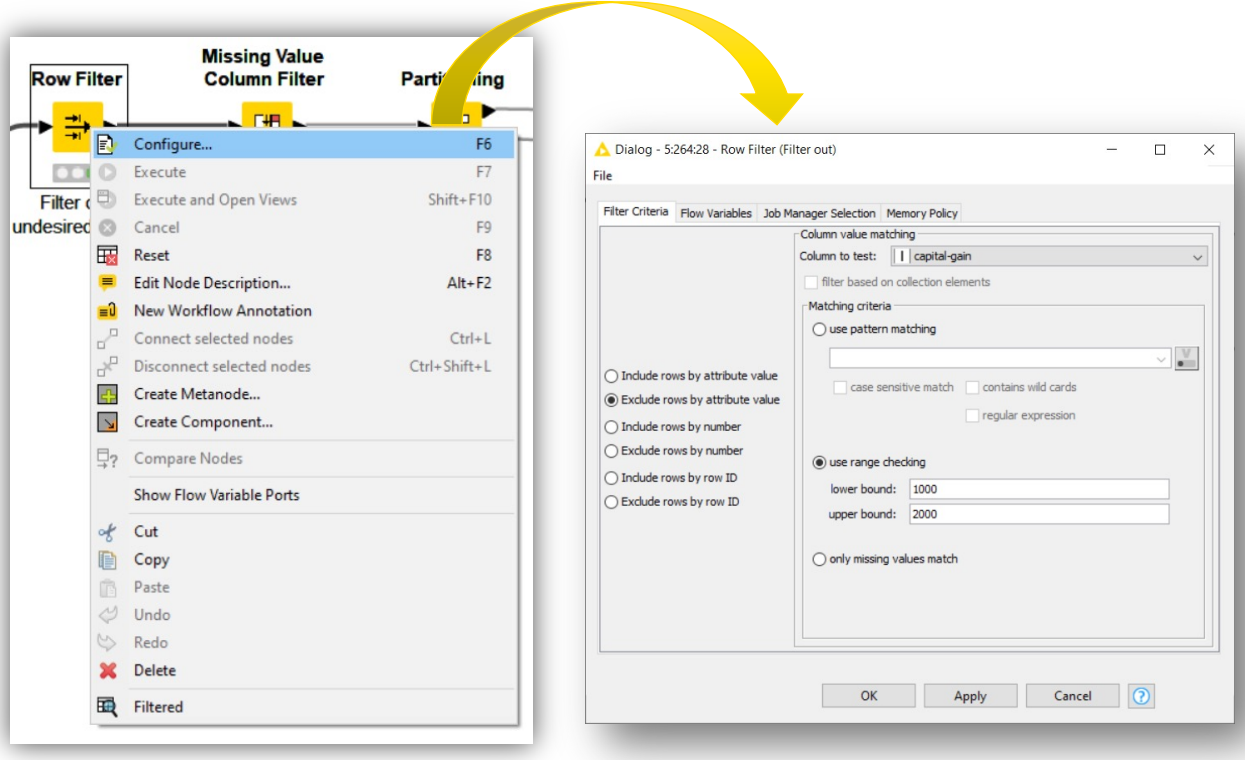


Düğüm Yapılandırması

- Çoğu düğüm yapılandırma gerektirir
- Bir düğüm yapılandırma (configure) penceresine erişmek için:
  - Düğümü çift tıklayın

VEYA

  - Sağ tıklayın > Configure (Yapılandır)

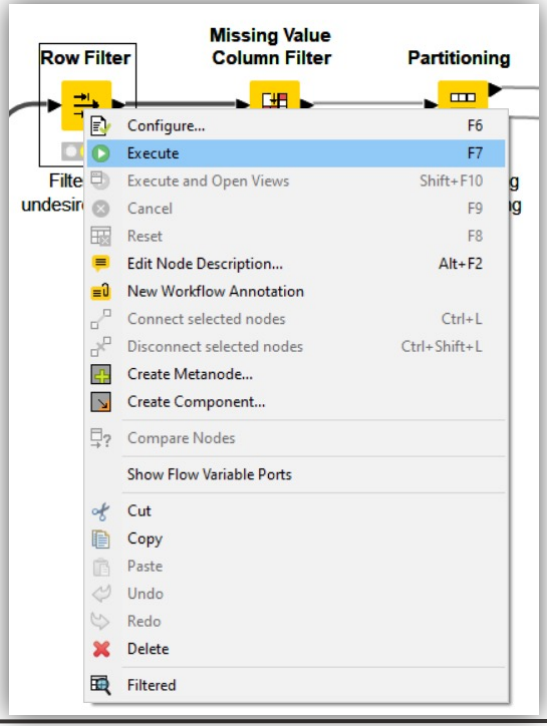


Düğüm Yürütme

- Sağ tıklama ve «Execute» seçeneği

veya

- Araç çubuğunda yeşil «Execute» düğmesini seçin

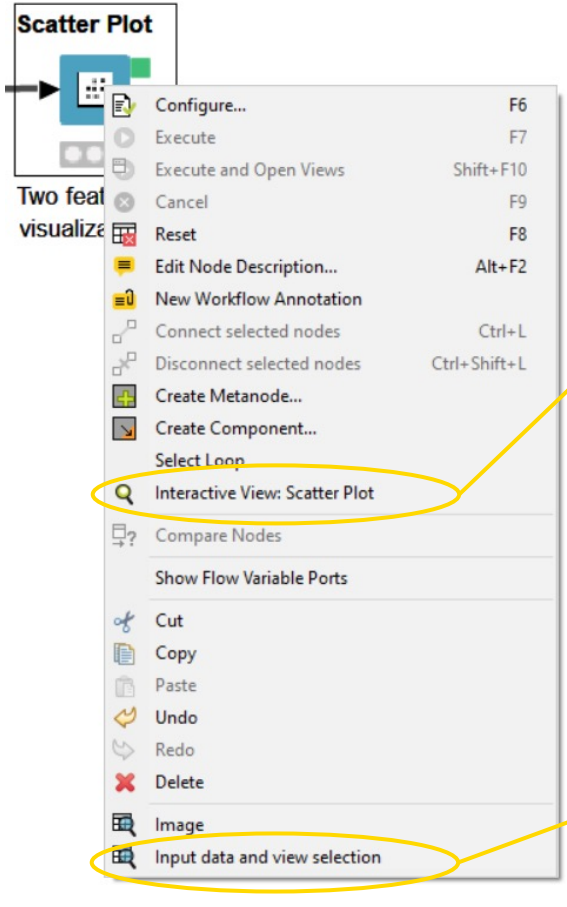


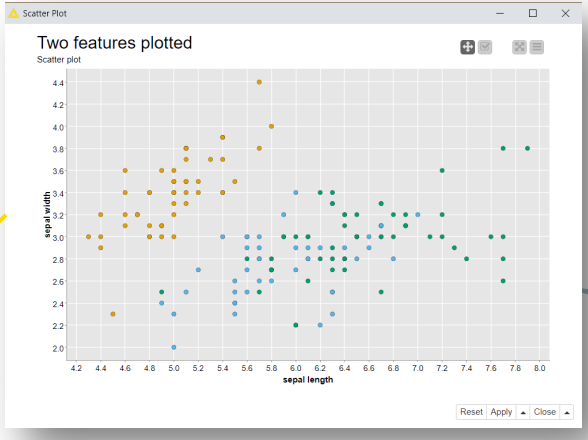
Yürütme başarılıysa durum yeşil ışık gösterir

Yürütme uyarı veriyorsa, durum sarı üçgeni gösterir

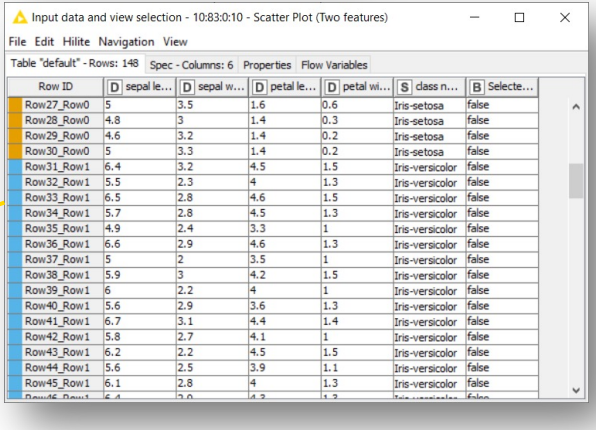
Yürütme hatalarla karşılaşırsa durum kırmızı bir X gösterir

Analiz Sonuçlarını Görme-Düğüm Görünümleri



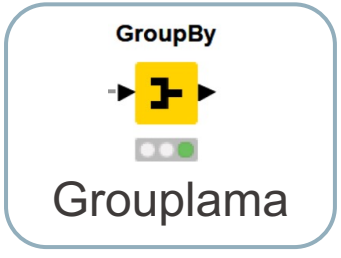


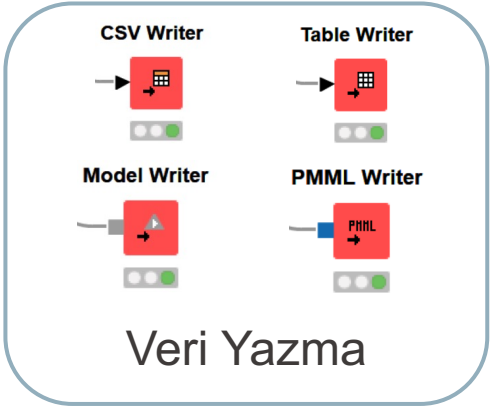
Etkileşimli Görünüm

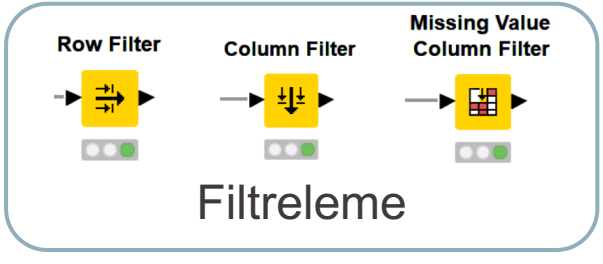


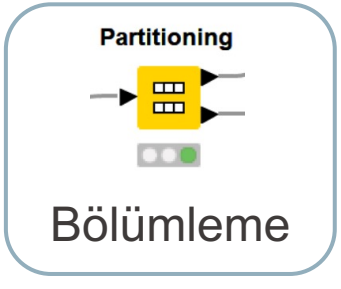
Veri görünümü

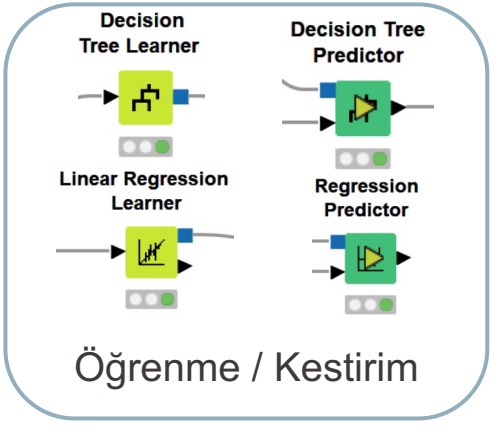
Sık Kullanılan Düğümler





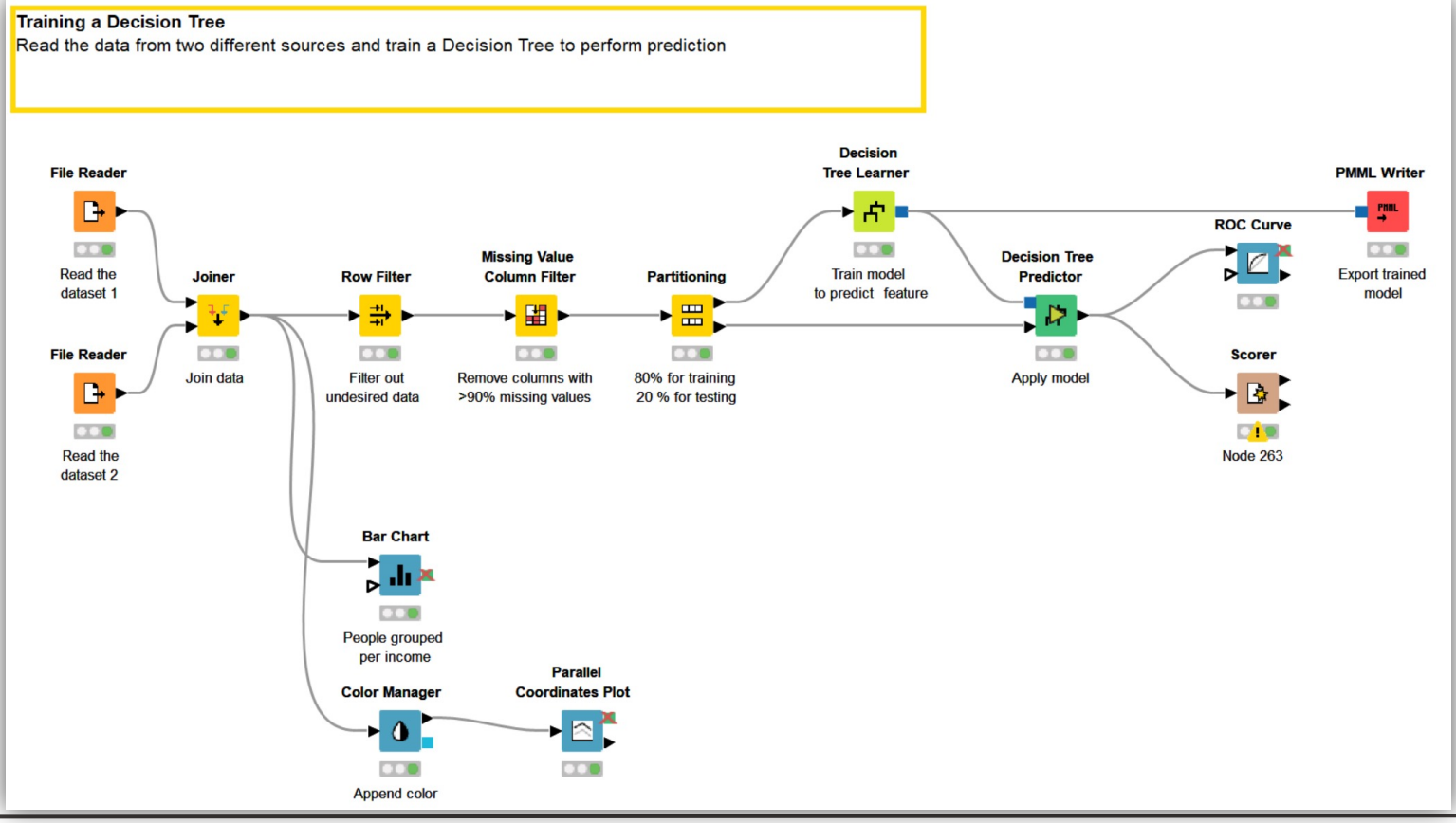






İş akışlarını düzenleyin

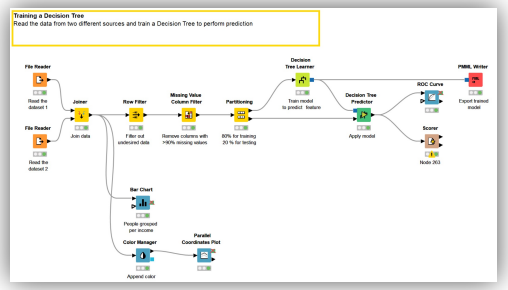
- İş akışı kolayca karmaşık ve anlaşılması zor hale gelebilir

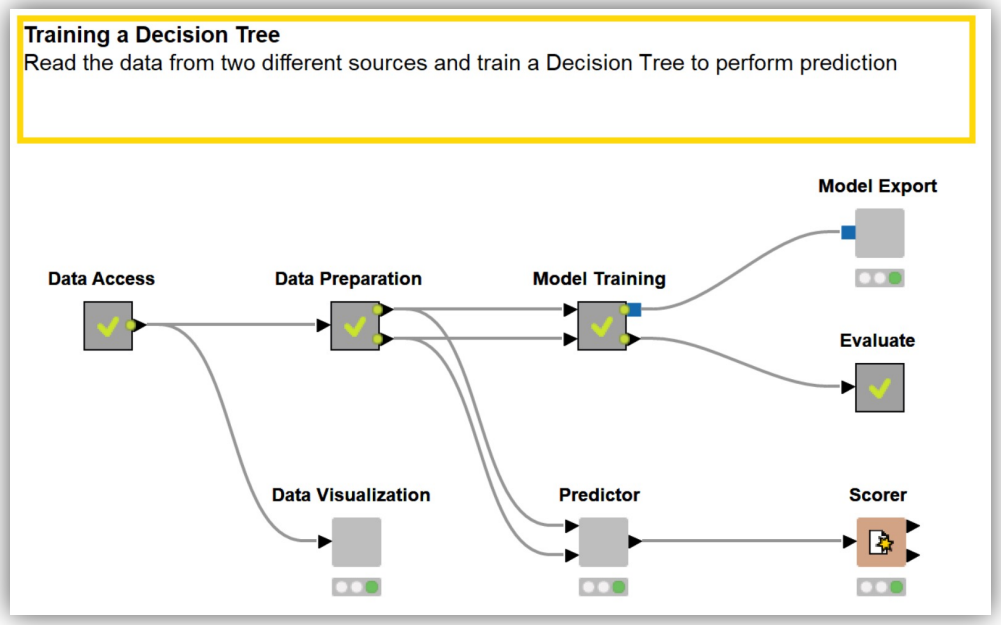


Metanodlar ve Bileşenler

İş akışlarını düzenleyin-MetaNodlar

- Metanodlar ve bileşenler, ortak işlemleri gerçekleştiren düğümleri kapsülleyerek toplamaya yardımcı olabilir

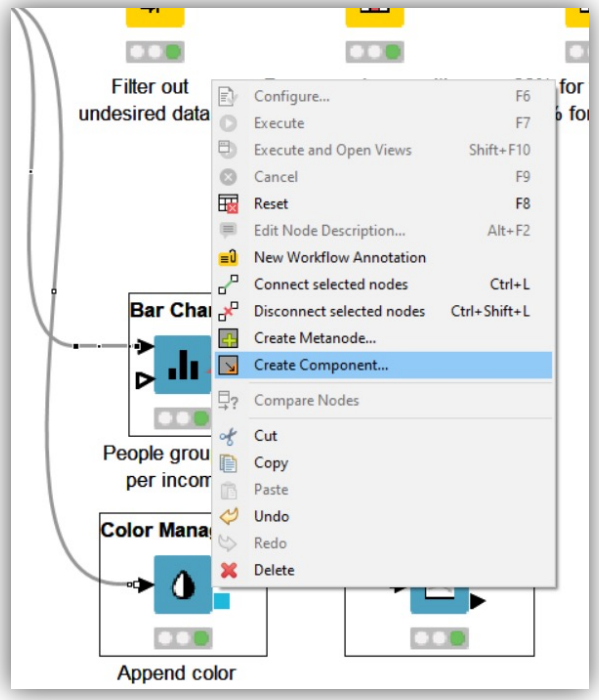




İş akışlarını düzenleyin-Bileşenler

Bir bileşen veya metanod oluşturma adımları

- Gruplamak istediğiniz ilgili düğümleri seçin
- Sağ tık
- *«Create Component…»* öğesini seçin . veya «Create *Metanode…»*
- Bir isim ver

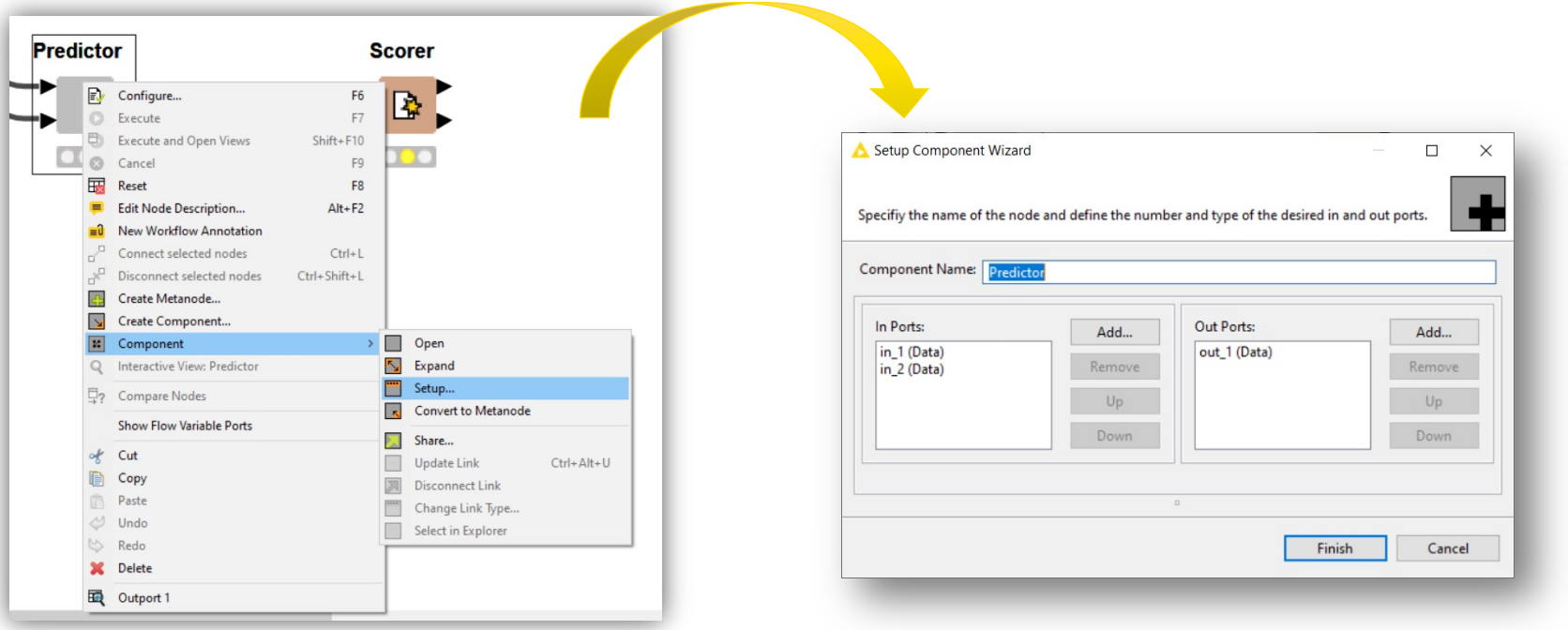


Bileşenler daha gelişmiş özelliklere sahiptir:

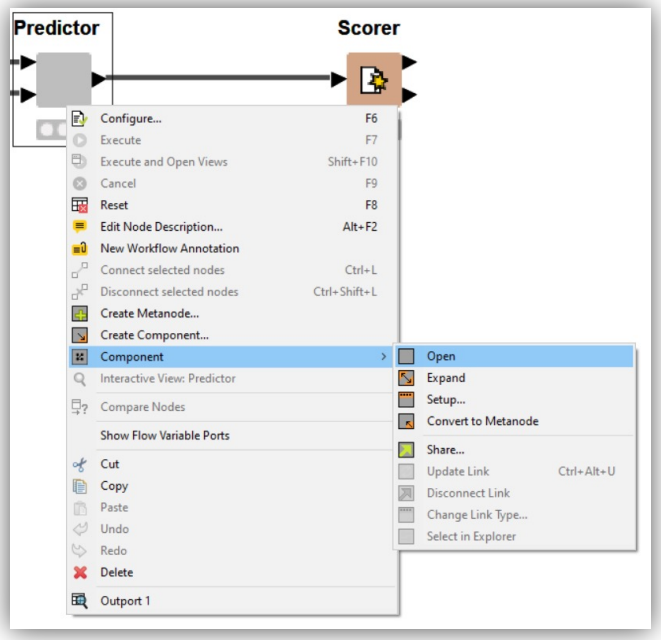
- «Akış değişkenleri»ni kapsülleyin, yani parametreler yalnızca bileşenin içinde işlev görür
- **«Konfigürasyon penceresi»** ile bu gerçekleştirilebilir : bileşen içindeki değişkenler ve parametreler üzerinde Sağ Tıkla -> «Configure»ü seç ve düzenle …
- **Bileşik** bir görünüm oluşturun : Bileşen içindeki bir görselleştirme, gösterge panosunda (Dashboard) gruplandırılabilir.

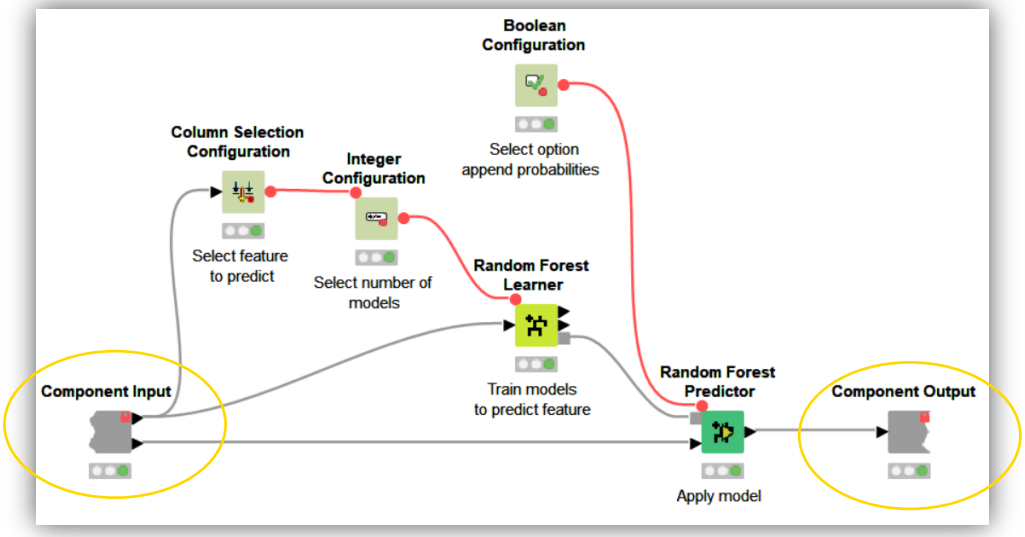
Alt Menü Bileşeni

- Bileşene sağ tıklayın ve bileşen adı ve bağlantı noktaları gibi daha fazla özelleştirme yapmak için «Component» alt menüsünden «Setup»ı işaretleyin



Bir bileşenin içinde



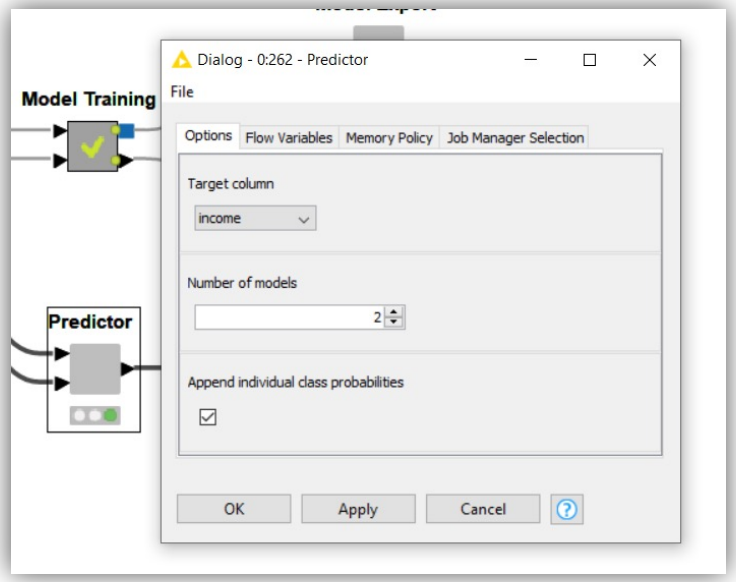


Kısayol:

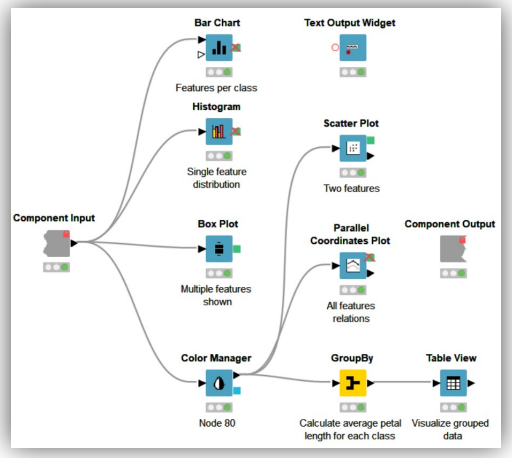
İçeriğini açmak için **Ctrl + bileşene çift tıklayın**

Bileşenler Yapılandırma Penceresi

- Bileşenler yapılandırılabilir
- «Configure» penceresinden (Sağ tıklama -> Configure …) kullanıcı bazı parametreleri girebilir
- Girilen parametreler, bileşen içindeki düğümlerin davranışını değiştirir



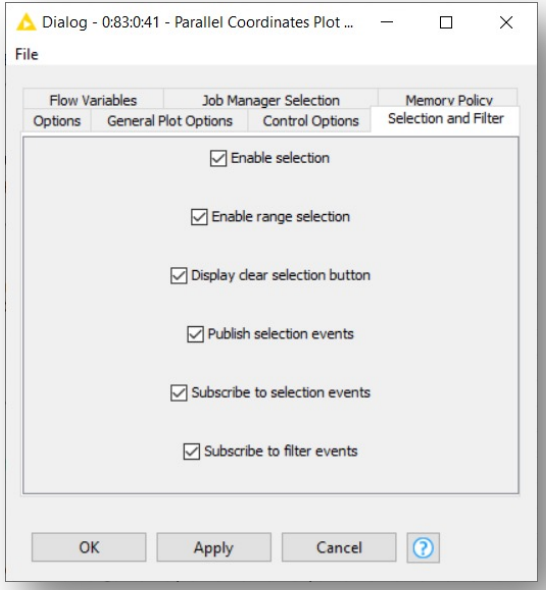
Bileşenler Kompozit Görünüm

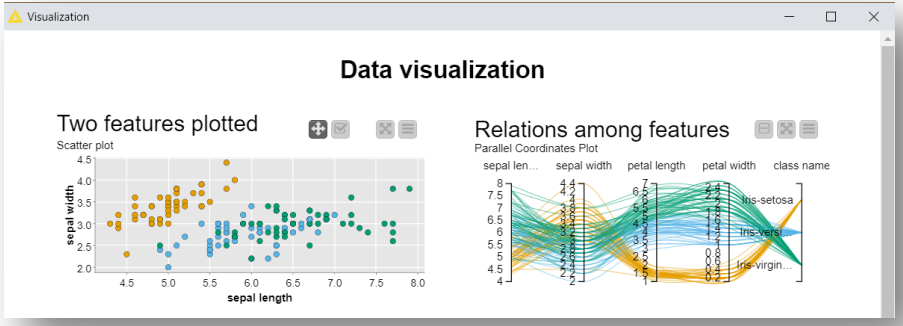


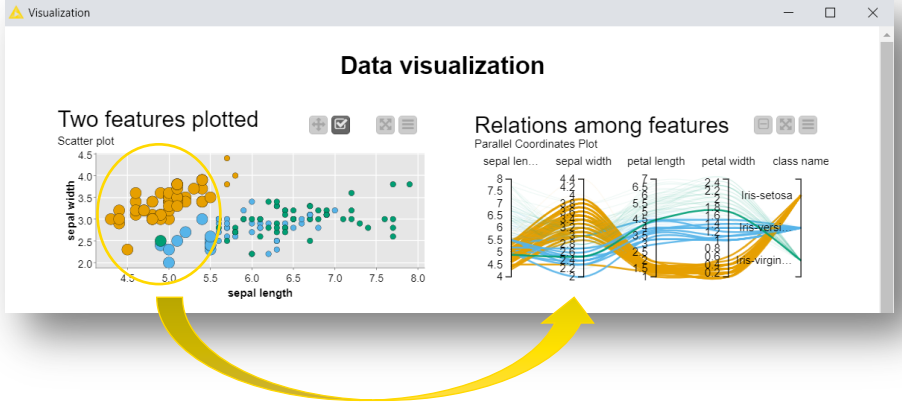
Bileşen içindeki görselleştirme düğümleri, etkileşimli bir bileşik görünüm oluşturmak için düzenlenebilir

Görsel Düzen penceresinden düğüm görünümlerini düzenleyebilir ve yeniden şekillendirebilirsiniz (bileşenin içinden, araç çubuğundaki son simge)

Bileşik görünüm etkileşimi





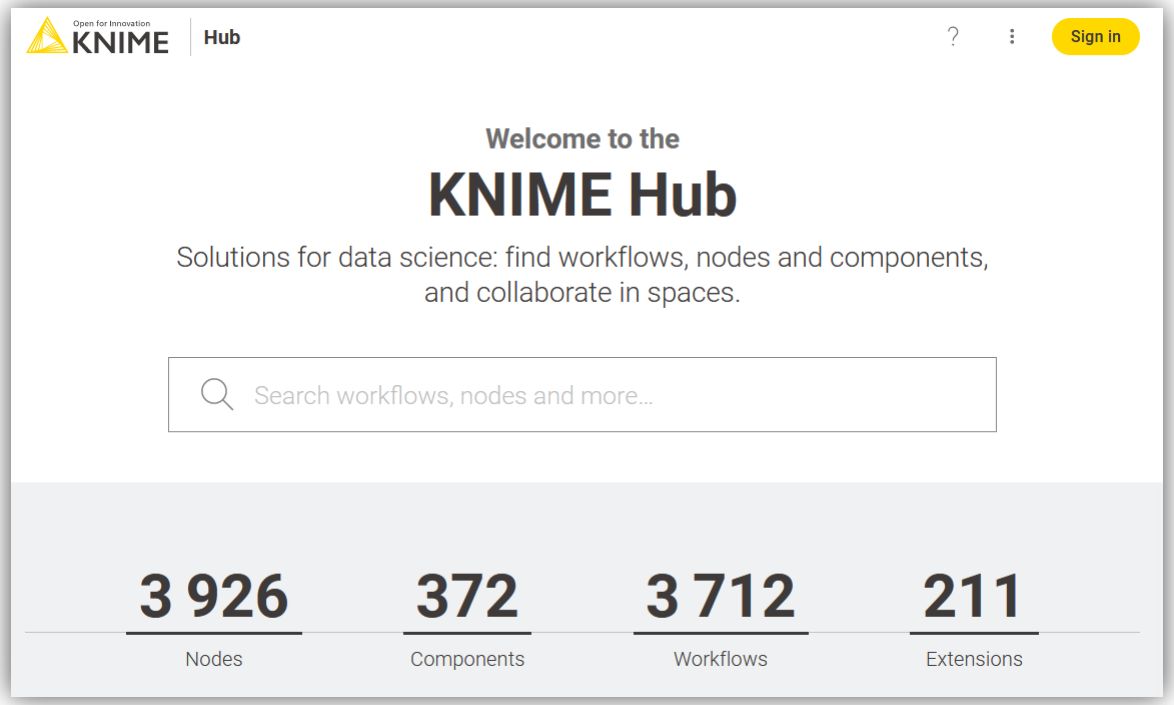


Bileşik görünümü etkileşimli hale getirmek için seçim olaylarını yayınlamayı ve aboneliği etkinleştirin: bir görünümde seçilen veriler diğerlerinde vurgulanır

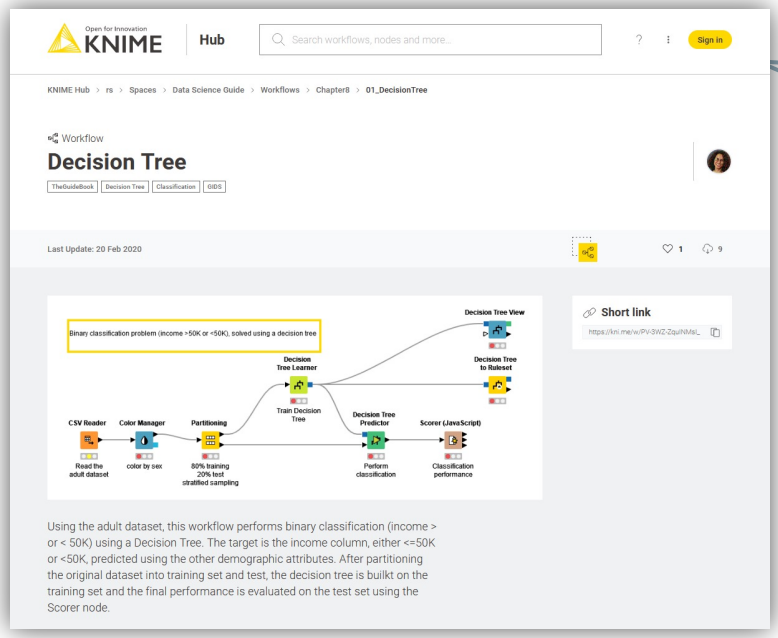
KNIME Merkezi

KNIME Merkezi

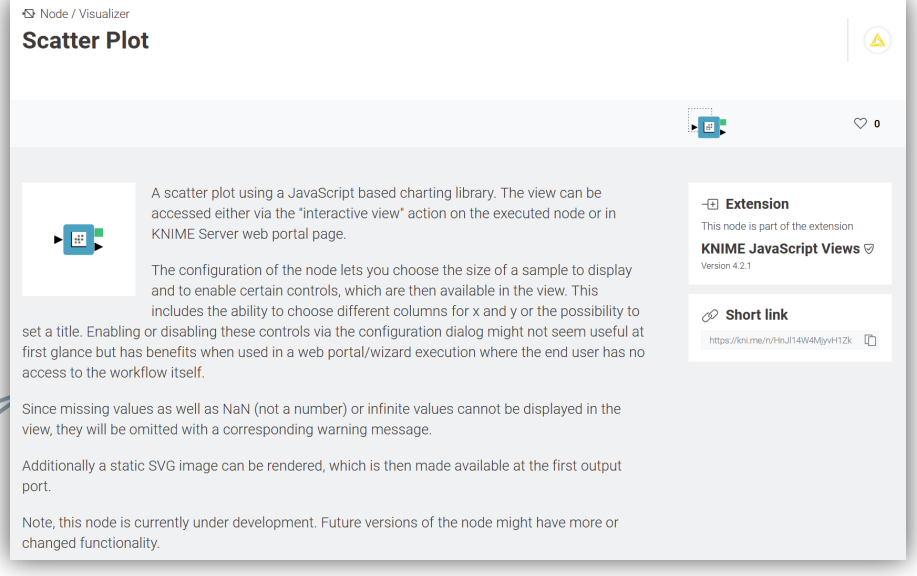
İş Akışları ve Düğümler hakkında bilgi paylaşımı için bir yer https://hub.knime.com



KNIME Merkezi

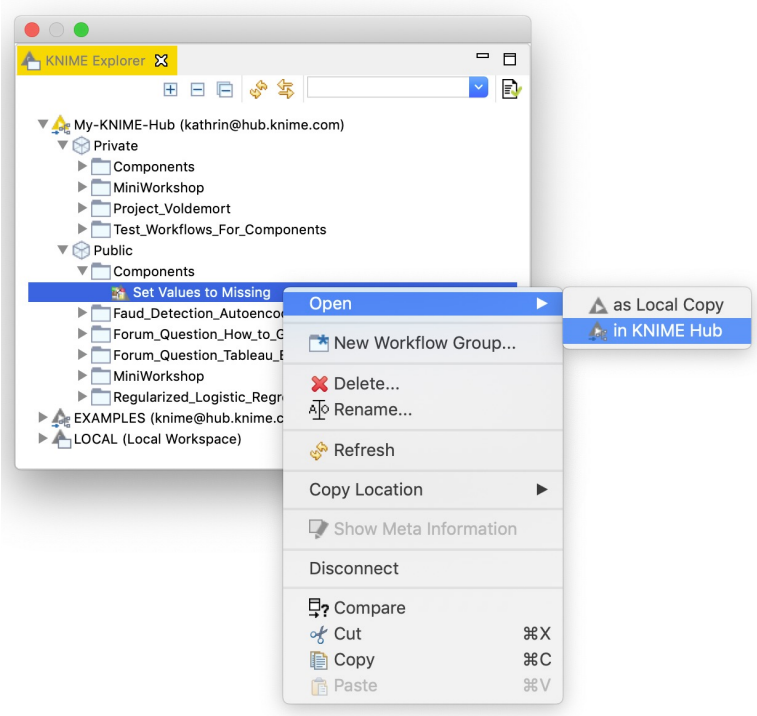


iş akışları



Düğümler, Paylaşılan Bileşenler ve Uzantılar

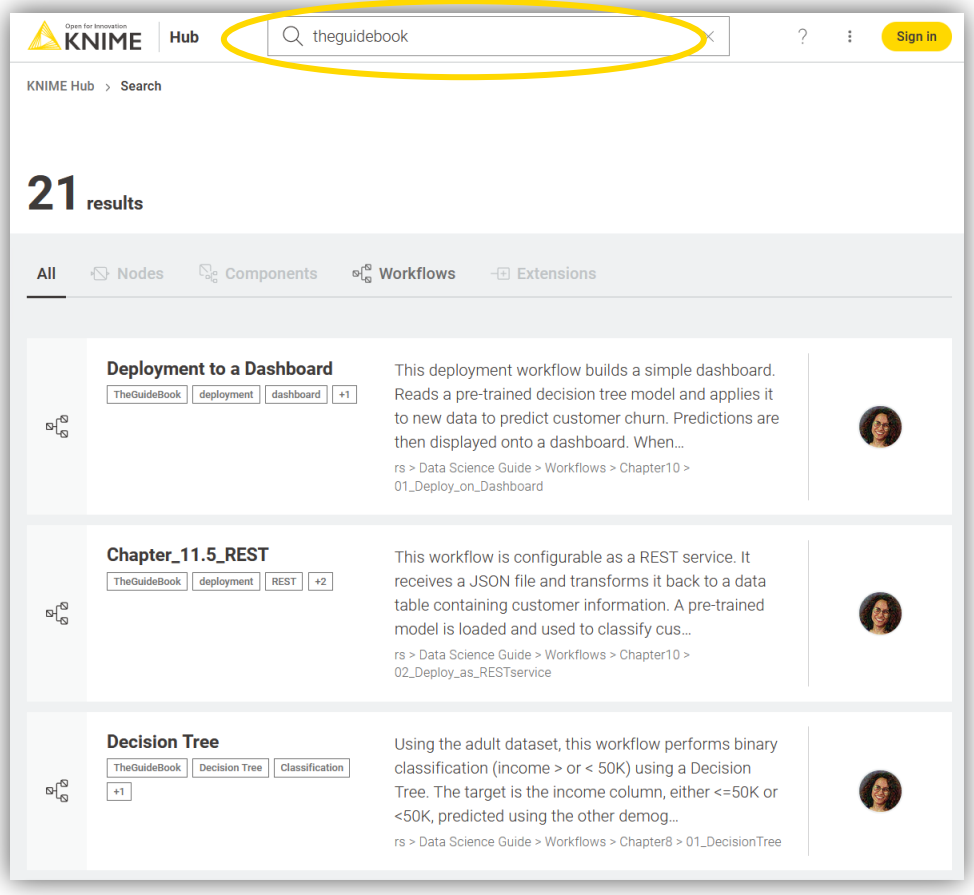
KNIME Hub Alanları



- **Özel alan**
  - Kişisel alanınız. İş akışlarınızı ve bileşenlerinizi (maks. 1 GB) her zaman merkezi bir yerde kullanabilmek için buraya yükleyin

- **Halka açık alan**
  - KNIME topluluğuyla paylaşılır. Herkes bunları KNIME Hub'dan bulabilir ve indirebilir

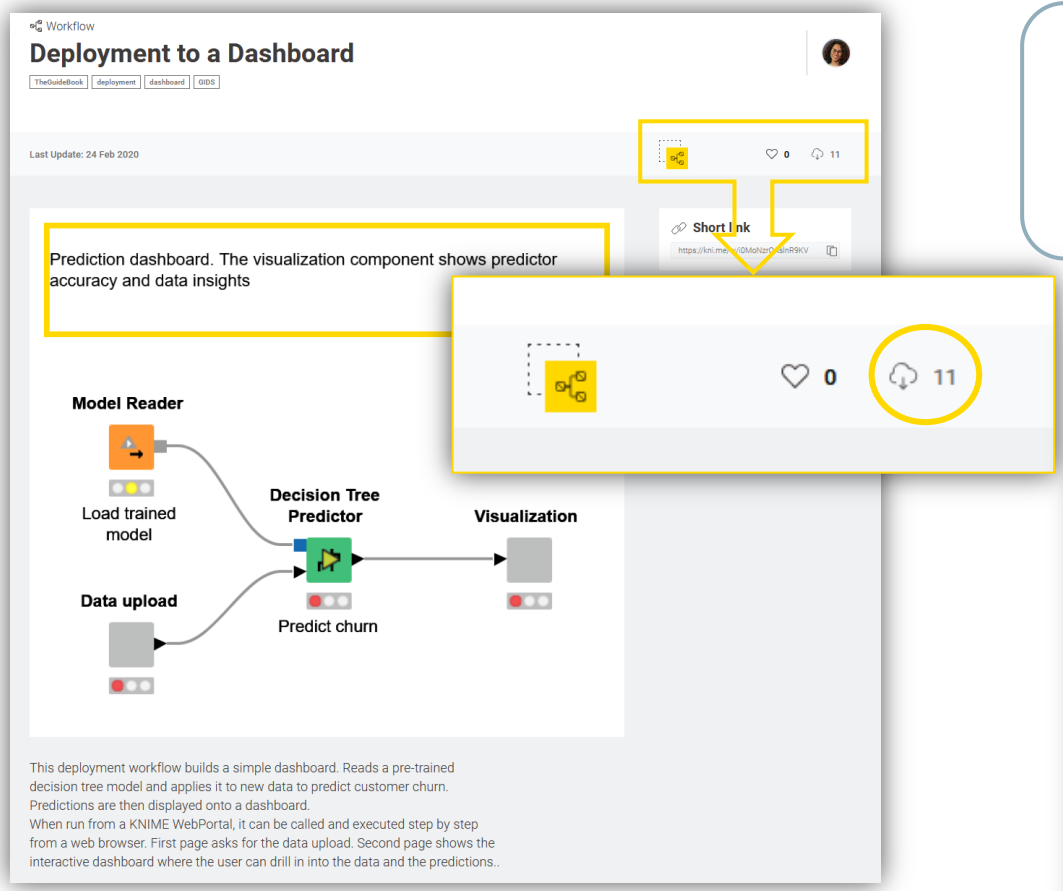
KNIME Hub'dan indirme ve içe aktarma



Etiket ***aramak***,

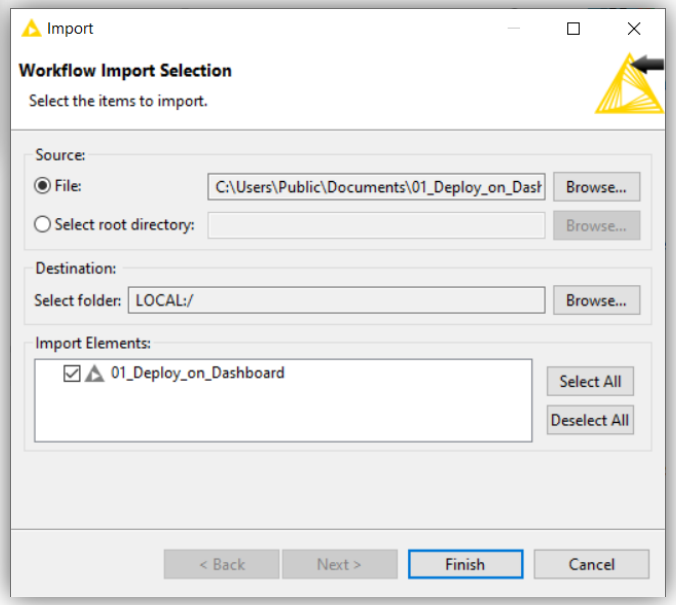
bu kullanım klavuzu ilgili tüm iş akışlarını size gösterecektir.

KNIME Hub'dan indirme ve içe aktarma

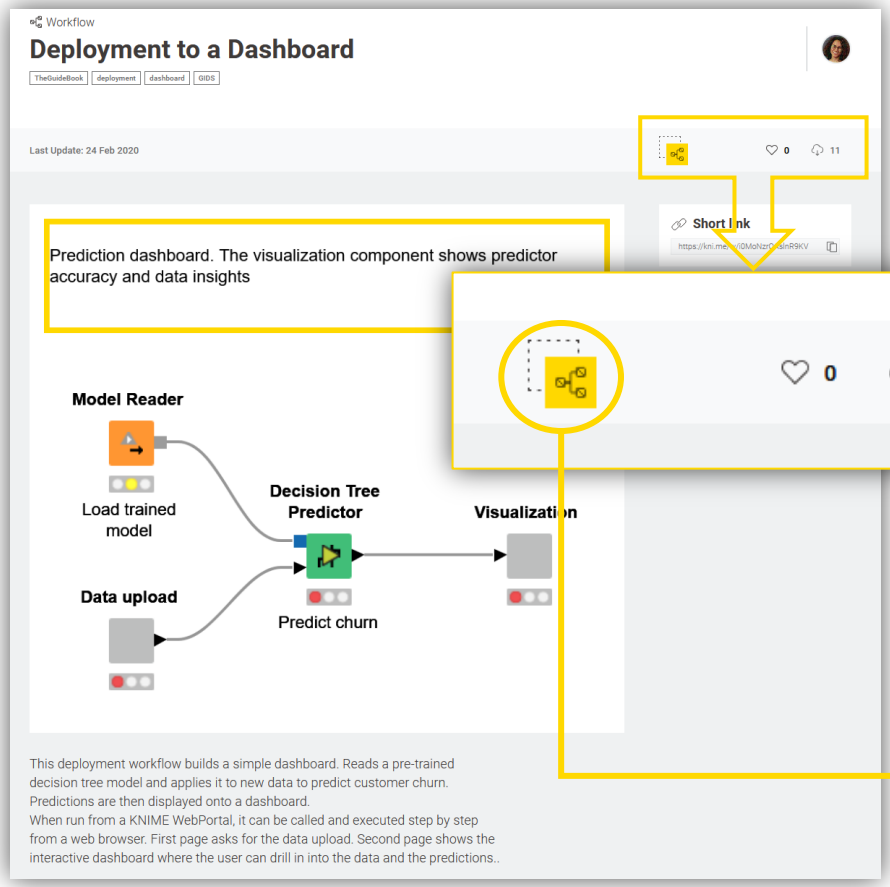


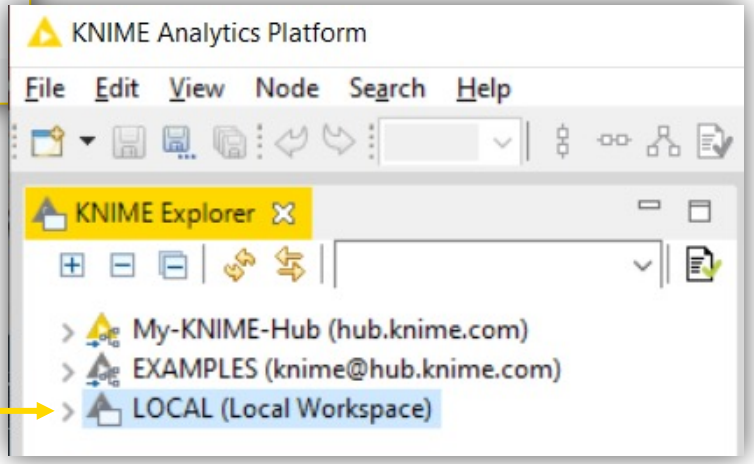
Yöntem 1

İş akışını indirin, makinenize yerleştirin ve daha önce görüldüğü gibi içe aktarın



KNIME Hub'dan indirme ve içe aktarma



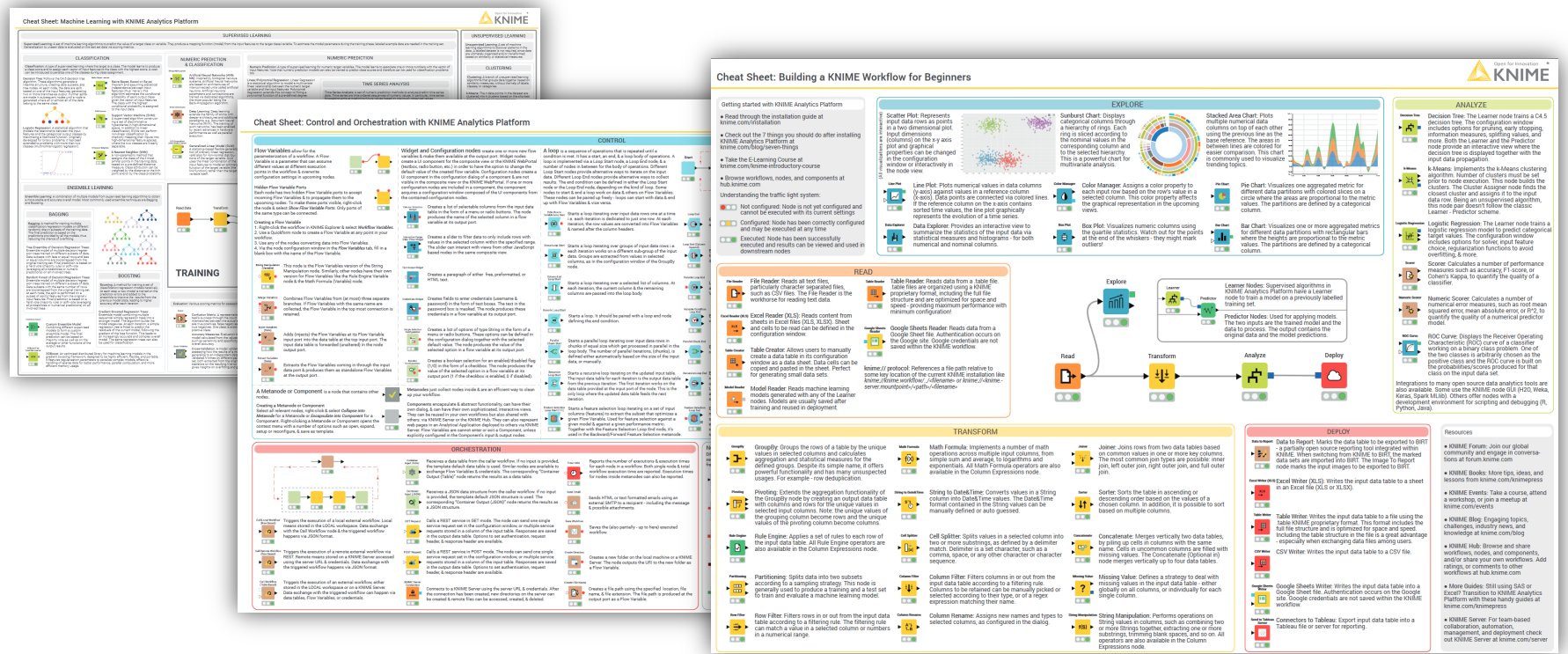


Yöntem 2

simgesini doğrudan istenen KNIME Explorer'daki konuma sürükleyip bırakın

KNIME Hile Sayfaları

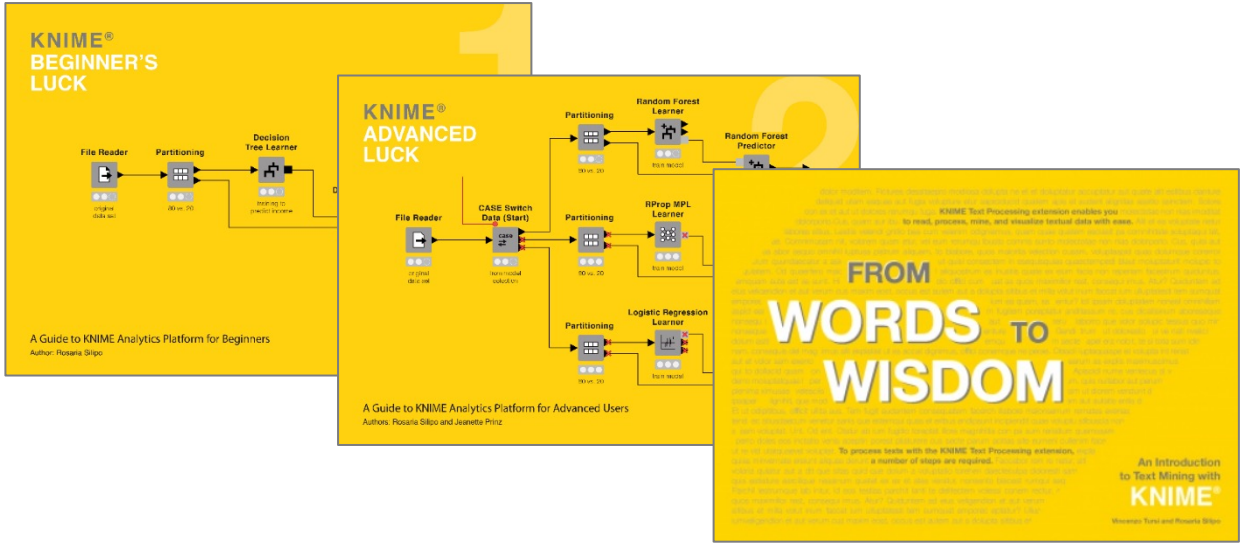
https://www.knime.com/cheat-sheets

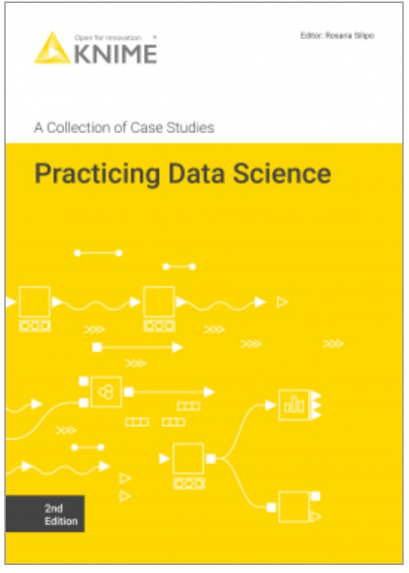


KNIME Kitapları

KNIME Press'ten e-kitap indirmeleri
https://www.knime.com/knimepress

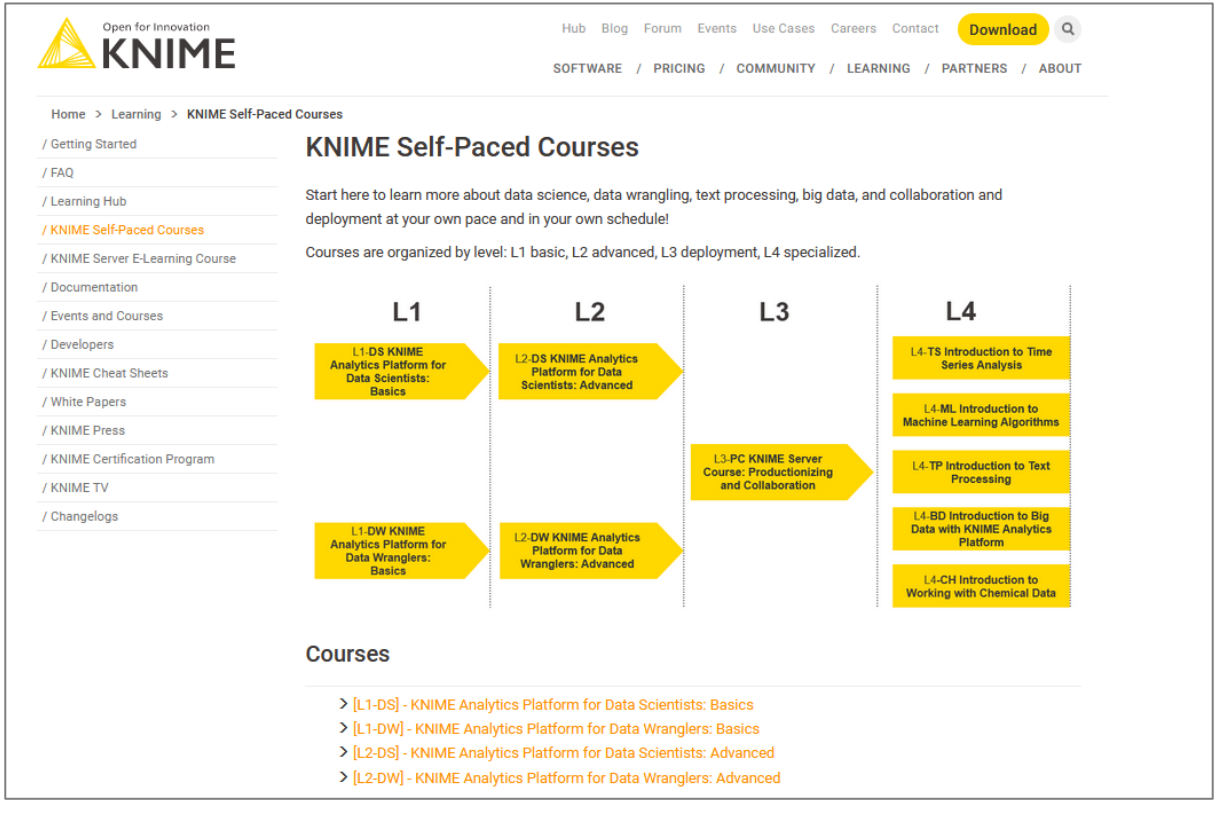
kod ile: < **Promosyon Kodu** >





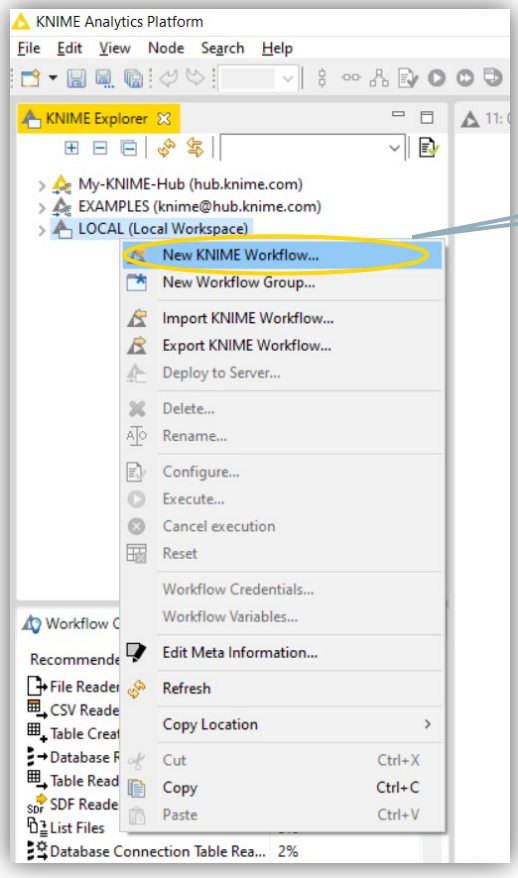
Ücretsiz Kendi Hızınızda Kurslar

- https://www.knime.com/knime-self-paced-courses



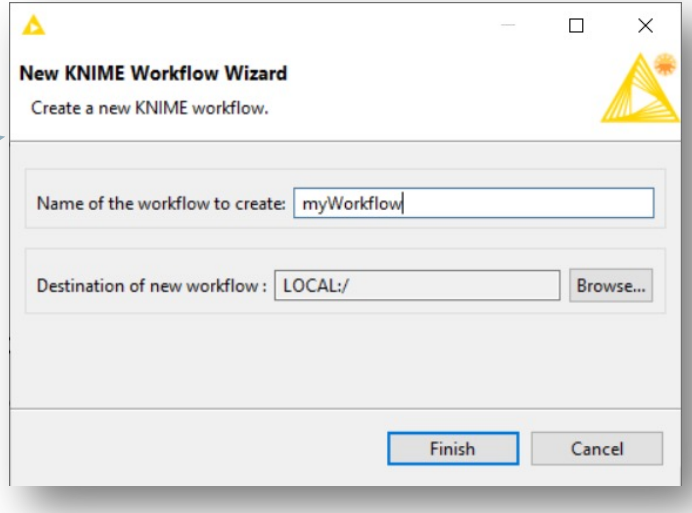
İlk İş Akışınızı oluşturun

İlk iş akışınızı (Workflow) oluşturun

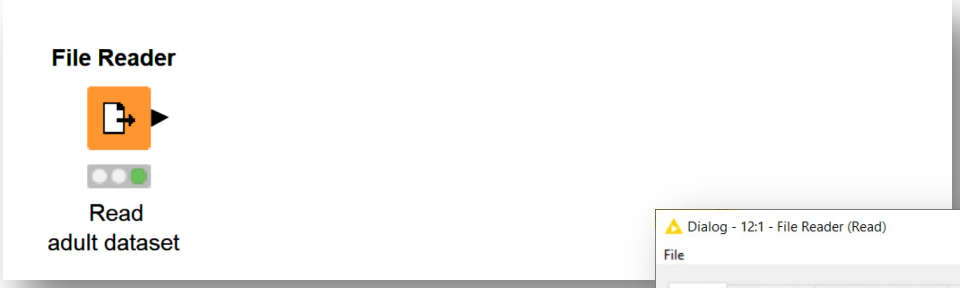


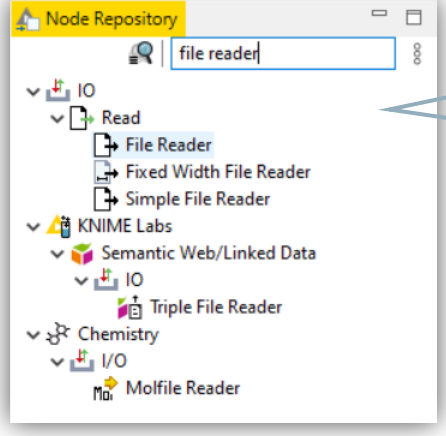
KNIME Explorer'da LOCAL klasöre sağ tıklayın ve *New KNIME Workflow'u seçin.*

Açılır pencereden ilk iş akışınızın adını girin



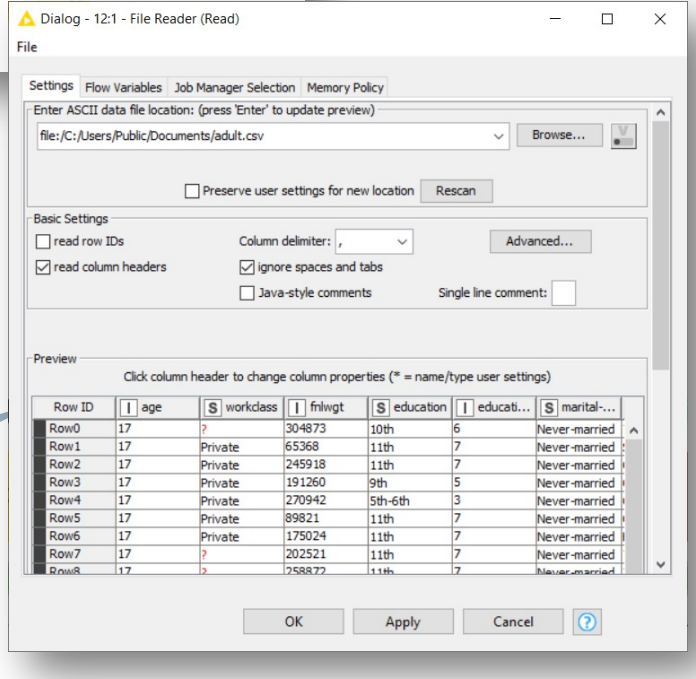
Veri kümesini okuyun



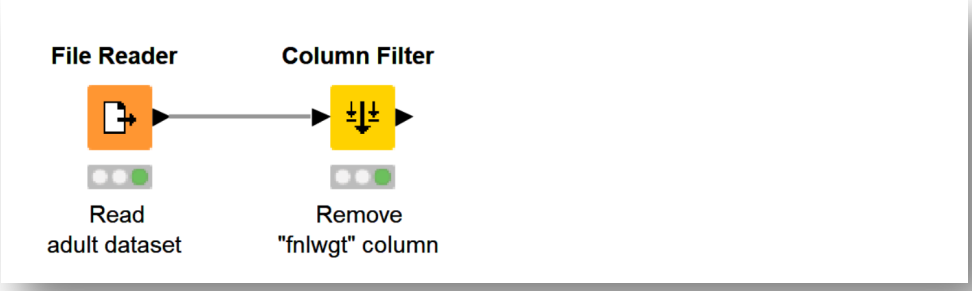


İş akışına eklemek için

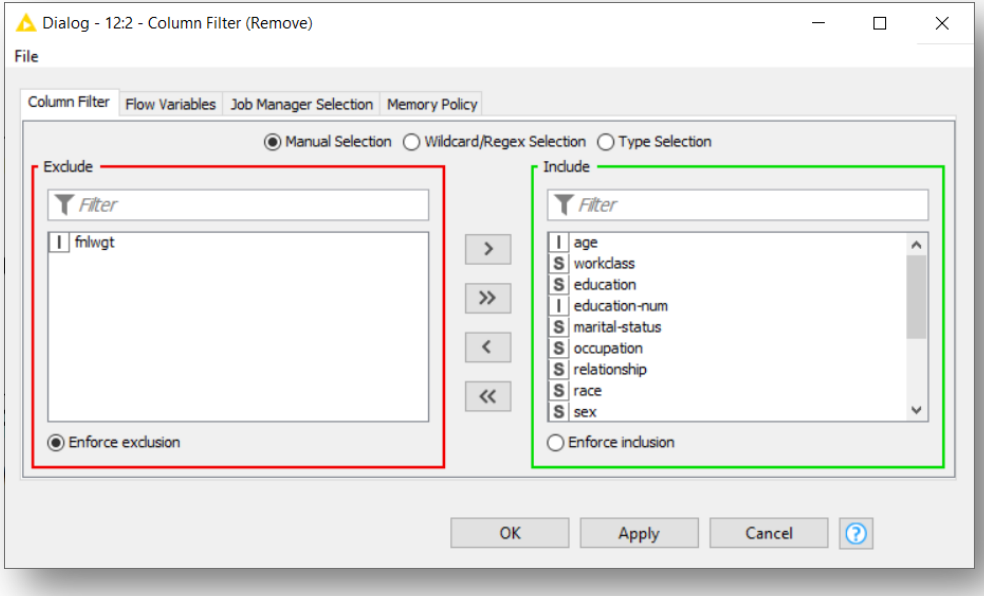
Yapılandırma penceresini açın (çift tıklayın) ve makinenizde «adult.csv» veri dosyasını seçin



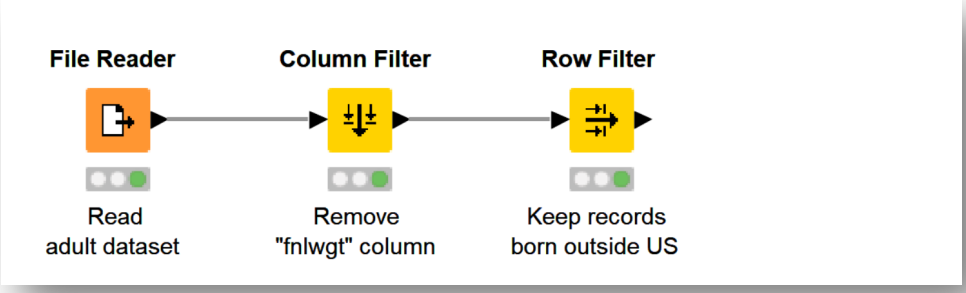
Sütunları kaldır

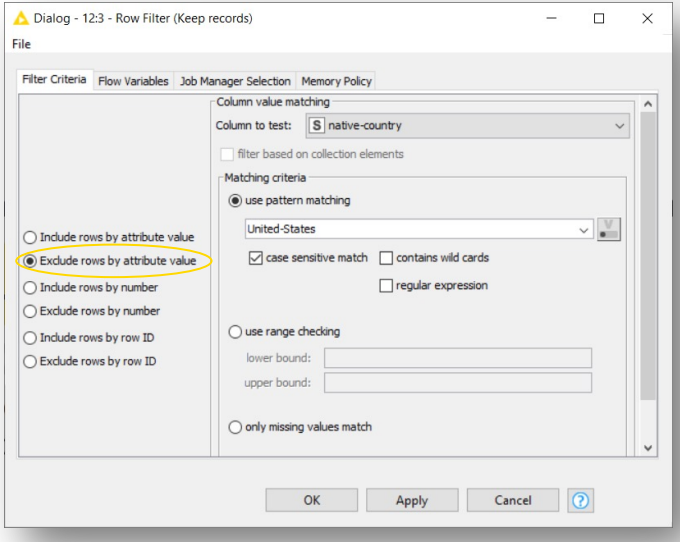


Bazı sütunlarda gereksiz bilgiler var. Bunları bir Sütun Filtresi düğümü ile kaldırın.



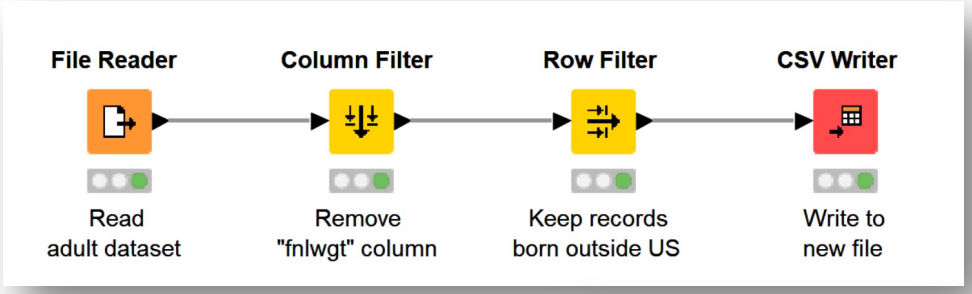
Satırları Kaldır





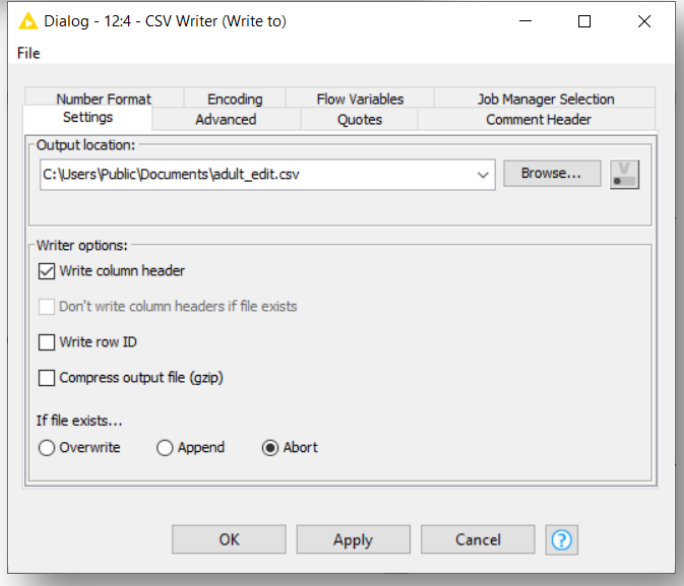
Bir Satır Filtresi düğümü ekleyin ve onu yalnızca değerleri "Amerika Birleşik Devletleri" olmayan girişleri tutacak şekilde yapılandırın

Yeni dosyaya yaz



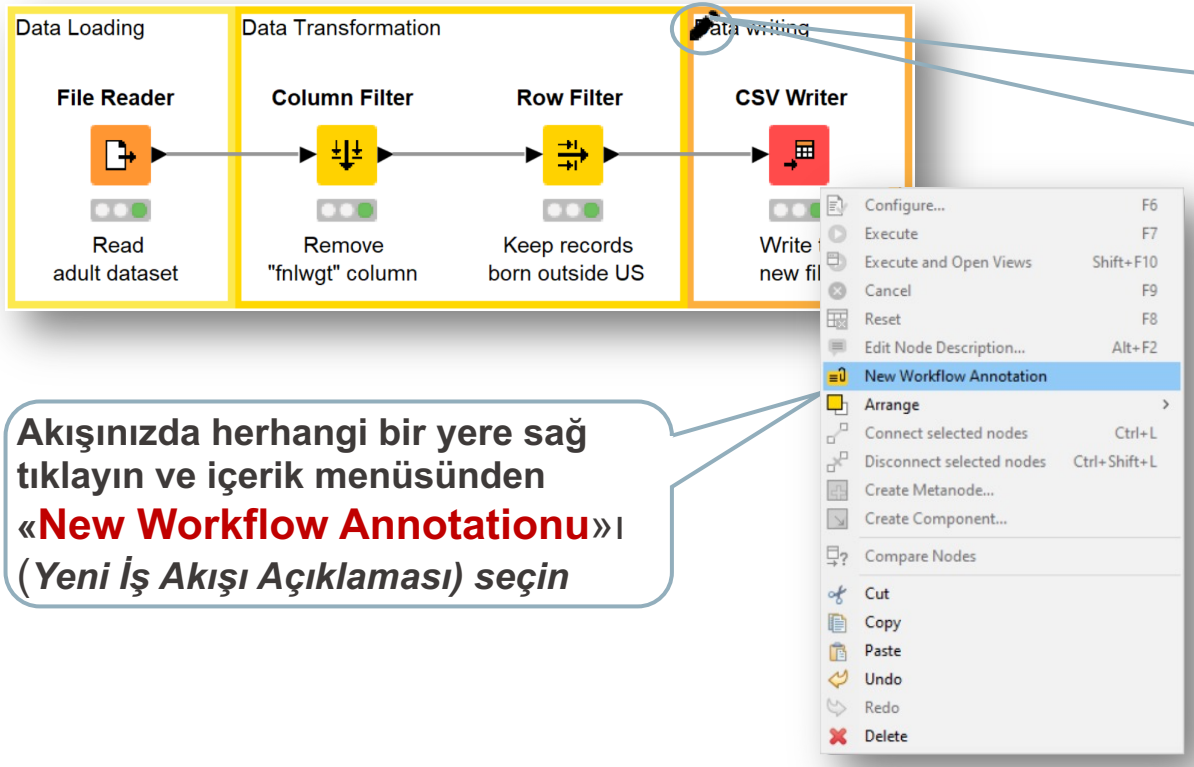
Son olarak, ardışık düzene bir CSV Writer düğümü ekleyin.

Dönüştürülen veri kümesini yeni bir dosyaya yazmak için yapılandırın ve çalıştırın



Ek açıklamalar (New Workflow Annotation)

- Ek açıklamalar, iş akışınıza ekleyebileceğiniz renkli düzenlenebilir kutulardır.
- Daha okunaklı ve görsel olarak hoş hale getirmenize yardımcı olurlar



Bir açıklamanın metnini ve görünümünü özelleştirmek için sol üst köşedeki «kalem simgesi»ne tıklayın

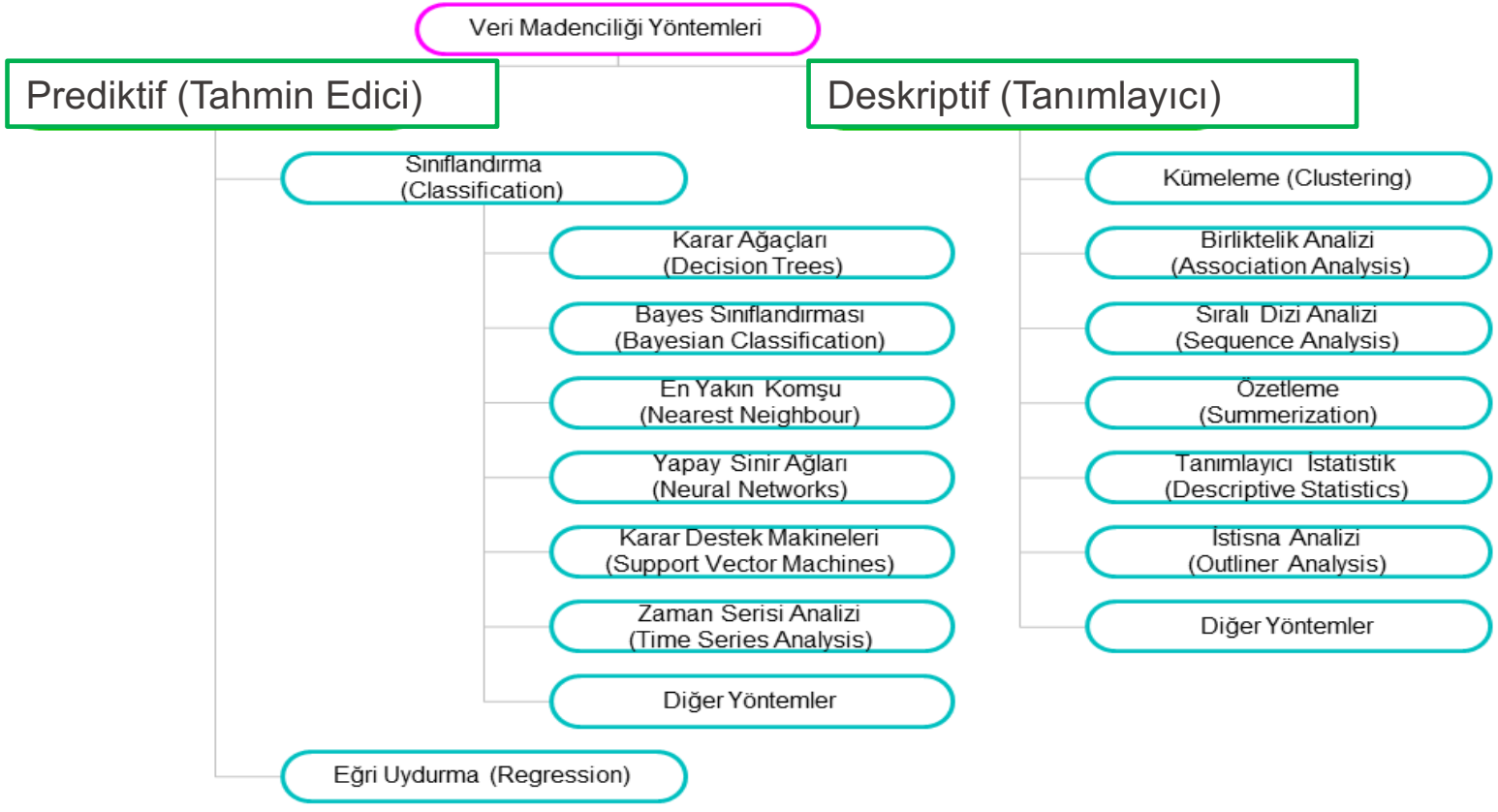
Akışınızda herhangi bir yere sağ tıklayın ve içerik menüsünden «New Workflow Annotationu»ı (*Yeni İş Akışı Açıklaması*) seçin

VERİ MADENCİLİĞİ

Veri Madenciliği Yöntemleri

Prof. Dr. Ünal Halit ÖZDEN

Veri Madenciliği Yöntemleri
-Genel Bilgi-1



Veri Madenciliği Yöntemleri
-Genel Bilgi-2

Prediktif (Tahmin Edici)

Tahmin edici modeller; eldeki verilerden hareketle bir model geliştirilmesi ve geliştirilen bu model kullanılarak önceden sonuçları bilinmeyen veri kümeleri için sonuçların tahmin edilmesini amaçlar. Kısaca bilinenden yola çıkarak bilinmeyeni tahmin etme çabasıdır. Örneğin bankalar, müşterilerinin önceki dönemlerde kullanmış oldukları kredilere ilişkin verilerine kendi veritabanlarından ulaşabilirler. Bu verilerden hareketle, müşterilerinin daha sonraki kredi taleplerinde kredi borcunu geri ödeyip ödemeyeceği, ya da ödemelerde düzenli olup olmayacağı konusunda tahminlerde bulunabilirler.

Deskriptif (Tanımlayıcı)

Tanımlayıcı modeller verilerdeki örüntü veya ilişkileri tanımlar. Bu modeller analiz edilen verilerin özelliklerini incelemek için kullanılan modellerdir. Veriler arasında çok rastlanan kurallar ortaya çıkarılır. Alış-veriş sepetindeki ürünler arasındaki ilişkiyi ortaya çıkarıp, müşterinin herhangi bir ürünü seçmesinin ardından müşteriye ilgisini çekecek bir başka ürünün önerilmesi. Bir diğer örnek olarak sigorta poliçesini yenilememiş müşterilerin benzer özelliklerini belirleyecek bir kümeleme çalışması verilebilir.

Veri Madenciliği Yöntemleri
-Tahmin Edici Yöntemler

Tahmin edici modeller kendi içinde *regresyon (eğri uydurma)* modelleri ve *sınıflandır- ma* modelleri biçiminde ikiye ayrılır.

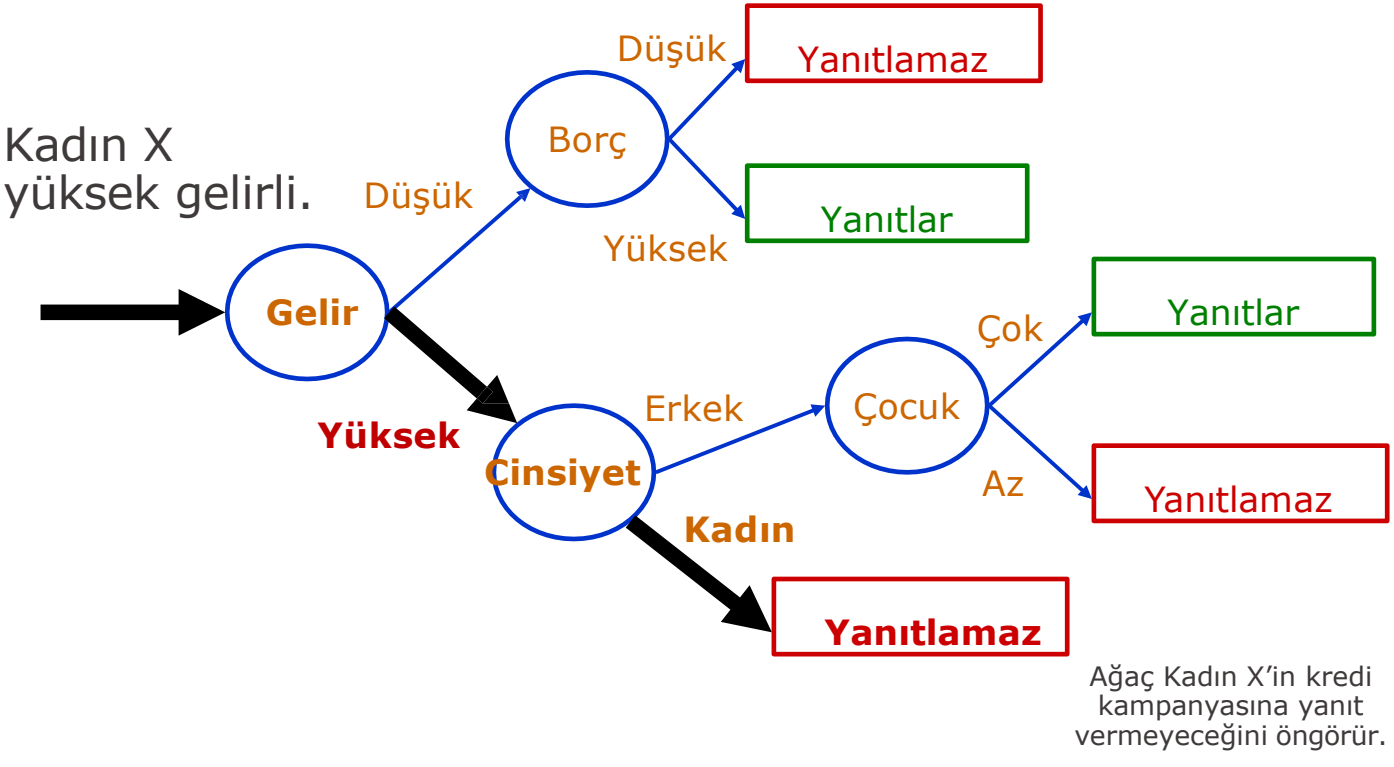
Regresyon Modelleri: Bilindiği gibi regresyon, bağımsız değişkenler ile bağımlı değiş- kenler arasındaki ilişkiyi en iyi tanımlayan fonksiyonu elde etmek için uygulanan ista- tistiksel tekniktir. Regresyon analizinde model, değişkenler arasındaki ilişkinin net bir biçimde gösterilebildiği bir fonksiyon ile temsil edilir. (Lojistik Regresyon, Lasso Regresyon vs.)

Sınıflandırma Modelleri: Sınıflama, veri sınıfı ve kavramlarını tanımlama ve ayırt et- meyi sağlayan bir model kümesini bulma sürecidir. Sınıflandırmada, veriler istatistik ve/ veya makine öğrenimi yöntemleri kullanılarak önceden belirlenen sınıflara atanır. Sınıfla- ma modelleri, sınıflar önceden incelenen veriler aracılığıyla oluşturulduğundan, denetim- li öğrenme modelleridir.

Regresyon ve sınıflandırma modellerinden en yaygın kullanılanlar; *karar ağaçları, yapay sinir ağları, genetik algoritmalar, zaman serisi analizi, k-en yakın komşu ve Bayes sınıflandırması* biçiminde sıralanabilir. Ünitenin izleyen kesiminde bu kavramlar kısaca açıklanmıştır.

Veri Madenciliği Yöntemleri
-Tahmin Edici Yöntemler
--Sınıflandırma
---Karar Ağaçları

Karar ağaçları biçiminde geliştirilen veri madenciliği modeli, kökleri yukarıda, ters çevrilmiş bir ağaca benzetilebilir. Ağaç karar verme noktalarını temsil eden düğümler ve bu düğümleri birbirine bağlayan dallardan oluşur. En üstte yer alan düğüm kök düğüm olarak adlandırılır. Kök düğümde bazı özellikler test edilerek bu testin farklı sonuçlarına göre kök düğümden farklı yönlere dallar oluşturulur. Her bir dal yeni bir karar düğümüne bağlanır ve burada yeni bir-takım özellikler test edilerek bu düğümlerden de yeni dallar türetilir. Ağaç yapısının en altında ise artık kendisinden yeni bir dal türemeyecek ve bu nedenle yaprak olarak adlandırılan düğümler bulunur. Buna göre veritabanındaki tüm kayıtlar bir ağaç yapısı biçiminde düzenlenerek ağaçta yer aldıkları dala göre sınıflandırılmış olur.



Ağaç Kadın X'in kredi kampanyasına yanıt vermeyeceğini öngörür.

Veri Madenciliği Yöntemleri
-Tahmin Edici Yöntemler
--Sınıflandırma
---Yapay Sinir Ağları

Yapay sinir ağları, özellikle bağımlı ve bağımsız değişkenler arasındaki karmaşık ve doğrusal olmayan ilişkileri modelleyebilmesi açısından tercih edilir. Bununla birlikte, bu yöntemle oluşturulan modellerin yorumlanması diğerlerine göre daha zordur. Yapay sinir ağları karmaşık hesaplamaları gerçekleştiren biyolojik sinir sistemlerini model alır. Bu anlamda biyolojik sinir sistemlerinin bir simülasyonudur.

---Genetik Algoritmalar

Genetik algoritmalar karmaşık eniyileme problemlerinin çözümünde kullanılan bir teknolojidir. Dolayısıyla aslında doğrudan bir veri madenciliği modeli değildir. Bununla birlikte veri madenciliğinde de kullanılabilen bir eniyileme yöntemidir. Genetik algoritmalar da yapay sinir ağları gibi biyolojik mekanizmalardan esinlenerek geliştirilmiş algoritmalardır. Genetik algoritmalar doğada gözlenen evrim sürecine benzer bir yapıda ele alınan problemi, sanal olarak evrimden geçirerek çözmektedir.

Veri Madenciliği Yöntemleri
-Tahmin Edici Yöntemler
--Sınıflandırma
---Zaman Serisi Analizi

Zaman serisi analizi, zaman değişkeni ile ilişkilendirilmiş verilerin tahmin edilmesnde kullanılır. Zaman serisi analizlerinin kullanıldığı en yaygın alan borsa işlemleridir. Bir hisse senedinin veya borsa endeksinin gelecek değeri tahmini zaman serisi problemlerine örnek oluşturur. Tahmin modellerinin oluşturulmasında geçmiş verilerden yararlanılması nedeniyle bu modeller denetimli öğrenme modellerindendir.

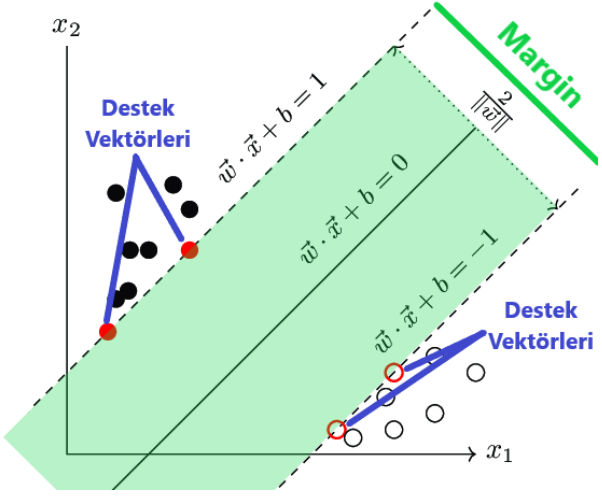
---Bayes Sınıflandırması (Naïve Bayes)

Bayes sınıflandırma yöntemi, elde var olan, mevcut sınıflanmış verileri kullanarak yeni bir verinin mevcut sınıflardan herhangi birine girme olasılığını hesaplayan yöntemdir. İstatistikte kullanılan Bayes kuralına dayalı olarak geliştirilmiş algoritma ve sınıflandırma teknikleri bu isimle anılır (Silahtaroğlu, s. 97).

Veri Madenciliği Yöntemleri
-Tahmin Edici Yöntemler
--Sınıflandırma
---Destek Vektör Makineleri (Support Vector Machine)

Destek Vektör Makineleri (Support Vector Machine) genellikle sınıflandırma problemlerinde kullanılan gözetimli öğrenme yöntemlerinden biridir. Bir düzlem üzerine yerleştirilmiş noktaları ayırmak için bir doğru çizer. Bu doğrunun, iki sınıfının noktaları için de maksimum uzaklıkta olmasını amaçlar. Karmaşık ama küçük ve orta ölçekteki veri setleri için uygundur. Daha açıklayıcı olması için görsel üzerinde tekrar inceleyelim.

Tabloda siyahlar ve beyazlar olmak üzere iki farklı sınıf var. Sınıflandırma problemlerindeki asıl amacımız gelecek verinin hangi sınıfta yer alacağını karar vermektir. Bu sınıflandırmayı yapabilmek için iki sınıfı ayıran bir doğru çizilir ve bu doğrunun ±1'i arasında kalan yeşil bölgeye Margin adı verilir. Margin ne kadar geniş ise iki veya daha fazla sınıf o kadar iyi ayrıştırılır.



Veri Madenciliği Yöntemleri
-Tahmin Edici Yöntemler
--Sınıflandırma
---k-En Yakın Komşu

k-en yakın komşu algoritması sıklıkla kullanılan bir algoritmadır. Sınıflandırma yapılırken her bir kayıdın diğer kayıtlarla olan uzaklığı hesaplanır. Ancak, bir kayıt için diğer kayıtlardan sadece *k* adedi göz önüne alınır. Algoritmanın isminden de anlaşılacağı gibi bu *k* adet kayıt, başka bir ifadeyle veritabanındaki nokta, mesafesi hesaplanan noktaya diğer kayıtlara nazaran en yakın olan kayıtlardır. Bu yöntem coğrafi bilgi sistemlerinde çok kullanılır, belirlenen bir noktaya en yakın şehir, istasyon vb. belirlenmesi aslında k-en yakın komşu algoritmasının temelini oluşturur. Algoritmada k değeri önceden seçilir; değerinin yüksek olması birbirlerine benzemeyen noktaların bir araya toplanmasına, çok küçük seçilmesiyse birbirine benzediği, yani aynı sınıfın noktaları oldukları hâlde, bazı noktaların ayrı sınıflara konmasına ya da o tür noktalar için ayrı sınıfların açılmasına neden olur (Silahtaroğlu, s. 118). Gözlem değerlerinin arasındaki uzaklıkların hesaplanmasında “Öklid” uzaklık formülü kullanılır.

Veri Madenciliği Yöntemleri
-Tanımlayıcı Yöntemler
--Kümeleme

Kümeleme (denetimsiz), verileri birbirlerine olan benzerliklerine göre anlamlı ve/ veya kullanışlı gruplara ayırmaktır. Eğer amaç anlamlı kümeler oluşturmaksa o zaman kümeler verilerin doğal yapısını yansıtmalıdır. Bazı durumlarda ise küme- leme veri özetleme gibi farklı amaçlar için kullanışlı bir başlangıç noktası oluş- turmaktadır. Kümeleme analizi bir hedef değişken içermediğinden, diğer bir ifade ile veriler bağımlı bir değişkene göre değil öznitelik değerlerine göre gruplandı- rıldığından, daha önce sözü edilen *sınıflama* analizinden farklı bir yaklaşımdır. Kümeleme analizinde, hedef değişkenin değerini belirlemeye yönelik sınıflama, tahmin etme veya kestirim yapılmaya çalışılmaz. Bunun yerine verinin tamamını bölümlere ayırmak için homojen alt gruplar veya kümeler araştırılır. Bu işlem ger-ekleştirilirken kümeler içindeki verilerin benzerliği göz önüne alınır. Oluşturulan kümeler önceden tanımlanmadığından ve verinin özelliklerine göre belirlendiğin- den kümelerin anlamı konuyla ilgili bir alan uzmanı tarafından yorumlanmalıdır. Verilerin kümeleme analizine göre modellenmesinde matematik, istatistik, makine öğrenimi ve yapay zekâ gibi birçok alandan yararlanılır. (K-means algoritması, hiyerarşik kümeleme)

Veri Madenciliği Yöntemleri
-Tanımlayıcı Yöntemler
--Birliktelik Kuralları

Birliktelik kuralları veriler arasındaki güçlü birliktelik özelliklerini tanımlayan örüntüleri keşfetmek için kullanılan analiz yöntemidir. Birliktelik kuralı, belirli türdeki veri ilişkilerini tanımladığı için tanımlayıcı modeller içinde yer almaktadır. Herhangi bir ürün alındığında bir başka ürünün de satın alınması bir birliktelik kuralı verir. İş dünyasında birliktelik analizi, pazar sepeti veya benzeşme analizi olarak da adlandırılır ve müşterilerin satın alma alışkanlıklarını analiz ederek, ilgili ürünler arasındaki potansiyel çapraz satış olanaklarını tanımlamak için kullanılır. Örneğin; “Kola satın alan müşteriler %80 olasılıkla cips de satın alırlar” biçimindeki sonuçlara birliktelik kuralları analizi ile ulaşılabilir. Raf düzenlemeleri bu sonuçlar temel alınarak yapıldığında satış oranları arttırılabilir.

--Sıra Örüntü Analizi

Sıra örüntü analizi birliktelik kurallarına benzer bir yapıda olup aynı zamanda olayların zaman sıralarıyla ilgilenir. Birliktelik kurallarında sözü edilen pazar sepeti analizinde, ürünlerin müşteri tarafından aynı anda alınmasıyla ilgilenilirken sıra örüntüleri analizinde belirli bir zaman aralığında satın alınan ürünler arasındaki ilişkilerle ilgilenilir. “A ameliyatı olan bir hastada, 10 gün içinde %40 olasılıkla B enfeksiyonu oluşacaktır”, ya da “Çekiç satın alan bir müşteri, ilk üç ay içerisinde %15, bu dönemi izleyen üç ay içerisinde %10 olasılıkla çivi satın alacaktır” biçiminde sıralanabilecek ilişki tanımlamaları, sıra örüntü analizi ile tanımlanabilecek ilişkilere örneklerdir.

--Özetleme

Karakteristik değerlerin hesaplanması veya genelleştirme olarak da adlandırılan özetleme, verileri basit tanımları yapılmış alt gruplar içine yerleştirme işlemidir. Özetleme veritabanı hakkında betimleyici bilgileri ortaya çıkarır ve verilerden elde edilen ortalama veya standart sapma gibi tüm veriyi temsil eden göstergelerin hesaplanmasını ifade eder. Özet bilgiler, veritabanı fonksiyonları ve tanımlayıcı veri madenciliği teknikleri kullanılarak elde edilebilir.

VERİ MADENCİLİĞİ

Sınıflandırma Yöntemleri (Classification)

Prof. Dr. Ünal Halit ÖZDEN

Veri Madenciliği Yöntemleri-3
-Sınıflandırma

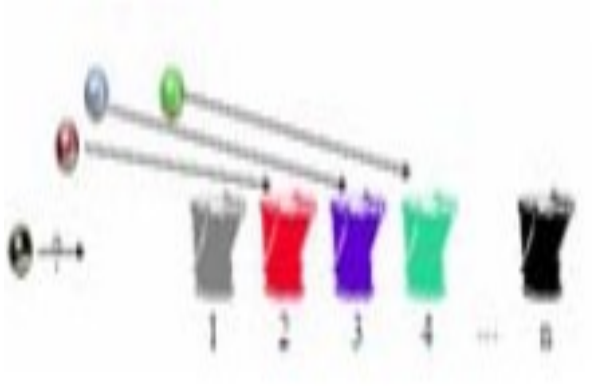
Veri madenciliği konusunda çok sayıda yöntem ve algoritma vardır. Bu yöntemlerin bir çok istatistik tabanlıdır. Bu yöntemler temel olarak 3 grupta toplanır: “Sınıflandırma”, “Kümeleme” ve “Birliktelik kuralları”.

Sınıflandırma	Kümeleme	Birliktelik Kuralları
Sınıflama veri mdenciliğinde sıkça kullanılan bir yöntem olup, veri tabanlarındaki gizli örüntüleri ortaya çıkarmakta kullanılır. Verilerin sınıflandırılması için belirli bir süreç izlenir. Öncelikle var olan veritabanının bir kısmı eğitim amacıyla kulanılarak sınıflandırma kurallarının oluşturulması sağlanır. Daha sonra bu kurallar yardımıyla yeni bir durum ortaya çıktığında nasıl karar verileceği belirlenir.	Kümelerin kendi aralarındaki benzerliklerin göz önüne alınarak gruplandırılması işlemidir. Bu özelliği nedeniyle pek çok alanda uygulanabilmektedir. Örn. Pazarlama araştırmaları, resim işleme, uzaysal harita verilerini işleme.	Veritabanı içinde yer alan kayıtların birbirleriyle olan ilişkilerini inceleyerek, hangi olayların eş zamanlı olarak birlikte gerçekleşebiileceklerini ortaya koymaya çalışan veri madenciliği yöntemleri bulunmaktadır. Bu ilişkilerin belirlenmesi ile “birliktelik kuralları” (association rules) elde edilir. Örn. Pazar sepeti analizi…

Sınıflandırma Yöntemleri
-Sınıflandırma-1

- Sınıflandırma problemi:
  - nesnelerden oluşan veri kümesi (öğrenme kümesi):
  - $D=\{t_1,t_2,\ldots,t_n\}$
  - her nesne niteliklerden oluşuyor, niteliklerden biri sınıf bilgisi
- Sınıf niteliğini belirlemek için diğer nitelikleri kullanarak bir model bulma
- Öğrenme kümesinde yer almayan nesneleri (test kümesi) mümkün olan en iyi şekilde doğru sınıflara atamak
- sınıflandırma=ayrık değişkenler için öngörüde (prediction) bulunmak.

Sınıflandırma Yöntemleri
-Sınıflandırma-2



- Amaç: Yeni bir kayıt geldiğinde, bu kaydı geliştirilen modeli kullanarak mümkün olduğunca doğru bir sınıfa atamak.
  - verinin dağılımına göre bir model bulunur
  - bulunan model, başarımı belirlendikten sonra niteliğin gelecekteki ya da bilinmeyen değerini tahmin etmek için kullanılır
  - Sınıflandırma: hangi topun hangi sepete koyulabileceği
  - Öngörü: Topun ağırlığı
    - model başarımı: doğru sınıflandırılmış sınama kümesi örneklerinin oranı
- Veri madenciliği uygulamasında:
  - ayrık nitelik değerlerini tahmin etmek: sınıflandırma
  - sürekli nitelik değerlerini tahmin etmek: öngörü

Sınıflandırma Yöntemleri
-Sınıflandırma
--Denetimli & Denetimsiz Öğrenme

- Denetimli (Gözetimli, Supervised) öğrenme= sınıflandırma
  - Sınıfların sayısı ve hangi nesnenin hangi sınıfta olduğu biliniyor.

- Denetimsiz (Gözetimsiz, Unsupervised) öğrenme=kümeleme (clustering)
  - Hangi nesnenin hangi sınıfta olduğu bilinmiyor. Genelde sınıf sayısı bilinmiyor.

Sınıflandırma Yöntemleri
-Sınıflandırma
--Sınıflandırma Uygulamaları

- Kredi başvurusu değerlendirme
- Kredi kartı harcamasının sahtekarlık olup olmadığına karar verm
- Hastalık teşhisi
- Ses tanıma
- Karakter tanıma
- Gazete haberlerini konularına göre ayırma
- Kullanıcı davranışları belirleme

--Sınıflandırma İşlemi: Modeli Kullanma

- 3. Modeli kullanma:
  - Model daha önce görülmemiş örnekleri sınıflandırmak için kullanılır
    - Örneklerin sınıf etiketlerini tahmin etme
    - Bir niteliğin değerini tahmin etme

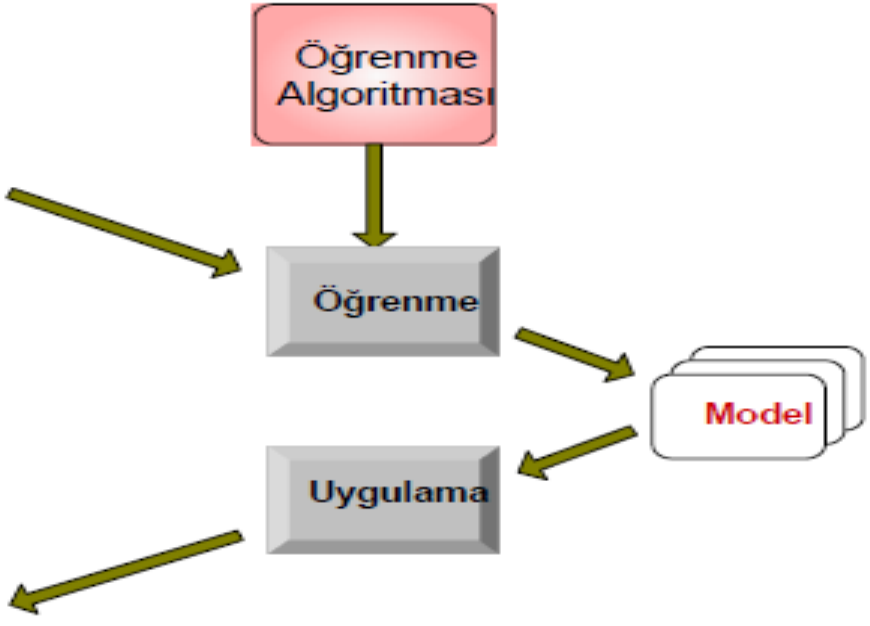
Sınıflandırma Yöntemleri
-Sınıflandırma
--Örnek

Tid	Nit1	Nit2	Nit3	Sınıf
1	1	Büyük	125K	0
2	0	Orta	100K	0
3	0	Küçük	70K	0
4	1	Orta	120K	0
5	0	Büyük	95K	1
6	0	Orta	60K	0
7	1	Büyük	220K	0
8	0	Küçük	85K	1
9	0	Orta	75K	0
10	0	Küçük	90K	1

Öğrenme Kümesi

Tid	Nit1	Nit2	Nit3	Sınıf
11	0	Küçük	55K	?
12	1	Orta	80K	?
13	1	Büyük	110K	?
14	0	Küçük	95K	?
15	0	Büyük	67K	?

Sınama Kümesi



Sınıflandırma Yöntemleri
-Sınıflandırma
--Sınıflandırıcı Başarımını Değerlendirme

Doğru sınıflandırma başarısı

- Hız
  - modeli oluşturmak için gerekli süre
  - sınıflandırma yapmak için gerekli süre
- Kararlı olması
  - veri kümesinde gürültülü ve eksik nitelik değerleri olduğu durumlarda da iyi sonuç vermesi
- Ölçeklenebilirlik
  - büyük miktarda veri kümesi ile çalışabilmesi
- Anlaşılabilir olması
  - kullanıcı tarafından yorumlanabilir olması
- Kuralların yapısı
  - birbiriyle örtüşmeyen kurallar

Sınıflandırma Yöntemleri
-Sınıflandırma
--Sınıflandırma Yöntemleri

- Karar ağaçları (decision trees)
- Yapay sinir ağları (artificial neural networks)
- Bayes sınıflandırıcılar (Bayes classifier)
- İlişki tabanlı sınıflandırıcılar (association-based classifier)
- k-en yakın komşu yöntemi (k- nearest neighbor method)
- Destek vektör makineleri (support vector machines)
- Genetik algoritmalar (genetic algorithms)
- ...

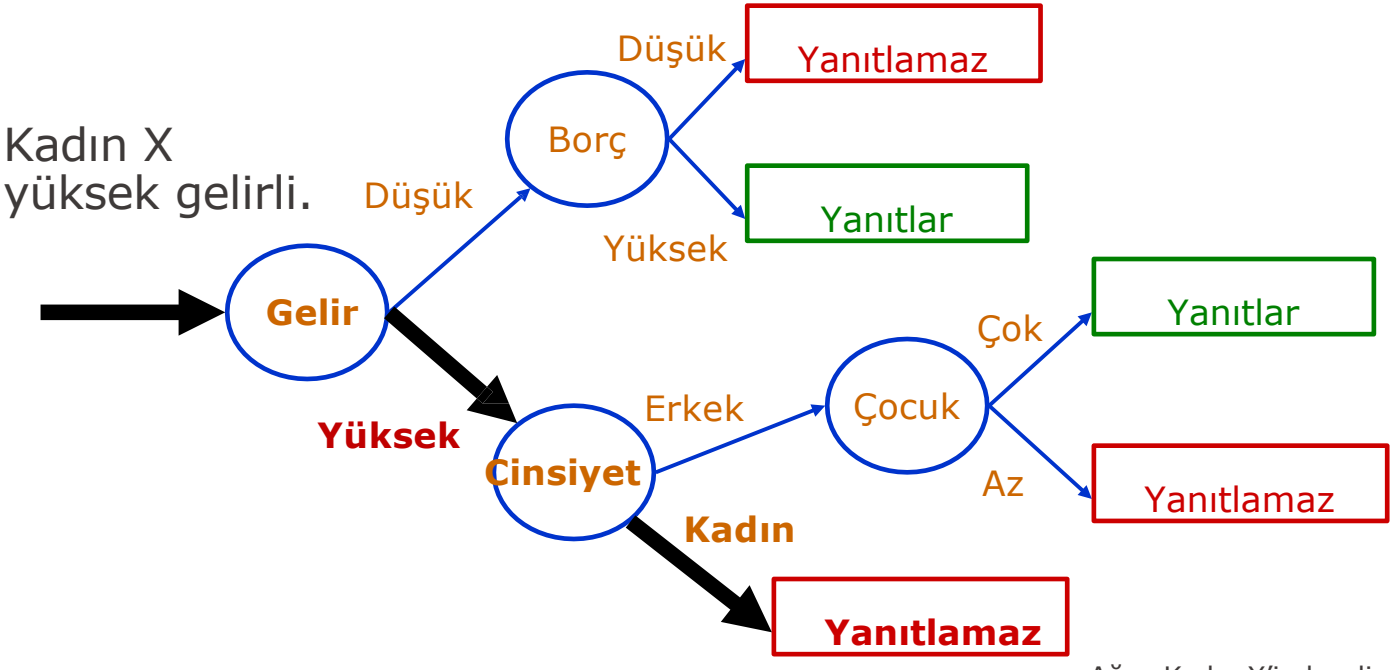
Sınıflandırma Yöntemleri
-Karar Ağaçları ile Sınıflandırma

- Karar Ağacı
  - Yaygın kullanılan öngörü yöntemlerinden bir tanesidir.
  - Ağaçtaki her düğüm bir özellikteki testi gösterir.
  - Düğüm dalları testin sonucunu belirtir.
  - Ağaç yaprakları sınıf etiketlerini içerir.
- Karar ağacı çıkarımı iki aşamadan oluşur
  - Ağaç inşası
    - Başlangıçta bütün öğrenme örnekleri kök düğümdedir.
    - Örnekler seçilmiş özelliklere tekrarlamalı olarak göre bölünür.
  - Ağaç Temizleme (Budama) (Tree pruning)
    - Gürültü ve istisna kararları içeren dallar belirlenir ve kaldırılır.
- Karar ağacı kullanımı: Yeni bilinmeyen örneğin sınıflandırılması
  - Bilinmeyen örneğin özellikleri karar ağacında test edilerek sınıfı bulunur.

Sınıflandırma Yöntemleri

-Karar Ağaçları ile Sınıflandırma

--Bir Kredi Kartı Kampanyasında Yeni Bir Örneğin Sınıflandırılması



Ağaç Kadın X'in kredi kampanyasına yanıt vermeyeceğini öngörür.

Sınıflandırma Yöntemleri
-Karar Ağaçları ile Sınıflandırma
--Karar Ağacı Yöntemleri

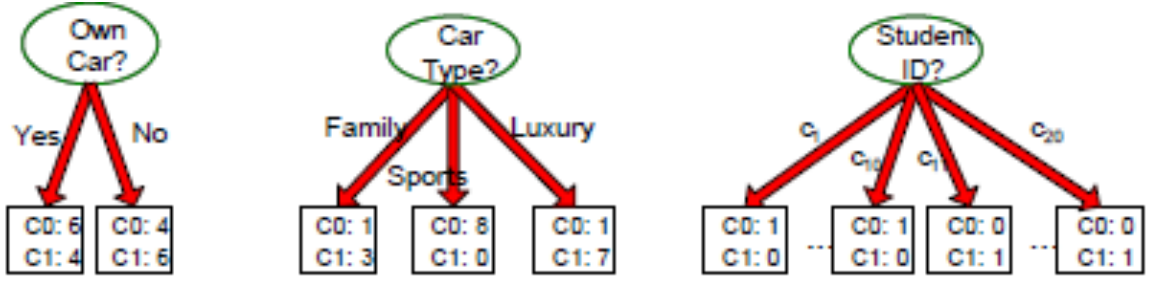
- Karar ağacı oluşturma yöntemleri genel olarak iki aşamadan oluşur:
  - **1. Ağaç oluşturma**
    - en başta bütün öğrenme kümesi örnekleri kökte
    - seçilen niteliklere bağlı olarak örnek yinelemeli olarak bölünüyor.
  - **2. Ağaç budama**
    - öğrenme kümesindeki gürültülü verilerden oluşan ve sınama kümesinde hataya neden olan dalları silme (sınıflandırma başarımını artırır)

---Karar Ağacı Oluşturma

- Yinelemeli işlem
  - ağaç bütün verinin oluşturduğu tek bir düğümle başlıyor
  - eğer örnekleri hepsi aynı sınıfa aitse düğüm yaprak olarak sonlanıyor ve sınıf etiketini alıyor
  - eğer değilse örnekleri sınıflara en iyi bölecek olan nitelik seçiliyor
  - işlem sona eriyor
    - örneklerin hepsi (çoğunluğu) aynı sınıfa ait
    - örnekleri bölecek nitelik kalmamış
    - kalan niteliklerin değerini taşıyan örnek yok

Örnekleri En İyi Bölen Nitelik Hangisi?

- Bölmeden önce:
  - 10 örnek C0 sınıfında
  - 10 örnek C1 sınıfında



Hangisi daha iyi?

En iyi Bölme Nasıl Belirlenir?

- "Greedy" (aç gözlü)yaklaşım
  - çoğunlukla aynı sınıfa ait örneklerin bulunduğu düğümler tercih edilir
- Düğümün kalitesini ölçmek için bir yöntem

En İyi Bölen Nitelik Nasıl Belirlenir?

- İyilik Fonksiyonu (Goodness Function)
- Farklı algoritmalar farklı iyilik fonksiyonları kullanabilir:
  - bilgi kazancı (information gain): ID3, C4.5
    - bütün niteliklerin ayrık değerler aldığı varsayılıyor
    - sürekli değişkenlere uygulamak için değişiklik yapılab
  - gini index (IBM IntelligentMiner)
    - her nitelik ikiye bölünüyor
    - her nitelik için olası bütün ikiyi bölünmeler sınanıyor

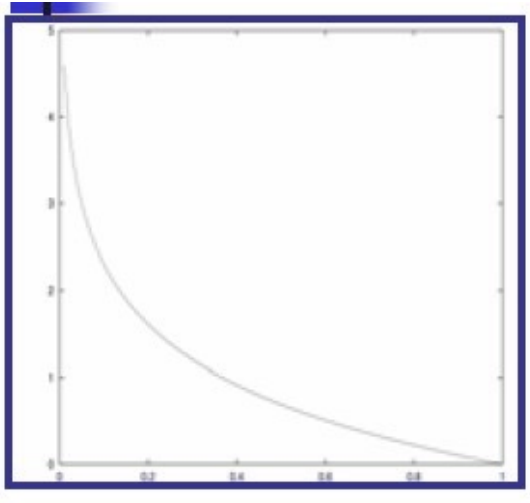
Bilgi / Entropi

- p_1, p_2,.., p_s toplamları 1 olan olasılıklar. Entropi (Entropy)

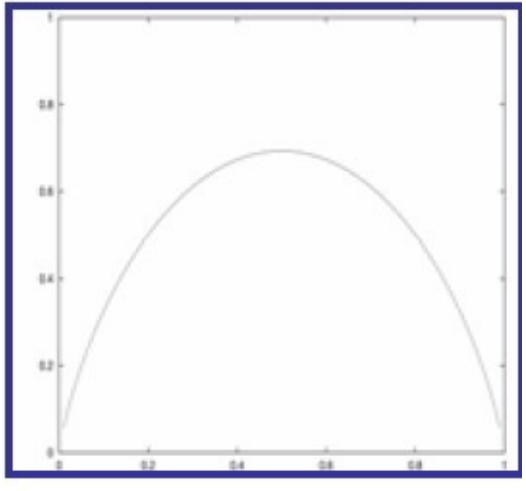
$$H(p_1, p_2, \ldots, p_s) = -\sum_{i=1}^{s} p_i \log(p_i)$$

- Entropi rastgeleliği, belirsizliği ve beklenmeyen durumun ortaya çıkma olasılığını gösterir.
- Sınıflandırmada
  - olayın olması beklenen bir durum
  - entropi=0

Entropi



log (p)



H(p,1-p)

- örnekler aynı sınıfa aitse entropi=0
- örnekler sınıflar arasında eşit dağılmışsa entropi=1
- örnekler sınıflar arasında rastgele dağılmışsa 0<entropi<1

Örnek

- S veri kümesinde 14 örnek: C0 sınıfına ait 9, C1 sınıfına ait 5 örnek.
- Entropi $$H(p_1, p_2, ..., p_s) = -\sum_{i=1}^{s} p_i \log(p_i)$$
- $H(p_1,p_2)$= - (9/14) Log_2 (9/14) - (5/14) Log_2 (5/14)
  = 0.940

Bilgi Kazancı (ID3 / C4.5)

- Bilgi kuramı kavramlarını kullanarak karar ağacı oluşturulur. Sınıflandırma sonucu için en az sayıda karşılaştırma yapmayı hedefler.
- Ağaç bir niteliğe göre dallandığında entropi ne kadar düşer?
- A niteliğinin S veri kümesindeki bilgi kazancı

$$Gain(S, A) = Entropy(S) - \sum_{v \in Values(A)} \frac{|Sv|}{|S|} Entropy(S_v)$$

Values(A), *A* niteliğinin alabileceği değerler, S_v, $A=v$ olduğu durumda S'nin altkümesi.

VERİ MADENCİLİĞİ

(Karar Ağaçları ile Sınıflandırma)

Genel İçerik

- Veri Madenciliğine Giriş
- Veri Madenciliğinin Adımları
- Veri Madenciliği Yöntemleri
  - Sınıflandırma
  - Kümeleme
  - İlişkilendirme/birliktelik kuralları
- Metin madenciliği
- WEB madenciliği
- Veri Madenciliği Uygulamaları

İçerik

- Sınıflandırma yöntemleri
  - Karar ağaçları ile sınıflandırma
    - Entropi Kavramı
    - ID3 Algoritması
    - C4.5 Algoritması

    } Entropiye dayalı algoritmalar

    - Twoing Algoritması
    - Gini Algoritması

    } Sınıflandırma ve regresyon ağaçları (CART)

    - k-en yakın komşu algoritması } Bellek tabanlı algoritmalar

Karar Ağaçları ile Sınıflandırma

- Sınıflandırma problemleri için yaygın kullanılan yöntemdir.
- Sınıflandırma doğruluğu diğer öğrenme metotlarına göre çok etkindir.
- Öğrenmiş sınıflandırma modeli ağaç şeklinde gösterilir ve karar ağacı (decision tree) olarak adlandırılır.
- Karar ağaçları akış şemalarına benzeyen yapılardır. Her bir nitelik bir düğüm tarafından temsil edilir. Dallar ve yapraklar ağaç yapısının elemanlarıdır. En son yapı yaprak en üst yapı kök ve bunların arasında kalan yapılar dal olarak isimlendirilir.

Karar Ağaçlarında Dallanma Kriterleri

- Karar ağaçlarında en önemli sorunlardan birisi hangi kökten itibaren bölümlemenin veya dallanmanın hangi kritere göre yapılacağıdır. Aslında her farklı kriter için bir karar ağacı algoritması karşılık gelmektedir.
- Bu algoritmalar şu şekilde gruplandırılabilir.
  - ID3 ve C4.5, entropiye dayalı sınıflandırma algoritmalarıdır.
  - Twoing ve Gini, CART (Classification And Regression Trees) sınıflandırma ve regresyon ağaçlarına dayalı sınıflandırma algoritmalarıdır.
  - k-en yakın komşu algoritması bellek tabanlı sınıflandırma yöntemleri arasında yer almaktadır.

Entropi

(1/3)

- Entropi, rastgele değere sahip bir değişken veya bir sistem için belirsizlik ölçütüdür.
- Enformasyon, rassal bir olayın gerçekleşmesi halinde ortaya çıkan bilgi ölçütüdür.
- Bir süreç için entropi, tüm örnekler tarafından içerilen enformasyonun beklenen değeridir.
- Eşit olasıklı durumlara sahip sistemler yüksek belirsizliğe sahiptirler.
- Shannon, bir sistemdeki durum değişikliğinde, entropideki değişimin enformasyon boyutunu tanımladığını öne sürmüştür.
- Buna göre bir sistemdeki belirsizlik arttıkça, bir durum gerçekleştiğinde elde edilecek enformasyon boyutu da artacaktır.

Entropi

(2/3)

- Shannon bilgiyi bitlerle ifade ettiği için, logaritmayı 2 tabanında kullanmıştır.

- S bir kaynak olsun. Bu kaynağın $\{m_1, m_2, \ldots m_n\}$ olmak üzere n mesaj üretildiğini varsayalım. Tüm mesajlar birbirinden bağımsız üretilmektedir ve m_i mesajlarının üretilme olasılıkları p_i 'dir. $P = \{p_1, p_2, \ldots p_n\}$ olasılık dağılımına sahip mesajları üreten S kaynağının entropisi $H(S)$ şu şekildedir.

$$H(S) = -\sum_{i=1}^{n} p_i \log_2 p_i \qquad (\ )$$

Entropi

(3/3)

- Bir paranın havaya atılması olayı rassal X sürecini göstersin. Yazı ve tura gelme olasılıkları eşit olduğundan elde edilecek entropi,

$$H(X) = -\sum_{i=1}^{n} p_i \log_2 (p_i) = -\left(\frac{1}{2}\log_2\frac{1}{2} + \frac{1}{2}\log_2\frac{1}{2}\right) = 1$$

Örnek

- Aşağıdaki 8 elemanlı S kümesi verilsin.
- S = {evet, hayır, evet, hayır, hayır, hayır, hayır, hayır}
- “evet “ ve “hayır” için olasılık,
- $p(evet) = \frac{2}{8}$, $p(hayır) = \frac{6}{8}$

$$H(S) = -\left(\frac{2}{8}\log_2\frac{2}{8} + \frac{6}{8}\log_2\frac{6}{8}\right) = 0.81128$$

ID3 Algoritması (1/4)

- Karar ağaçları yardımıyla sınıflandırma işlemlerini yerine getirmek üzere Quinlan tarafından birçok algoritma geliştirilmiştir. Bunlar arasında ID3 ve C4.5 algoritması yer almaktadır.
- ID3(Iterative Dichotomiser 3) algoritması sadece *kategorik* verilerle çalışmaktadır.
- Karar ağaçları çok boyutlu veriyi belirlenmiş bir niteliğe göre parçalara böler.
- Her adımda verinin hangi özelliğine göre ne tür işlem yapılacağına karar verilir.
- Oluşturulabilecek tüm ağaçların kombinasyonu çok fazladır.
- Karar ağaçlarının en az düğüm ve yaprak ile oluşturulması için farklı algoritmalar kullanılarak bölme işlemi yapılır.

ID3 Algoritması

(2/4)

- Karar Ağacında Entropi
  - Bir eğitim kümesindeki sınıf niteliğinin alacağı değerler kümesi T, her bir sınıf değeri $C_i$olsun.
  - T sınıf değerini içeren küme için P_T sınıfların olasılık dağılımı,

$$P_T = \left(\frac{|C_1|}{|T|}, \frac{|C_2|}{|T|}, \ldots, \frac{|C_k|}{|T|} \right)$$

şeklinde ifade edilir.

  - T sınıf kümesi için ortalama entropi değeri ise

$$H(T) = -\sum_{i=1}^{n} p_i \log_2 p_i \qquad (\ )$$

şeklinde ifade edilir.

ID3 Algoritması (3/4)

- Karar ağaçlarında bölümlemeye hangi düğümden başlanacağı çok önemlidir.
- Uygun düğümden başlanmazsa ağacın içerisindeki düğümlerin ve yaprakların sayısı çok fazla olacaktır.
- Bir risk kümesi aşağıdaki gibi tanımlansın. C_1="var", C_2="yok"
  - *RISK = {var, var, var, yok, var, yok, yok, var, var, yok}*

$$|C_1| = 6 \qquad |C_2| = 4 \qquad p_1 = 6/10 = 0{,}6 \qquad p_2 = 4/10 = 0{,}4$$

$$P_{RISK} = \left(\frac{6}{10}, \frac{4}{10}\right)$$

$$H(RISK) = -\sum_{i=1}^{n} p_i \log_2(p_i) = -\left(\frac{6}{10}\log_2\frac{6}{10} + \frac{4}{10}\log_2\frac{4}{10}\right) = 0{,}97$$

ID3 Algoritması (4/4)

- Dallanma için niteliklerin seçimi
  - Öncelikle **sınıf niteliğinin entropisi** hesaplanır.

$$H(T) = -\sum_{i=1}^{n} p_i \log_2(p_i)$$

  - Sonra **özellik vektörlerinin sınıfa bağımlı entropileri** hesaplanır.

$$H(X_k) = -\sum_{i=1}^{n} \frac{|T_i|}{|X_k|} \log \frac{|T_i|}{|X_k|} \qquad H(X,T) = \sum_{k=1}^{n} \frac{|X_k|}{|X|} H(X_k)$$

  - Son olarak sınıf niteliğinin entropisinden tüm özellik vektörlerinin entropisi çıkartılarak her özellik için **kazanç ölçütü hesaplanır.**

$$Kazanç(X,T) = H(T) - H(X,T)$$

  - **En büyük kazanca sahip özellik vektörü** o iterasyon için dallanma düğümü olarak seçilir.

Örnek

- Aşağıdaki tablo için karar ağacı oluşturulsun.

MÜŞTERİ	BORÇ	GELİR	STATÜ	RİSK
1	YÜKSEK	YÜKSEK	İŞVEREN	KÖTÜ
2	YÜKSEK	YÜKSEK	ÜCRETLİ	KÖTÜ
3	YÜKSEK	DÜŞÜK	ÜCRETLİ	KÖTÜ
4	DÜŞÜK	DÜŞÜK	ÜCRETLİ	İYİ
5	DÜŞÜK	DÜŞÜK	İŞVEREN	KÖTÜ
6	DÜŞÜK	YÜKSEK	İŞVEREN	İYİ
7	DÜŞÜK	YÜKSEK	ÜCRETLİ	İYİ
8	DÜŞÜK	DÜŞÜK	ÜCRETLİ	İYİ
9	DÜŞÜK	DÜŞÜK	İŞVEREN	KÖTÜ
10	DÜŞÜK	YÜKSEK	İŞVEREN	İYİ

$$H(T) = H(RISK) = -\sum_{i=1}^{n} p_i \log_2(p_i) = -\left(\frac{5}{10}\log_2\frac{5}{10} + \frac{5}{10}\log_2\frac{5}{10}\right) = 1$$

Örnek

$$H(BORÇ_{YÜKSEK}) = -\left(\frac{3}{3}\log_2\frac{3}{3} + \frac{0}{3}\log_2\frac{0}{3}\right) = 0$$

$$H(BORÇ_{DUSUK}) = -\left(\frac{5}{7}\log_2\frac{5}{7} + \frac{2}{7}\log_2\frac{2}{7}\right) = 0{,}863$$

$$H(BORÇ, RISK) = \frac{3}{10}H(BORÇ_{YÜKSEK}) + \frac{7}{10}H(BORÇ_{DUSUK})$$

$$= \frac{3}{10}(0) + \frac{7}{10}(0{,}863) = 0{,}64$$

$$Kazanç(BORÇ, RISK) = 1 - 0{,}64 = 0{,}36$$

Örnek

$$H(GELIR_{YÜKSEK}) = -\left(\frac{2}{5}\log_2\frac{2}{5} + \frac{3}{5}\log_2\frac{3}{5}\right) = 0{,}971$$

$$H(GELIR_{DUSUK}) = -\left(\frac{3}{5}\log_2\frac{3}{5} + \frac{2}{5}\log_2\frac{2}{5}\right) = 0{,}971$$

$$H(GELIR, RISK) = \frac{5}{10}H(GELIR_{YÜKSEK}) + \frac{5}{10}H(GELIR_{DUSUK})$$

$$= \frac{5}{10}(0{,}971) + \frac{5}{10}(0{,}971) = 0{,}971$$

$$Kazanç(GELIR, RISK) = 1 - 0{,}971 = 0{,}029$$

Örnek

$$H(STATU_{ISVEREN}) = -\left(\frac{3}{5}\log_2\frac{3}{5} + \frac{2}{5}\log_2\frac{2}{5}\right) = 0{,}971$$

$$H(STATU_{DUSUK}) = -\left(\frac{3}{5}\log_2\frac{3}{5} + \frac{2}{5}\log_2\frac{2}{5}\right) = 0{,}971$$

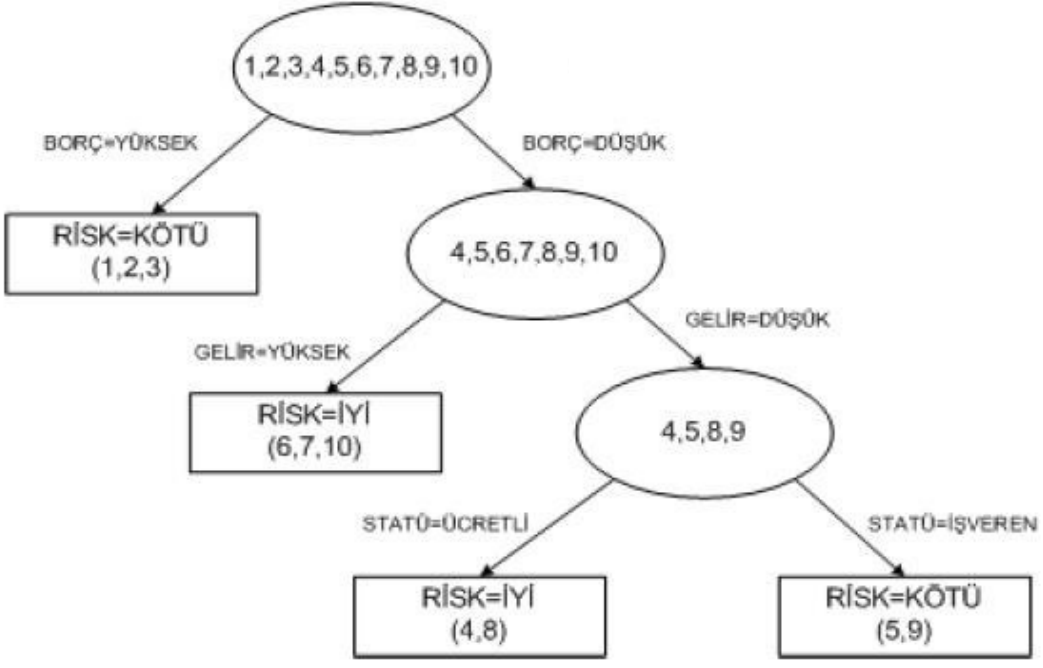
$$H(STATU, RISK) = \frac{5}{10}H(STATU_{YÜKSEK}) + \frac{5}{10}H(STATU_{DUSUK})$$

$$= \frac{5}{10}(0{,}971) + \frac{5}{10}(0{,}971) = 0{,}971$$

$$Kazanç(STATU, RISK) = 1 - 0{,}971 = 0{,}029$$

İlk dallanma için uygun seçim BORÇ niteliğidir.

Örnek


1,2,3,4,5,6,7,8,9,10
BORÇ=YÜKSEK
BORÇ=DÜŞÜK
RİSK=KÖTÜ
(1,2,3)
4,5,6,7,8,9,10
GELİR=YÜKSEK
GELİR=DÜŞÜK
RİSK=İYİ
(6,7,10)
4,5,8,9
STATÜ=ÜCRETLİ
STATÜ=İŞVEREN
RİSK=İYİ
(4,8)
RİSK=KÖTÜ
(5,9)

Örnek

- Karar ağacından elde edilen kurallar
- **1.EĞER**(BORÇ = YÜKSEK) **İSE** (RİSK = KÖTÜ)
- **2.EĞER**(BORÇ = DÜŞÜK) **VE** (GELİR = YÜKSEK) **İSE** (RİSK = İYİ)
- **3.EĞER**(BORÇ = DÜŞÜK) **VE** (GELİR = DÜŞÜK) **VE** (STATÜ = ÜCRETLİ) **İSE** (RİSK = İYİ)
- **4.EĞER**(BORÇ = DÜŞÜK) **VE** (GELİR = DÜŞÜK) **VE** (STATÜ = İŞVEREN) **İSE**(RİSK = KÖTÜ)

Uygulama: Hava problemi örneği

Eğitim kümesi

HAVA	ISI	NEM	RÜZGAR	OYUN
güneşli	sıcak	yüksek	hafif	Hayır
güneşli	sıcak	yüksek	kuvvetli	Hayır
bulutlu	sıcak	yüksek	hafif	Evet
yağmurlu	ılık	yüksek	hafif	Evet
yağmurlu	soğuk	normal	hafif	Evet
yağmurlu	soğuk	normal	kuvvetli	Hayır
bulutlu	soğuk	normal	kuvvetli	Evet
güneş!i	ılık	yüksek	hafif	Hayır
güneşli	soğuk	normal	hafif	Evet
yağmurlu	ılık	normal	hafif	Evet
güneşli	ılık	normal	kuvvetli	Evet
bulutlu	ılık	yüksek	kuvvetli	Evet
bulutlu	sıcak	normal	hafif	Evet
yağmurlu	ılık	yüksek	kuvvetli	Hayır

Uygulama: Hava problemi

- *OYUN = {hayır, hayır, hayır, hayır, hayır, evet, evet, evet, evet, evet, evet, evet, evet, evet}*
- C1, sınıfı **"hayır",** C2, sınıfı ise **"evet"**
- P1=5/14, P2=9/14

$$H(OYUN) = -\left(\frac{5}{14}log_2\frac{5}{14} + \frac{9}{14}log_2\frac{9}{14}\right) = 0.940$$

Adım1: Birinci dallanma

ISI niteliği için kazanç ölçütü:

ISI	OYUN
soğuk	evet
soğuk	hayır
soğuk	evet
soğuk	evet
ılık	evet
ılık	hayır
ılık	evet
ılık	evet
ılık	evet
ılık	hayır
sıcak	hayır
sıcak	hayır
sıcak	evet
sıcak	evet

$$|ISI_{soğuk}| = 4$$

$$|ISI_{ılık}| = 6$$

$$|ISI_{sıcak}| = 4$$

$$H(X,T) = \sum_{i=1}^{n} \frac{|T_i|}{|T|} H(T_i)$$

$$H(ISI_{soğuk}) = -\left(\frac{1}{4}log_2\frac{1}{4} + \frac{3}{4}log_2\frac{3}{4}\right) = 0.811$$

$$H(ISI_{ılık}) = -\left(\frac{2}{6}log_2\frac{2}{6} + \frac{4}{6}log_2\frac{4}{6}\right) = 0.918$$

$$H(ISI_{sıcak}) = -\left(\frac{2}{4}log_2\frac{2}{4} + \frac{2}{4}log_2\frac{2}{4}\right) = 1.00$$

$$H(ISI, OYUN) = \frac{4}{14}(0.811) + \frac{6}{14}(0.918) + \frac{4}{14}(1.00) = 0.911$$

$$H(ISI, OYUN) = \frac{4}{14}H(ISI_{soğuk}) + \frac{6}{14}H(ISI_{ılık}) + \frac{4}{14}H(ISI_{sıcak})$$

$$Kazanç(ISI,OYUN) = H(OYUN) - H(ISI,OYUN)$$
$$= 0.940 - 0.911 = 0.029$$

Adım1: Birinci dallanma

HAVA niteliği için kazanç ölçütü:

$$|HAVA_{güneşli}| = 5 \qquad |HAVA_{yağmurlu}| = 5 \qquad |HAVA_{bulutlu}| = 4$$

$$H(HAVA, OYUN) = \frac{5}{14}H(HAVA_{güneşli}) + \frac{4}{14}H(HAVA_{bulutlu}) + \frac{5}{14}H(HAVA_{yağmurlu})$$

$$H(HAVA_{güneşli}) = -\left(\frac{3}{5}log_2\frac{3}{5} + \frac{2}{5}log_2\frac{2}{5}\right) = 0.971$$

$$H(HAVA_{yağmurlu}) = -\left(\frac{2}{5}log_2\frac{2}{5} + \frac{3}{5}log_2\frac{3}{5}\right) = 0.971$$

$$H(HAVA_{bulutlu}) = -\left(\frac{4}{4}log_2\frac{4}{4}\right) = 0$$

HAVA	OYUN
güneşli	hayır
güneşli	hayır
güneşli	hayır
güneşli	evet
güneşli	evet
yağmurlu	evet
yağmurlu	evet
yağmurlu	hayır
yağmurlu	evet
yağmurlu	hayır
bulutlu	evet
bulutlu	evet
bulutlu	evet
bulutlu	evet

$$H(HAVA, OYUN) = \frac{5}{14}H(HAVA_{güneşli}) + \frac{4}{14}H(HAVA_{bulutlu}) + \frac{5}{14}H(HAVA_{yağmurlu})$$

$$H(HAVA, OYUN) = \frac{5}{14}(0.971) + \frac{4}{14}(0) + \frac{5}{14}(0.971) = 0.694$$

$$Kazanç(HAVA,OYUN) = H(OYUN) - H(HAVA,OYUN)$$
$$= 0.940 - 0.693 = 0.247$$

Adım1: Birinci dallanma

NEM niteliği için kazanç ölçütü:

$|NEM_{yüksek}| = 7$

$|NEM_{normal}| = 7$

$$H(NEM, OYUN) = \frac{7}{14} H(NEM_{yüksek}) + \frac{7}{14} H(NEM_{normal})$$

$$H(NEM_{yüksek}) = -\left(\frac{4}{7} log_2 \frac{4}{7} + \frac{3}{7} log_2 \frac{3}{7}\right) = 0.985$$

$$H(NEM_{normal}) = -\left(\frac{1}{7} log_2 \frac{1}{7} + \frac{6}{7} log_2 \frac{6}{7}\right) = 0.592$$

NEM	OYUN
yüksek	hayır
yüksek	hayır
yüksek	evet
yüksek	evet
yüksek	hayır
yüksek	evet
yüksek	hayır
normal	evet
normal	hayır
normal	evet
normal	evet
normal	evet
normal	evet
normal	evet

$$H(NEM, OYUN) = \frac{7}{14} H(NEM_{yüksek}) + \frac{7}{14} H(NEM_{normal})$$

$$H(NEM, OYUN) = \frac{7}{14}(0.985) + \frac{7}{14}(0.592) = 0.789$$

$$Kazanç(NEM,OYUN) = H(OYUN) - H(NEM,OYUN)$$
$$= 0.940 - 0.789 = 0.151$$

Adım1: Birinci dallanma

RÜZGAR niteliği için kazanç ölçütü:

RÜZGAR	OYUN
hafif	hayır
hafif	evet
hafif	evet
hafif	evet
hafif	hayır
hafif	evet
hafif	evet
hafif	evet
kuvvetli	hayır
kuvvetli	hayır
kuvvetli	evet
kuvvetli	evet
kuvvetli	evet
kuvvetli	hayır

$$|RÜZGAR_{hafif}| = 8$$

$$|RÜZGAR_{kuvvetli}| = 6$$

$$H(RÜZGAR, OYUN) = \frac{8}{14}H(RÜZGAR_{hafif}) + \frac{6}{14}H(RÜZGAR_{kuvvetli})$$

$$H(RÜZGAR_{hafif}) = -\left(\frac{2}{8}log_2\frac{2}{8} + \frac{6}{8}log_2\frac{6}{8}\right) = 0.811$$

$$H(RÜZGAR_{kuvvetli}) = -\left(\frac{3}{6}log_2\frac{3}{6} + \frac{3}{6}log_2\frac{3}{6}\right) = 1.00$$

$$H(RÜZGAR, OYUN) = \frac{8}{14}H(RÜZGAR_{hafif}) + \frac{6}{14}H(RÜZGAR_{kuvvetli})$$

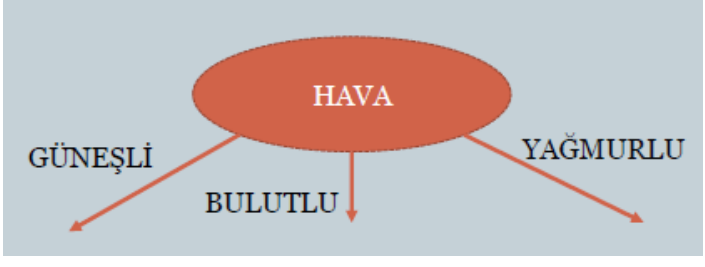
$$H(RÜZGAR, OYUN) = \frac{8}{14}(0.811) + \frac{6}{14}(1.00) = 0.892$$

$$Kazanç(RÜZGAR,OYUN) = H(OYUN) - H(,OYUN)$$
$$= 0.940 - 0.892 = 0.048$$

Nitelik	Kazanç
HAVA	0.246
ISI	0.029
NEM	0.151
RÜZGAR	0.048

Adım1: Birinci dallanma

- Birinci dallanma sonucu karar ağacı:



Adım 2: HAVA niteliğinin "güneşli" değeri için dallanma

HAVA=güneşli için gözlem değerleri				
HAVA	ISI	NEM	RÜZGAR	OYUN
güneşli	sıcak	yüksek	hafif	hayır
güneşli	sıcak	yüksek	kuvvetli	hayır
güneşli	ılık	yüksek	hafif	hayır
güneşli	soğuk	normal	hafif	evet
güneşli	ılık	normal	kuvvetli	evet

Adım 2: HAVA niteliğinin "güneşli" de için dallanma

- Oyun için entropi:

$$H(OYUN) = -\left(\frac{3}{5}log_2\frac{3}{5} + \frac{2}{5}log_2\frac{2}{5}\right) = 0.970$$

Adım 2: HAVA niteliğinin "güneşli" de için dallanma

ISI niteliği için kazanç ölçütü:

$$|ISI_{soğuk}| = 1$$

$$H(ISI_{soğuk}) = -\left(\frac{1}{1}log_2\frac{1}{1}\right) = 0$$

$$H(ISI_{sıcak}) = -\left(\frac{2}{2}log_2\frac{2}{2}\right) = 0$$

$$H(ISI_{ılık}) = -\left(\frac{1}{2}log_2\frac{1}{2} + \frac{1}{2}log_2\frac{1}{2}\right) = 1$$

$$H(ISI, OYUN) = \frac{1}{5}(0) + \frac{1}{5}(0) + \frac{1}{5}(1) = 0.4$$

$$Kazanç(ISI, OYUN) = H(OYUN) - H(ISI, OYUN) = 0.970 - 0.4 = 0.570$$

ISI	OYUN
soğuk	evet
sıcak	hayır
sıcak	hayır
ılık	hayır
ılık	evet

Adım 2: HAVA niteliğinin "güneşli" de için dallanma

NEM niteliği için kazanç ölçütü:

$$H(NEM_{yüksek}) = -\left(\frac{3}{3}log_2\frac{3}{3}\right) = 0$$

$$H(NEM_{normal}) = -\left(\frac{2}{2}log_2\frac{2}{2}\right) = 0$$

$$H(NEM, OYUN) = \frac{3}{5}(0) + \frac{2}{5}(0) = 0$$

NEM	OYUN
yüksek	hayır
yüksek	hayır
yüksek	hayır
normal	evet
normal	evet

$$Kazanç(NEM, OYUN) = H(OYUN) - H(NEM, OYUN) = 0.970 - 0 = 0.970$$

Adım 2: HAVA niteliğinin "güneşli" de için dallanma

RÜZGAR niteliği için kazanç ölçütü:

$$H(RÜZGAR_{hafif}) = -\left(\frac{2}{3}\log_2\frac{2}{3} + \frac{1}{3}\log_2\frac{1}{3}\right) = 0.918$$

$$H(RÜZGAR_{kuvvetli}) = -\left(\frac{1}{2}\log_2\frac{1}{2} + \frac{1}{2}\log_2\frac{1}{2}\right) = 1$$

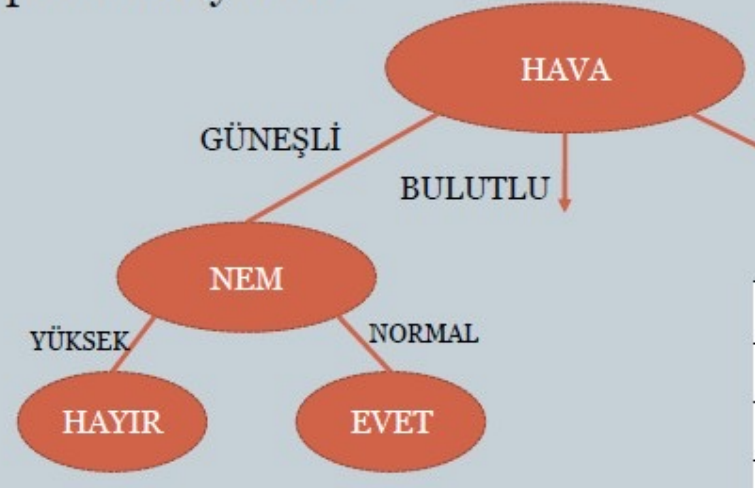
$$H(RÜZGAR, OYUN) = \frac{3}{5}(0.918) + \frac{2}{5}(1) = 0.951$$

RÜZGAR	OYUN
hafif	hayır
hafif	hayır
hafif	evet
kuvvetli	hayır
kuvvetli	evet

$$Kazanç(RÜZGAR, OYUN) = H(OYUN) - H(RÜZGAR, OYUN) = 0.970 - 0.951 = 0.019$$

Adım 2: HAVA niteliğinin "güneşli" de için dallanma

Elde edilen kazanç ölçütlerini aşağıdaki tabloda topluca veriyoruz:



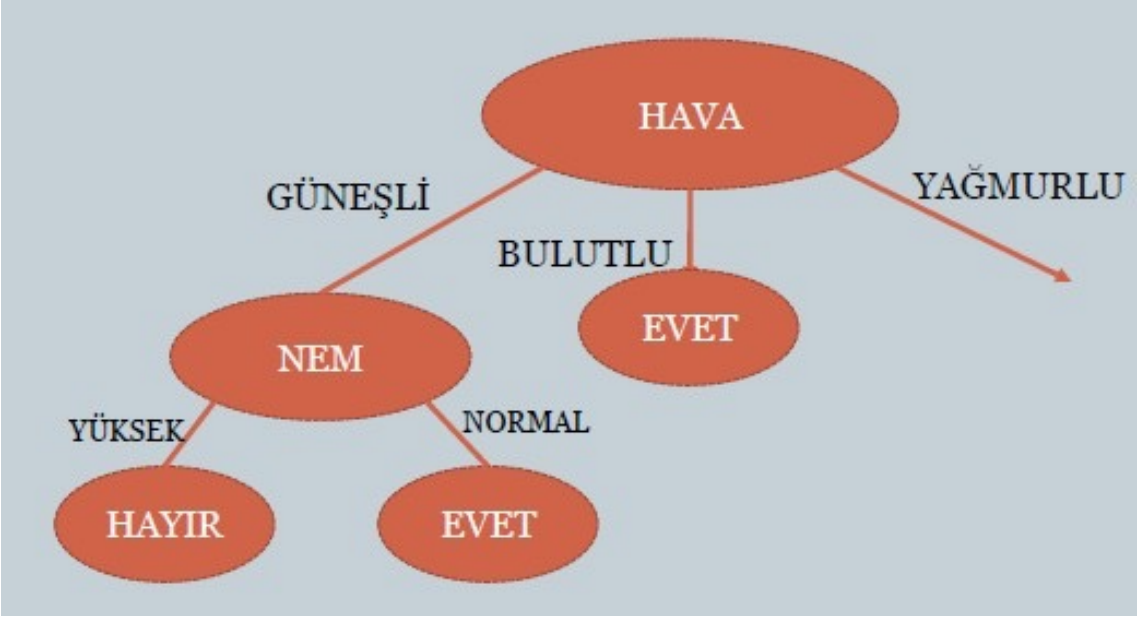
Nitelik	Kazanç
ISI	0.570
NEM	0.970
RÜZGAR	0.019

Adım 3: HAVA niteliğinin “bulutlu” de için dallanma:

Görüldüğü gibi tüm karar değerleri **"evet"** olduğu için herhangi bir analize gerek yoktur.

HAVA	ISI	NEM	RÜZGAR	OYUN
bulutlu	sıcak	yüksek	hafif	evet
bulutlu	soğuk	normal	kuvvetli	evet
bulutlu	ılık	yüksek	kuvvetli	evet
bulutlu	sıcak	normal	hafif	evet

Adım 3: HAVA niteliğinin “bulutlu” de için dallanma:



Adım 3:HAVA niteliğinin "yağmurlu" değeri için dall

OYUN için entropi:

HAVA	ISI	NEM	RÜZGAR	OYUN
yağmurlu	ılık	yüksek	hafif	evet
yağmurlu	soğuk	normal	hafif	evet
yağmurlu	soğuk	normal	kuvvetli	hayır
yağmurlu	ılık	normal	hafif	evet
yağmurlu	ılık	yüksek	kuvvetli	hayır

$$H(OYUN) = -\left(\frac{3}{5}log_2\frac{3}{5} + \frac{2}{5}log_2\frac{2}{5}\right) = 0.970$$

Adım 3:HAVA niteliğinin “yağmurlu” değeri için dall

ISI niteliği için kazanç ölçütü:

ISI	OYUN
soğuk	evet
soğuk	hayır
ılık	evet
ılık	evet
Ilık	hayır

$$|ISI_{soğuk}| = 2 \qquad |ISI_{ılık}| = 3$$

$$H(ISI_{soğuk}) = -\left(\frac{1}{2}log_2\frac{1}{2} + \frac{1}{2}log_2\frac{1}{2}\right) = 1$$

$$H(ISI_{ılık}) = -\left(\frac{2}{3}log_2\frac{2}{3} + \frac{1}{3}log_2\frac{1}{3}\right) = 0.918$$

$$H(ISI, OYUN) = \frac{2}{5}(1) + \frac{3}{5}(0.918) = 0.951$$

$$Kazanç(ISI, OYUN) = H(OYUN) - H(ISI, OYUN) = 0.970 - 0.951 = 0.019$$

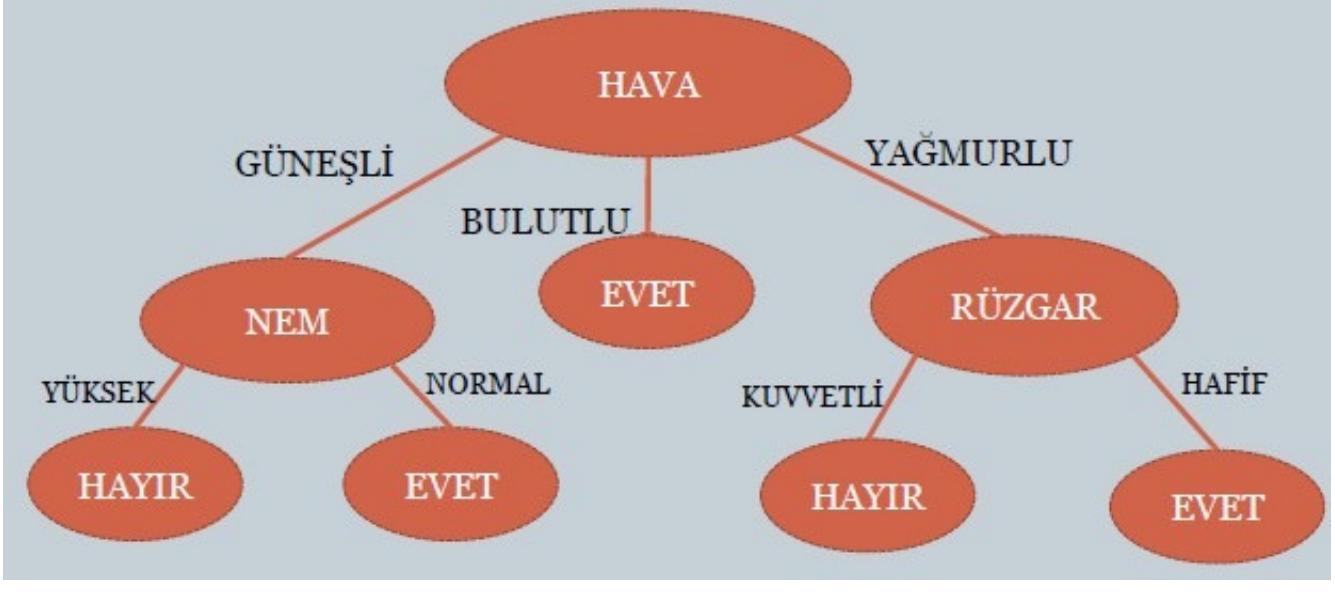
Adım 3:HAVA niteliğinin “yağmurlu” değeri için dall

RÜZGAR niteliği için kazanç ölçütü:

$|RÜZGAR_{hafif}| = 3 \quad |RÜZGAR_{güçlü}| = 2$

RÜZGAR	OYUN
hafif	evet
hafif	evet
hafif	evet
kuvvetli	hayır
kuvvetli	hayır

Oluşturulan Karar Ağacı



C4.5 Algoritması

- C4.5 ile sayısal değerlere sahip nitelikler için karar ağacı oluşturmak için Quinlan tarafından geliştirilmiştir.
- ID3 algoritmasından tek farkı nümerik değerlerin kategorik değerler haline dönüştürülmesidir.
- En büyük bilgi kazancını sağlayacak biçimde bir eşik değer belirlenir.
- Eşik değeri belirlemek için tüm değerler sıralanır ve ikiye bölünür.
- Eşik değer için $[v_i, v_{i+1}]$ aralığının orta noktası alınabilir.

$$t_i = \frac{v_i + v_{i+1}}{2}$$

- Nitelikteki değerler eşik değere göre iki kategoriye ayrılmış olur.

Örnek

NİTELİK1	NİTELİK2	NİTELİK3	SINIF
a	eşit veya küçük	doğru	sınıf1
a	büyük	doğru	sınıf2
a	büyük	yanlış	sınıf2
a	büyük	yanlış	sınıf2
a	eşit veya küçük	yanlış	sınıf1
b	büyük	doğru	sınıf1
b	eşit veya küçük	yanlış	sınıf1
b	eşit veya küçük	doğru	sınıf1
b	eşit veya küçük	yanlış	sınıf1
c	eşit veya küçük	doğru	sınıf2
c	eşit veya küçük	doğru	sınıf2
c	eşit veya küçük	yanlış	sınıf1
c	eşit veya küçük	yanlış	sınıf1
c	büyük	yanlış	sınıf1

Tabloda örneğe ait eğitim kümesi ele alındığında sayısal değerlere sahip olan NİTELİK2 niteliğinin seçilmesi durumunda bilgi kazancının bulunması istenmektedir.

Örnek

Eşik değerinin belirlenmesi

- Nitelik 2 = {65, 70, 75, **80, 85**, 90, 95, 96} için eşik değer **(80+85)/2 = 83** alınmıştır.

NİTELİK1	NİTELİK2	NİTELİK3	SINIF
a	70	doğru	sınıf1
a	90	doğru	sınıf2
a	85	yanlış	sınıf2
a	95	yanlış	sınıf2
a	70	yanlış	sınıf1
b	90	doğru	sınıf1
b	78	yanlış	sınıf1
b	65	doğru	sınıf1
b	75	yanlış	sınıf1
c	80	doğru	sınıf2
c	70	doğru	sınıf2
c	80	yanlış	sınıf1
c	70	yanlış	sınıf1
c	96	yanlış	sınıf1

NİTELİK2≤ 83 veya NİTELİK2>83 testi uygulanarak düzenleme yapıldığında yandaki tablo elde edilir.

Örnek

$$H(SINIF) = -\left(\frac{5}{14}\log_2\frac{5}{14} + \frac{9}{14}\log_2\frac{9}{14}\right) = 0{,}940$$

$$H(NITELIK1_a) = -\left(\frac{2}{5}\log_2\frac{2}{5} + \frac{3}{5}\log_2\frac{3}{5}\right) = 0{,}971$$

$$H(NITELIK1_b) = -\left(\frac{4}{4}\log_2\frac{4}{4} + \frac{0}{4}\log_2\frac{0}{4}\right) = 0$$

$$H(NITELIK1_c) = -\left(\frac{3}{5}\log_2\frac{3}{5} + \frac{2}{5}\log_2\frac{2}{5}\right) = 0{,}971$$

Entropi değerleri ve Bilgi kazancı hesaplanır

b	eşit veya küçük	doğru	sınıf1
b	eşit veya küçük	yanlış	sınıf1
c	eşit veya küçük	doğru	sınıf2
c	eşit veya küçük	doğru	sınıf2
c	eşit veya küçük	yanlış	sınıf1
c	eşit veya küçük	yanlış	sınıf1
c	büyük	yanlış	sınıf1

$$H(NITELIK1, SINIF) = \frac{5}{14}H(NITELIK1_a) + \frac{4}{14}H(NITELIK1_b) + \frac{5}{14}H(NITELIK1_c)$$

$$= \frac{5}{14}0{,}971 + \frac{4}{14}0 + \frac{5}{14}0{,}971 = 0{,}694$$

$$Kazanç(NITELIK1, SINIF) = 0{,}940 - 0{,}694 = 0{,}246$$

Örnek

$$H(NITELIK2_{ek}) = -\left(\frac{7}{9}\log_2\frac{7}{9} + \frac{2}{9}\log_2\frac{2}{9}\right) = 0{,}765$$

$$H(NITELIK2_{b}) = -\left(\frac{2}{5}\log_2\frac{2}{5} + \frac{3}{5}\log_2\frac{3}{5}\right) = 0{,}971$$

$$H(NITELIK2, SINIF) = \frac{9}{14}H(NITELIK2_{ek}) + \frac{5}{14}H(NITELIK1_{b})$$

$$= \frac{9}{14}0{,}765 + \frac{5}{14}0{,}971 = 0{,}836$$

$$Kazanc(NITELIK2, SINIF) = 0.940 - 0.836 = 0.104$$

Örnek

$$H(NITELIK3_d) = -\left(\frac{3}{6}\log_2\frac{3}{6} + \frac{3}{6}\log_2\frac{3}{6}\right) = 1$$

$$H(NITELIK3_y) = -\left(\frac{6}{8}\log_2\frac{6}{8} + \frac{2}{8}\log_2\frac{2}{8}\right) = 0{,}811$$

$$H(NITELIK3, SINIF) = \frac{6}{14}H(NITELIK3_d) + \frac{8}{14}H(NITELIK3_y)$$

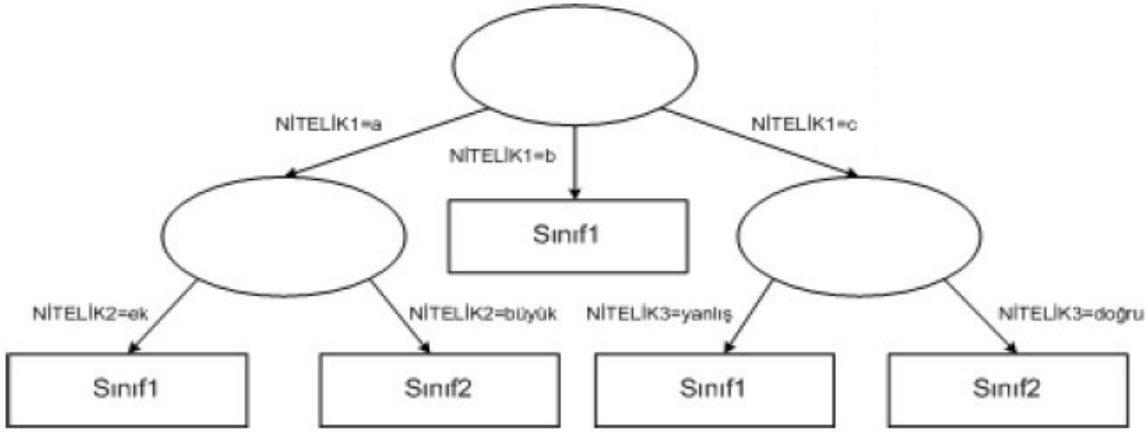
$$= \frac{6}{14}1 + \frac{8}{14}0{,}811 = 0{,}892$$

$$Kazanç(NITELIK\,3, SINIF) = 0{,}940 - 0{,}892 = 0{,}048$$

$$Kazanç(NITELIK\,3, SINIF) < Kazanç(NITELIK\,2, SINIF) < Kazanç(NITELIK1, SINIF)$$

Örnek

Oluşturulan karar ağacı



Örnek

- Karar ağacından elde edilen kurallar
- **1.EĞER**(NİTELİK1 = a) **VE**(NİTELİK2 = Eşit veya Küçük) **İSE**(SINIF = Sınıf1)
- **2.EĞER**(NİTELİK1 = a) **VE**(NİTELİK2 = Büyük) **İSE**(SINIF = Sınıf2)
- **3.EĞER**(NİTELİK1 = b) **İSE**(SINIF = Sınıf1)
- **4.EĞER**(NİTELİK1 = c) **VE**(NİTELİK3 = yanlış) **İSE**(SINIF = Sınıf1)
- **5.EĞER**(NİTELİK1 = c) **VE**(NİTELİK3 = doğru) **İSE**(SINIF = Sınıf2)

VERİ MADENCİLİĞİ

(Karar Ağaçları ile Sınıflandırma)

İçerik

- Sınıflandırma yöntemleri
  - Karar ağaçları ile sınıflandırma
    - Entropi Kavramı
    - ID3 Algoritması
    - C4.5 Algoritması

    } Entropiye dayalı algoritmalar

    - Twoing Algoritması
    - Gini Algoritması

    } Sınıflandırma ve regresyon ağaçları (CART)

    - k-en yakın komşu algoritması } Bellek tabanlı algoritmalar

Sınıflandırma ve Regresyon Ağaçları (CART)

- Sınıflandırma ve regresyon ağaçları veri madenciliğinin sınıflandırma ile ilgili konuları arasında yer alır. Bu yöntem 1984'te Breiman tarafından ortaya atılmıştır. CART karar ağacı, herbir karar düğümünden itibaren ağacın iki dala ayrılması ilkesine dayanır. Yani bu tür karar ağaçlarında ikili dallanmalar söz konusudur.
- CART algoritmasında bir düğümde belirli bir kriter uygulanarak bölünme işlemi gerçekleştirilir. Bunun için önce tüm niteliklerin var olduğu değerler gözönüne alınır ve tüm eşleşmelerden sonra iki bölünme elde edilir. Bu bölünmeler üzerinde seçme işlemi uygulanır. Bu kapsamdaki iki algoritma bulunmaktadır.
  - Twoing Algoritması
  - Gini Algoritması

Twoing Algoritması

- Twoing algoritmasında eğitim kümesi her adımda iki parçaya ayrılarak bölümleme yapılır.
- Aday bölünmelerin sağ ve sol kısımlarının her birisi için nitelik değerinin ilgili sütundaki tekrar sayısı alınır.
- Aday bölünmelerin sağ ve sol kısımlarındaki her bir nitelik değeri için sınıf değerlerinin olma olasılığı hesaplanır.
- Her bölünme için uygunluk değeri en yüksek olan alınır.

$$\Phi(B \mid d) = 2 \frac{|B_{sol}|}{|T|} \frac{|B_{sag}|}{|T|} \sum_{j=1}^{n} abs\left(\frac{|\text{Tsinif}_j|}{|B_{sol}|} - \frac{|\text{Tsinif}_j|}{|B_{sag}|} \right)$$

- Burada, T eğitim kümesindeki kayıt sayısını, B aday bölünmeyi, d düğümü, Tsinif_j ise j.sınıf değerini gösterir.

Örnek

(1/8)

- Tabloda çalışanların maaş, deneyim, görev niteliklerine göre hedef niteliği olan memnun olma durumlarına ait 11 gözlem verilmiştir. Twoing algoritmasını kullanarak sınıflandırma yapınız.

PERSONEL	MAAŞ	DENEYİM	GÖREV	MEMNUN
1	NORMAL	ORTA	UZMAN	EVET
2	YÜKSEK	YOK	UZMAN	EVET
3	DÜŞÜK	YOK	YÖNETİCİ	EVET
4	YÜKSEK	ORTA	YÖNETİCİ	EVET
5	DÜŞÜK	ORTA	YÖNETİCİ	EVET
6	YÜKSEK	İYİ	YÖNETİCİ	EVET
7	DÜŞÜK	İYİ	YÖNETİCİ	EVET
8	YÜKSEK	ORTA	UZMAN	HAYIR
9	DÜŞÜK	ORTA	UZMAN	HAYIR
10	YÜKSEK	İYİ	UZMAN	HAYIR
11	DÜŞÜK	İYİ	UZMAN	HAYIR

Örnek

(2/8)

- Aday bölünmeler aşağıdaki gibidir.

BÖLÜNME	SOL	SAĞ
1	MAAŞ = NORMAL	MAAŞ = {DÜŞÜK, YÜKSEK}
2	MAAŞ = YÜKSEK	MAAŞ = {DÜŞÜK, NORMAL}
3	MAAŞ = DÜŞÜK	MAAŞ = {NORMAL, YÜKSEK}
4	DENEYİM = YOK	DENEYİM = {ORTA, İYİ}
5	DENEYİM = ORTA	DENEYİM = {YOK, İYİ}
6	DENEYİM = İYİ	DENEYİM = {YOK, ORTA}
7	GÖREV = UZMAN	GÖREV = YÖNETİCİ
8	GÖREV = YÖNETİCİ	GÖRE = UZMAN

Örnek

- MAAŞ = NORMAL için

$$P_{sol} = \frac{|B_{sol}|}{|T|} = \frac{1}{11} = 0{,}09 \qquad P_{(EVET/t_{sol})} = \frac{|\text{Tsinif}_{EVET}|}{|B_{sol}|} = \frac{1}{1} = 1 \qquad P_{(HAYIR/t_{sol})} = \frac{|\text{Tsinif}_{HAYIR}|}{|B_{sol}|} = \frac{0}{1} = 0$$

BÖLÜNME	$\|B_{sol}\|$	P_{Sol}	$\|sinif_{EVET}\|$	$\|sinif_{HAYIR}\|$	$P(EVET\|t_{Sol})$	$P(HAYIR\|t_{Sol})$
1	1	0,09	1	0	1	0
2	5	0,45	3	2	0,6	0,4
3	5	0,45	3	2	0,6	0,4
4	2	0,18	2	0	1	0
5	5	0,45	3	2	0,6	0,4
6	4	0,36	2	2	0,5	0,5
7	6	0,55	2	4	0,33	0,67
8	5	0,45	5	0	1	0

Örnek

- MAAŞ = {DÜŞÜK, YÜKSEK} için

$$P_{sag} = \frac{|B_{sag}|}{|T|} = \frac{10}{11} = 0{,}91$$

$$P_{(EVET|t_{sag})} = \frac{|\text{Tsinif}_{EVET}|}{|B_{sag}|} = \frac{6}{10} = 0{,}6$$

$$P_{(HAYIR|t_{sag})} = \frac{|\text{Tsinif}_{HAYIR}|}{|B_{sag}|} = \frac{4}{10} = 0{,}4$$

BÖLÜNME	$\|B_{sag}\|$	P_{sag}	$\|sinif_{EVET}\|$	$\|sinif_{HAYIR}\|$	$P(EVET\|t_{Sag})$	$P(HAYIR\|t_{Sag})$
1	10	0,91	6	4	0,6	0,4
2	6	0,55	4	2	0,67	0,33
3	6	0,55	4	2	0,67	0,33
4	9	0,82	5	4	0,56	0,44
5	6	0,55	4	2	0,67	0,33
6	7	0,64	5	2	0,71	0,29
7	5	0,45	5	0	1	0
8	6	0,55	2	4	0,33	0,67

Örnek

Uygunluk değeri (1. aday bölünme için)

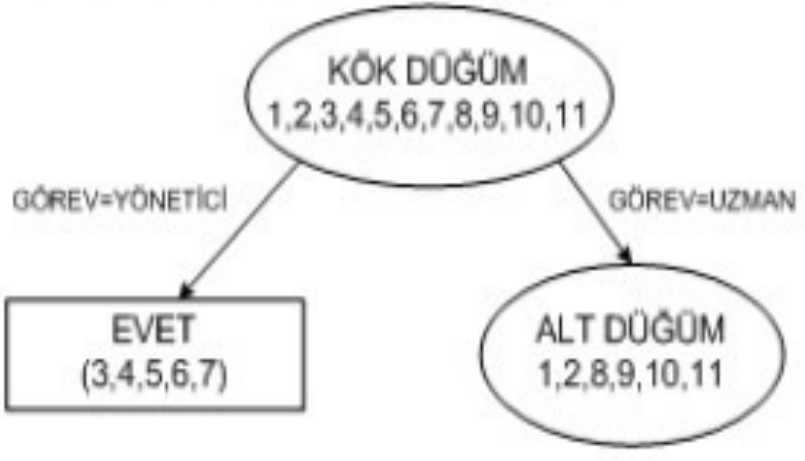
$$\Phi(1|d) = 2\frac{|B_{sol}|}{|T|}\frac{|B_{sag}|}{|T|}\sum_{j=1}^{n} abs\left(\frac{|\text{Tsinif}_j|}{|B_{sol}|} - \frac{|\text{Tsinif}_j|}{|B_{sag}|}\right)$$

$$= 2(0{,}09)(0{,}91)[|1-0{,}6|+|0-0{,}4|] = 0{,}13$$

BÖLÜNME	P_{Sol}	$P_{Sağ}$	$2P_{Sol}P_{Sağ}$	$\Phi(B\|d)$
1	0,09	0,91	0,17	0,13
2	0,45	0,55	0,5	0,07
3	0,45	0,55	0,5	0,07
4	0,18	0,82	0,3	0,26
5	0,45	0,55	0,5	0,07
6	0,36	0,64	0,46	0,2
7	0,55	0,45	0,5	0,66
8	0,45	0,55	0,5	0,66

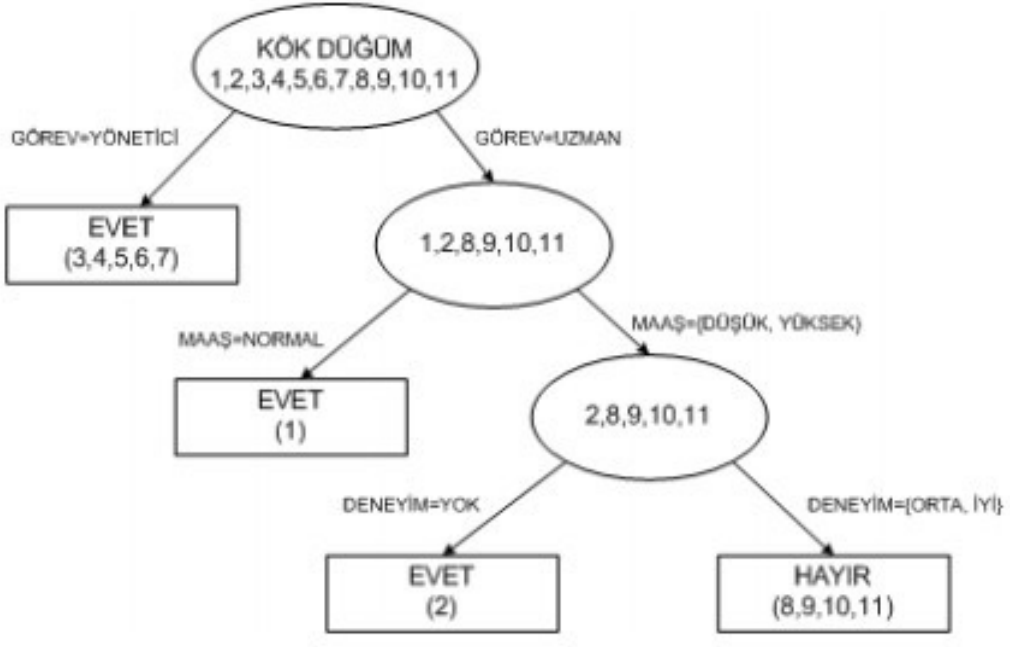
Örnek

- Aynı işlemler ALT DÜĞÜM için tekrarlanır.



Örnek

- Sonuç karar ağacı.



Örnek (8/8)

- Karar ağacından elde edilen kurallar
  - 1. EĞER (GÖREV = YÖNETİCİ) İSE (MEMNUN = EVET)
  - 2. EĞER (GÖREV = UZMAN) VE (MAAŞ = NORMAL) İSE (MEMNUN =EVET)
  - 3. EĞER (GÖREV = UZMAN) VE (MAAŞ = DÜŞÜK VEYA MAAŞ = YÜKSEK) VE (DENEYİM=YOK) İSE (MEMNUN = EVET)
  - 4. EĞER (GÖREV = UZMAN) VE (MAAŞ = DÜŞÜK VEYA MAAŞ = YÜKSEK) VE (DENEYİM = ORTA VEYA DENEYİM = İYİ) İSE (MEMNUN = HAYIR)

Gini Algoritması

- **Gini algoritmasında nitelik değerleri iki parçaya ayrılarak bölümleme yapılır.**
- Her bölünme için $Gini_{sol}$ ve $Gini_{sağ}$ değerleri hesaplanır.

$$Gini_{sol} = 1 - \sum_{i=1}^{k} \left(\frac{|\text{Tsinif}_i|}{|B_{sol}|} \right)^2 \qquad Gini_{sag} = 1 - \sum_{i=1}^{k} \left(\frac{|\text{Tsinif}_i|}{|B_{sag}|} \right)^2$$

- Burada, $Tsinif_i$ soldaki bölümdeki her bir sınıf değerini, $Tsinif_i$ sağdaki bölümdeki her bir sınıf değerini, $|B_{sol}|$ sol bölümdeki tüm değer sayısını, $|B_{sağ}|$ sağ bölümdeki tüm değer sayısını gösterir.

$$Gini_j = \frac{1}{n}\left(|B_{sol}| Gini_{sol} + |B_{sag}| Gini_{sag} \right)$$

- Her bölümlemeden sonra **Gini değeri en küçük olan seçilir.**

Örnek

SIRA	EĞİTİM	YAŞ	CİNSİYET	SONUÇ
1	ORTA	YAŞLI	ERKEK	EVET
2	İLK	GENÇ	ERKEK	HAYIR
3	YÜKSEK	ORTA	KADIN	HAYIR
4	ORTA	ORTA	ERKEK	EVET
5	İLK	ORTA	ERKEK	EVET
6	YÜKSEK	YAŞLI	KADIN	EVET
7	İLK	GENÇ	KADIN	HAYIR
8	ORTA	ORTA	ERKEK	EVET

	EĞİTİM		YAŞ		CİNSİYET	
SONUÇ	İLK	ORTA,YÜKSEK	GENÇ	ORTA,YAŞLI	KADIN	ERKEK
EVET	1	4	0	5	1	4
HAYIR	2	1	2	1	2	1

Örnek

SONUÇ	EĞİTİM		YAŞ		CİNSİYET	
	İLK	ORTA,YÜKSEK	GENÇ	ORTA,YAŞLI	KADIN	ERKEK
EVET	1	4	0	5	1	4
HAYIR	2	1	2	1	2	1

EĞİTİM için

$$Gini_{sol} = 1 - \left[\left(\frac{1}{3}\right)^2 + \left(\frac{2}{3}\right)^2\right] = 0{,}444$$

$$Gini_{sag} = 1 - \left[\left(\frac{4}{5}\right)^2 + \left(\frac{1}{5}\right)^2\right] = 0{,}320$$

Örnek

	EĞİTİM		YAŞ		CİNSİYET	
SONUÇ	İLK	ORTA,YÜKSEK	GENÇ	ORTA,YAŞLI	KADIN	ERKEK
EVET	1	4	0	5	1	4
HAYIR	2	1	2	1	2	1

YAŞ için

$$Gini_{sol} = 1 - \left[\left(\frac{0}{2}\right)^2 + \left(\frac{2}{2}\right)^2\right] = 0$$

$$Gini_{sag} = 1 - \left[\left(\frac{5}{6}\right)^2 + \left(\frac{1}{6}\right)^2\right] = 0{,}278$$

Örnek

	EĞİTİM		YAŞ		CİNSİYET	
SONUÇ	İLK	ORTA,YÜKSEK	GENÇ	ORTA,YAŞLI	KADIN	ERKEK
EVET	1	4	0	5	1	4
HAYIR	2	1	2	1	2	1

CİNSİYET için

$$Gini_{sol} = 1 - \left[\left(\frac{1}{3}\right)^2 + \left(\frac{2}{3}\right)^2\right] = 0{,}444$$

$$Gini_{sag} = 1 - \left[\left(\frac{4}{5}\right)^2 + \left(\frac{1}{5}\right)^2\right] = 0{,}320$$

Örnek

Gini değerleri

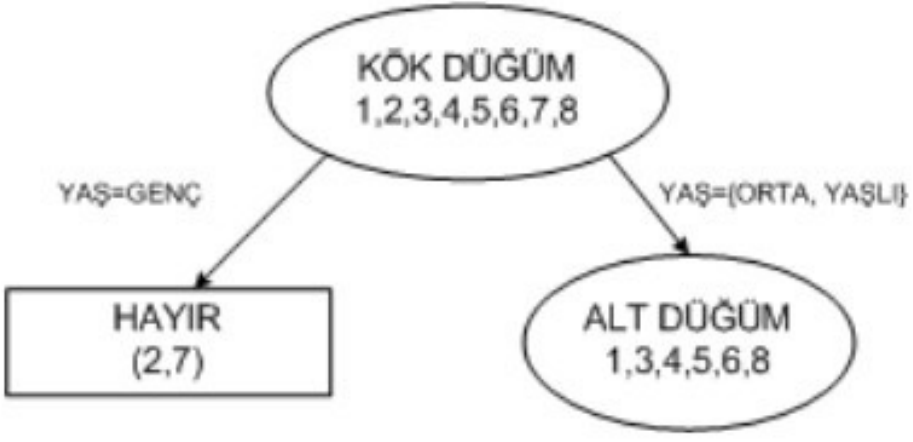
$$Gini_{EGITIM} = \frac{3(0,444) + 5(0,320)}{8} = 0,367$$

$$Gini_{YAS} = \frac{2(0) + 6(0,278)}{8} = 0,209$$

$$Gini_{CINSIYET} = \frac{3(0,444) + 5(0,320)}{8} = 0,367$$

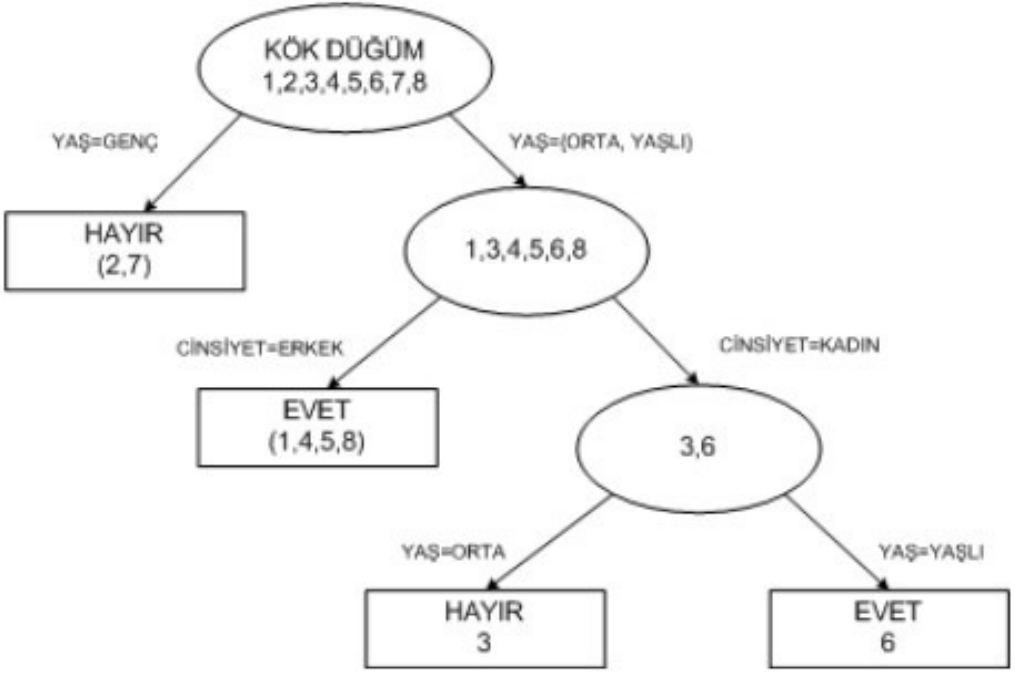
İlk bölünme YAŞ niteliğine göre yapılacaktır.

Örnek



Aynı işlemler ALT DÜĞÜM için tekrarlanır.

Örnek


KÖK DÜĞÜM
1,2,3,4,5,6,7,8
YAŞ=GENÇ
YAŞ={ORTA, YAŞLI}
HAYIR
(2,7)
1,3,4,5,6,8
CİNSİYET=ERKEK
CİNSİYET=KADIN
EVET
(1,4,5,8)
3,6
YAŞ=ORTA
YAŞ=YAŞLI
HAYIR
3
EVET
6

Örnek (8/8)

- Karar ağacından elde edilen kurallar
  - 1. EĞER (YAŞ = GENÇ) İSE (SONUÇ = HAYIR)
  - 2. EĞER (YAŞ = ORTA VEYA YAŞ = YAŞLI) VE (CİNSİYET = ERKEK) İSE (SONUÇ = EVET)
  - 3. EĞER (YAŞ = ORTA VEYA YAŞ = YAŞLI) VE (CİNSİYET = KADIN) VE (YAŞ = YAŞLI) İSE (SONUÇ = EVET)
  - 4. EĞER (YAŞ = ORTA VEYA YAŞ = YAŞLI) VE (CİNSİYET = KADIN) VE (YAŞ = ORTA) İSE (SONUÇ = HAYIR)

Bellek Tabanlı Algoritmalar

- K-en yakın komşu algoritması (K-nearest neighbor algorithm).

K-en yakın komşu algoritması

- Sınıflandırma yöntemlerinden birisi de K-en yakın komşu algoritmasıdır.
- Bu yöntem sınıfları belli olan bir örnek kümesindeki gözlem değerlerinden yararlanarak örneğe katılacak yeni bir gözlemin hangi sınıfa ait olduğunu belirlemek amacıyla kullanılır.
- Bu yöntem örnek kümedeki gözlemlerin her birinin, sonradan belirlenen bir gözlem değerine olan uzaklıklarının ve en küçük uzaklığa sahip k sayıda gözlemin seçilmesi esasına dayanmaktadır. Uzaklıkların hesaplanmasında i ve j noktaları için örneğin Öklid uzaklık formülü kullanılabilir. (Diğer uzaklıklar veri önişleme kısmında açıklanmıştı)

$$d(ij) = \sqrt{\sum_{k=1}^{p} (x_{jk} - x_{ik})^2}$$

K-en yakın komşu algoritması

- K-en yakın komşu algoritması, gözlem değerlerinden oluşan bir küme için aşağıdaki adımları içerir.
  - a) K parametresi belirlenir. Bu parametre verilen bir noktaya en yakın komşuların sayısıdır.
  - b) Bu algoritma verilen bir noktaya en yakın komşuları belirleyeceği için söz konusu nokta ile diğer tüm noktalar arasındaki uzaklıklar tek tek hesaplanır.
  - c) Yukarıda hesaplanan uzaklıklara göre satırlar sıralanır ve bunlar arasından en küçük olan k tanesi seçilir.
  - d) Seçilen satırların hangi kategoriye ait oldukları belirlenir ve en çok tekrarlanan kategori değeri seçilir.
  - e) Seçilen kategori, tahmin edilmesi beklenen gözlem değerinin kategorisi olarak kabul edilir.

Örnek 1.

- Aşağıda verilen gözlem tablosu X1 ve X2 nitelikleri ve Y sınıfından oluşmaktadır. Bu gözlem değerine bağlı olarak yeni bir gözlem değeri olan X1=8, X2=4 değerlerinin yani (8,4) gözleminin hangi sınıfa dahil olduğunu k-en yakın komşu algoritması ile bulunuz.

X1	X2	Y
2	4	KÖTÜ
3	6	İYİ
3	4	İYİ
4	10	KÖTÜ
5	8	KÖTÜ
6	3	İYİ
7	9	İYİ
9	7	KÖTÜ
11	7	KÖTÜ
10	2	KÖTÜ

Örnek 1.

- a) **K'nın belirlenmesi:** k=4 kabul edilir.
- b) **Uzaklıkların hesaplanması:** (8,4) noktası ile gözlem değerlerinin her biri arasındaki uzaklıklar Öklid uzaklığına göre hesaplanır.

$$d(ij) = \sqrt{\sum_{k=1}^{p} \left(x_{ik} - x_{jk} \right)^2}$$

Biçiminde birinci gözlem olan (2,4) noktası ile (8,4) noktası arasındaki uzaklık,

$$d(ij) = \sqrt{(2-8)^2 + (4-4)^2} = 6.00$$

Benzer şekilde uzaklıklar hesaplandığında tablodaki sonuç ortaya çıkacaktır.

Örnek 1.

- (8,4) noktasının gözlem değerlerine olan uzaklıkları,

X1	X2	Uzaklık
2	4	6
3	6	5,39
3	4	5
4	10	7,21
5	8	5
6	3	2,24
7	9	5,1
9	7	3,16
11	7	4,24
10	2	2,83

- c) **En küçük uzaklıkların belirlenmesi:** Satırlar sıralanarak en küçük k=4 tanesi belirlenir. Bu dört nokta verilen (8,4) noktasına en yakın gözlem değerleridir.

X1	X2	Uzaklık	Sıra
2	4	6	9
3	6	5,39	8
3	4	5	6
4	10	7,21	10
5	8	5	5
6	3	2,24	1
7	9	5,1	7
9	7	3,16	3
11	7	4,24	4
10	2	2,83	2

Örnek 1.

- d) **Seçilen satırların ilişkin sınıfların belirlenmesi:** (8,4) noktasına en yakın olan gözlem değerlerinin Y sınıfları göz önüne alınır ve içinde hangi değerin baskın olduğu araştırılır. Bu dört gözlem içinde bir tane **İYİ** 3 tane **KÖTÜ** sınıfı vardır.

X1	X2	Uzaklık	Sıra	k komşunun Y değeri
2	4	6	9	
3	6	5,39	8	
3	4	5	6	
4	10	7,21	10	
5	8	5	5	
6	3	2,24	1	İYİ
7	9	5,1	7	
9	7	3,16	3	KÖTÜ
11	7	4,24	4	KÖTÜ
10	2	2,83	2	KÖTÜ

- e) **Yeni gözlemin sınıfı: KÖTÜ** değerlerinin sayısı **İYİ** değerlerinin sayısından fazla olduğu için (8,4) noktasının sınıfı **KÖTÜ** olarak belirlenir.

Örnek 2.

- Aşağıda verilen gözlem tablosunda Y sınıf niteliğini ifade etmektedir. Bu verilere dayanarak (7,8,5) noktasının hangi sınıf değerine sahip olduğunu belirleyelim. Gözlemlerin gerçek değerleri değil normalize edilmiş değerleri kullanılacaktır. Gözlem değerlerini (0,1) aralığına çekmek için min-max normalleştirmesi kullanılacaktır.

X1	X2	X3	Y
10	5	19	EVET
8	2	4	HAYIR
18	16	6	HAYIR
12	15	8	EVET
3	15	15	EVET

Örnek 2.

- Min-max normalleştirmesi sonucu dönüştürülen değerler aşağıdadır.
- $X^* = \frac{X - X_{min}}{X_{max} - X_{min}}$ (min-max normalizasyonu

X1	X2	X3	Y
0,47	0,21	1	EVET
0,33	0	0	HAYIR
1	1	0,13	HAYIR
0,6	0,93	0,27	EVET
0	0,93	0,73	EVET

- Aday noktanın normalizasyon değeri (0.27,0.43, 0.07)

Örnek 2.

- a) **K'nın belirlenmesi:** k=3 kabul edilir.
- b) **Uzaklıkların hesaplanması:** (0,27, 0,43, 0,07) noktası ile gözlem değerlerinin her biri arasındaki uzaklıklar Öklid uzaklığına göre hesaplanır.

$$d(i,j) = \sqrt{(0{,}47-0{,}27)^2 + (0{,}21-0{,}43)^2 + (1-0{,}07)^2} = 0{,}98$$

X1	X2	X3	Uzaklık
0,47	0,21	1	0,98
0,33	0	0	0,44
1	1	0,13	0,93
0,6	0,93	0,27	0,63
0	0,93	0,73	0,87

Örnek 2.

- c) **En küçük uzaklıkların belirlenmesi:** Satırlar sıralanarak en küçük k=3 tanesi belirlenir.

X1	X2	X3	Uzaklık	Sıra
0,47	0,21	1	0,98	5
0,33	0	0	0,44	1
1	1	0,13	0,93	4
0,6	0,93	0,27	0,63	2
0	0,93	0,73	0,87	3

Örnek 2.

- d) **Seçilen satırların ilişkin sınıfların belirlenmesi:** (0,27, 0,43, 0,07) noktasına en yakın olan gözlem değerlerinin Y sınıfları göz önüne alınır ve içinde hangi değerin baskın olduğu araştırılır. Bu üç gözlem içinde bir tane **HAYIR** 2 tane **EVET** sınıfı vardır.

X1	X2	X3	Uzaklık	Sıra	k komşunun Y değeri
0,47	0,21	1	0,98	5	
0,33	0	0	0,44	1	HAYIR
1	1	0,13	0,93	4	
0,6	0,93	0,27	0,63	2	EVET
0	0,93	0,73	0,87	3	EVET

- e) **Yeni gözlemin sınıfı: EVET** değerlerinin sayısı **HAYIR** değerlerinin sayısından fazla olduğu için (7,8,5) gözleminin sınıfı **EVET** olarak kabul edilir.

Ağırlıklı Oylama

- K-en yakın komşu algoritması sınıfı bilinmeyen gözlem değeri için k gözlem içindeki en fazla tekrar eden sınıfın seçilmesi esasına dayanmaktadır. Ancak seçilen bu sınıf sadece k komşunun göz önüne alınması nedeniyle her zaman uygun olmayabilir. Bu son aşamada k komşu arasında en çok tekrarlanan sınıfı seçme yöntemi yerine **ağırlıklı oylama** (weighted voting) denilen bir yöntem uygulanabilir.
- Söz konusu ağırlıklı oylama yöntemi gözlem değerleri için aşağıdaki bağıntıya göre ağırlıklı uzaklıkların hesaplanmasına dayanır.

$$d(i,j)' = \frac{1}{d(i,j)^2}$$

- *d(i,j)* ifadesi *i* ve *j* gözlemleri arasındaki Öklid uzaklığıdır. Her bir sınıf değeri için bu uzaklıkların toplamı hesaplanarak ağırlıklı oylama değeri elde edilir. En büyük ağırlıklı oylama değerine sahip olan sınıf değeri yeni gözlemin ait olduğu sınıf olarak kabul edilir.

Örnek 2. Ağırlıklı Oylama Sonucu

- Ağırlıklı Oylama sonucunda Örnek 2.'deki değerin sınıfının HAYIR olduğu görülür.

X1	X2	X3	Uzaklık	Sıra	k komşunun Y değeri	Ağırlıklı Oylama
0,47	0,21	1	0,98	5		
0,33	0	0	0,44	1	HAYIR	5,17
1	1	0,13	0,93	4		
0,6	0,93	0,27	0,63	2	EVET	2,52
0	0,93	0,73	0,87	3	EVET	3,84

VERİ MADENCİLİĞİ

(Küm

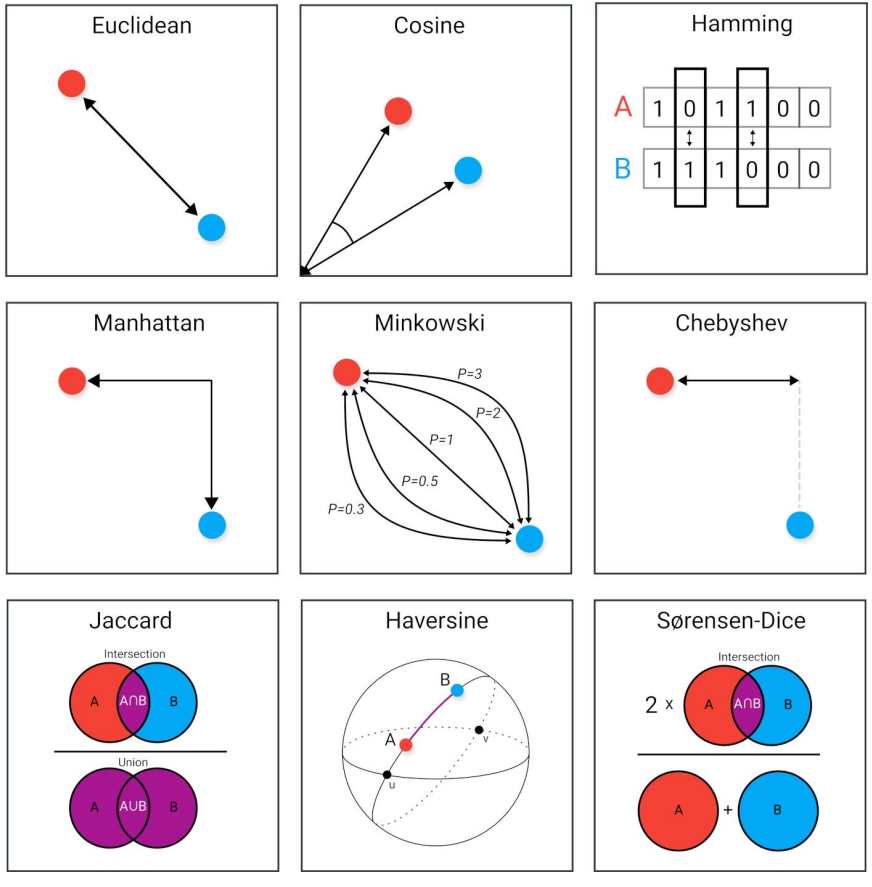
Benzerlik ve Farklılık

Benzerlik ve Farklılık

- Benzerlik
  - iki nesnenin benzerliğini ölçen sayısal değer
  - nesneler birbirine daha benzer ise daha büyük
  - genelde 0-1 aralığında değer alır
- Farklılık
  - iki nesnenin birbirinden ne kadar farklı olduğunu gösteren sayısal değer
  - nesneler birbirine daha benzer ise daha küçük
  - en küçük farklılık genelde 0
  - üst sınır değişebilir.

Uzaklık Çeşitleri

- Öklid(Euclid)
- Minkowski
- Manhattan



Uzaklık Özellikleri

- q=1⇒ Manhattan Uzaklığı
- q=2 ⇒ Öklid Uzaklığı
- Uzaklık ölçütünün sağlaması gereken özellikler:
  1. d(i,j)≥0
  2. d(i,i)=0
  3. d(i,j)=d(j,i)
  4. d(i,j)≤d(i,h)+d(h,j)
- Uzaklıklar ağırlıklı olarak da hesaplanabilir:

$$d(i,j)=\sqrt{w_1\left|x_{i1}-x_{j1}\right|^2+w_2\left|x_{i2}-x_{j2}\right|^2+\ldots+w_p\left|x_{ip}-x_{jp}\right|^2}$$

Benzerlik Özellikleri

- İki nesne arası benzerlik özellikleri
- 1. sim(i,j)>=0
- 2. sim(i,j)=sim(j,i)

İçerik

- Kümeleme İşlemleri
- Kümeleme Tanımı
- Kümeleme Uygulamaları
- Kümeleme Yöntemleri

Kümeleme (Clustering)

- Kümeleme birbirine benzeyen veri parçalarını ayırma işlemidir ve kümeleme yöntemlerinin çoğu veri arasındaki uzaklıkları kullanır.

- Nesneleri kümelere (gruplara) ayırma
- Küme: birbirine benzeyen nesnelerden oluşan grup
  - Aynı kümedeki nesneler birbirine daha çok benzer
  - Farklı kümedeki nesneler birbirine daha az benzer

Kümeleme

- Danışmansız öğrenme: Hangi nesnenin hangi sınıfa ait olduğu ve sınıf sayısı belli değil
- Uygulamaları:
  - verinin dağılımını anlama
  - başka veri madenciliği uygulamaları için ön hazırlık

Kümeleme Uygulamaları

- Örüntü tanıma
- Görüntü işleme
- Ekonomi
- Aykırılıkları belirleme
- WWW
  - Doküman kümeleme
  - Kullanıcı davranışlarını kümeleme
  - Kullanıcıları kümeleme
- Diğer veri madenciliği algoritmaları için bir önişleme adımı
- Veri azaltma – küme içindeki nesnelerin temsil edilmesi için küme merkezlerinin kullanılması

Veri Madenciliğinde Kümeleme

- Ölçeklenebilirlik
- Farklı tipteki niteliklerden oluşan nesneleri kümeleme
- Farklı şekillerdeki kümeleri oluşturabilme
- En az sayıda giriş parametresi gereksinimi
- Hatalı veriler ve aykırılıklardan en az etkilenme
- Model oluşturma sırasında örneklerin sırasından etkilenmeme
- Çok boyutlu veriler üzerinde çalışma
- Kullanıcıların kısıtlarını göz önünde bulundurma
- Sonucun yorumlanabilir ve anlaşılabilir olması

İyi Kümeleme

- İyi kümeleme yöntemiyle elde edilen kümelerin özellikleri
  - aynı küme içindeki nesneler arası benzerlik fazla
  - farklı kümelerde bulunan nesneler arası benzerlik az
- Oluşan kümelerin kalitesi seçilen benzerlik ölçütüne ve bu ölçütün gerçeklenmesine bağlı
  - Uzaklık / Benzerlik nesnelerin nitelik tipine göre değişir
    - Nesneler arası benzerlik: s(i,j)
    - Nesneler arası uzaklık: d(i,j) = 1 − s(i,j)
- İyi bir kümeleme yöntemi veri içinde gizlenmiş örüntüleri bulabilmeli
- Veriyi gruplama için uygun kümeleme kriteri bulunmalı
  - kümeleme= aynı kümedeki nesneler arası benzerliği en büyüten, farklı küm arası benzerliği en küçülten fonksiyon
- Kümeleme sonucunun kalitesi seçilen kümelerin şekline ve temsil edilme yöntemine bağlı

Kümeleme Yöntemlerinde Kullanılan Uzaklıklar

- Öklid

$$d(ij) = \sqrt{\sum_{k=1}^{p} (x_{ik} - x_{jk})^2}$$

- Minkowski

$$d(ij) = \left[\sum_{k=1}^{p} \left(|x_{ik} - x_{jk}| \right)^{\square} \right]^{\frac{1}{\square}}$$

- Manhattan

$$d(ij) = \sum_{k=1}^{p} \left(|x_{ik} - x_{jk}| \right)$$

Kümeleme Yöntemleri

- Hiyerarşik Kümeleme
  - Birleştirici Hiyerarşik Yöntemler
    - En yakın komşu algoritması
    - En uzak komşu algoritması
- Hiyerarşik Olmayan Kümeleme
  - K-Ortalamalar Yöntemi (K-Means)

En yakın komşu algoritması

- En yakın komşu yöntemine «tek bağlantı kümeleme yöntemi» adı da verilmektedir. Başlangıçta tüm gözlem değerleri birer küme olarak değerlendirilir. Adım adım bu kümeler birleştirilerek yeni kümeler elde edilir.
- Bu yöntemde öncelikle gözlemler arasındaki uzaklıklar belirlenir. Öklid uzaklık bağıntısı kullanılabilir.
- Uzaklıklar göz önüne Min d(i,j) seçilir. Söz konusu uzaklıkla ilgili satırlar birleştirilerek yeni bir küme elde edilir. Bu duruma göre uzaklıkların yeniden hesaplanması gerekir.
- Tek bir gözlemden oluşan kümeler arasındaki uzaklıkları doğrudan hesaplayabiliriz. Ancak birden fazla gözlem değerine sahip olan iki küme arasındaki uzaklığın belirlenmesi gerektiğinde farklı bir yol izlenir. İki kümenin içerdiği gözlemler arasında «birbirine en yakın olanların uzaklığı» iki kümenin birbirine olan uzaklığı olarak kabul edilir.

Örnek 1.

- Aşağıdaki tabloda verilen beş gözlem değeri, en yakın komşu algoritması ile kümelenmek isteniyor.

Gözlemler	X_1	X_2
1	4	2
2	6	4
3	5	1
4	10	6
5	11	8

- Adım1. Öncelikle uzaklık tablosu oluşturulur. Her bir gözlemin birbiriyle arasındaki öklid uzaklığı hesaplanır.

Örnek 1.

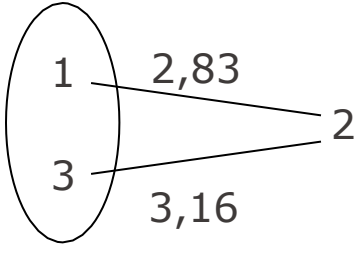
$$d(1,2) = \sqrt{(4-6)^2 + (2-4)^2} = 2,83$$
$$d(1,3) = \sqrt{(4-5)^2 + (2-1)^2} = 1,41$$
$$d(1,4) = \sqrt{(4-10)^2 + (2-6)^2} = 7,21$$

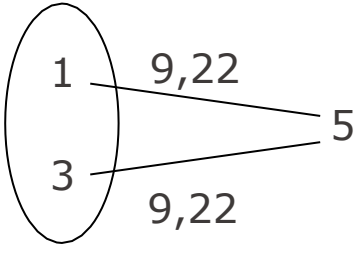
...

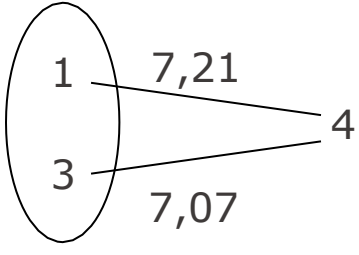
Gözlemler	1	2	3	4	5
1					
2	2,83				
3	1,41	3,16			
4	7,21	4,47	7,07		
5	9,22	6,4	9,22	2,24	

Örnek 1.

- Adım 2. Uzaklıklar tablosunda Min d(i,j) değerinin 1,41 olduğu görülmektedir. İlgili gözlemler 1 ve 3 gözlemleridir. Bu iki değer birleştirilerek (1,3) kümesi elde edilir. Sonrasında bu kümeye göre uzaklıklar matrisi yeniden incelenir.







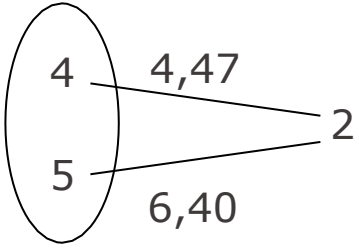
Örnek 1.

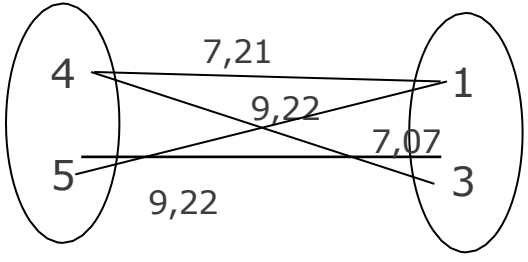
- Yeni uzaklık tablosu,

Gözlemler	(1,3)	2	4	5
(1,3)				
2	2,83			
4	7,07	4,47		
5	9,22	6,4	2,24	

- Bu tabloya bakıldığında *Min d(i,j)=2,24* olduğu görülür. Bu değerin 4 ve 5 gözlemleri arasındaki uzaklığı görülür. (4,5) yeni bir küme oluşturur. Bu durumda (1,3), 2 ve (4,5) kümeleri arasındaki uzaklık tablosu yeniden oluşturulur.

Örnek 1.





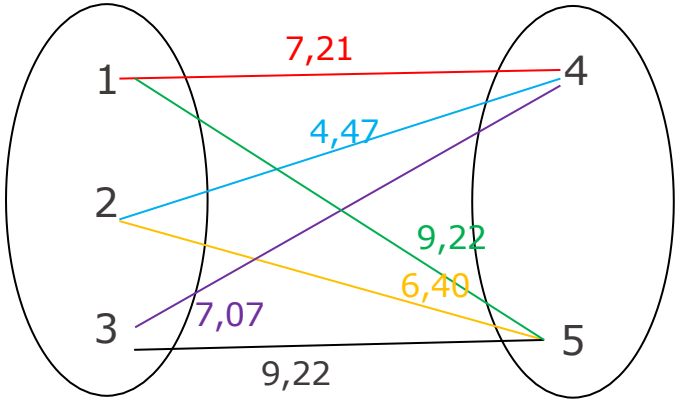
Örnek 1.

- Bu durumdaki uzaklık tablosu,

Gözlemler	(1,3)	2	(4,5)
(1,3)			
2	2,83		
(4,5)	7,07	4,47	

- Adım 4. En son uzaklıklar tablosu incelendiğinde Min d(i,j)=2,83 olduğu görülür. O halde bu uzaklık ile ilgili olan 2 gözlemi ile (1,3) kümesi birleştirilecektir. Elde edilen (1,2,3) kümesi ile (4,5) kümesi arasındaki uzaklığı belirlemek için kümeler içindeki her bir değer eşlenir ve en küçük olan belirlenir. En küçük uzaklık 4,47 olduğuna göre iki küme arasındaki uzaklığın bu değer olduğu kabul edilir.

Örnek 1.



- Adım 5. Elde edilen iki küme birleştirilerek sonuç küme bulunur. Bu küme (1,2,3,4,5) gözlemlerinden oluşan kümedir. Uzaklık düzeyi göz

önüne alınarak kümeler şu şekilde belirlenir.

Uzaklık	Kümeler
1,41	(1,3)
2,24	(4,5)
2,83	(1,2,3)
4,47	(1,2,3,4,5)

En uzak komşu algoritması

- En yakın komşu algoritması ile benzer adımları içerir. Gözlemler arasındaki uzaklıklar hesaplanır ve minimum değerli olan birleştirilir. Sonraki küme uzaklıkları tablosu oluşturulurken en uzak mesafe kullanılır.

K-Ortalamalar Yöntemi (K-Means) (1/2)

- Bu yöntemde daha başlangıçta belli sayıdaki küme için toplam ortalama hatayı minimize etmek amaçlanır.
- N noyutlu uzayda N örnekli kümelerin verildiğini varsayalım. Bu uzay $C_1, C_2, \ldots, C_k$ biçiminde K kümeye ayrılsın. O zaman $n_k = N$ (k=1,2,...,k) olmak üzere C_k kümesinin ortalama vektörü M_k şu şekilde hesaplanır.

$$M_k = \frac{1}{n_k} \sum_{i=1}^{n_k} X_{ik}$$

- Burada X_k değeri C_k kümesine ait olan *i.* örnektir. C_k kümesi için kare-hata, her bir C_k örneği ile onun merkezi (centroid) arasındaki Öklid uzaklıkları toplamıdır. Bu hataya «küme içi değişme» adı da verilir.

K-Ortalamalar Yöntemi (K-Means) (2/2)

- Küme içi değişmeler şu şekilde hesaplanır.

$$e_i^2 = X_{ik} - M(k)^2 \quad \text{(sum over } i=1 \text{ to } n_k\text{)}$$

- K kümesini içeren bütün kümeler uzayı için kare-hata içindeki değişmelerin toplamıdır. O halde söz konusu kare-hata şu şekilde hesaplanır.

$$E_k^2 = \sum_{k=1}^{K} e_k^2$$

- Kare-hata kümeleme yönteminin amacı verilen K değeri için E_k^2 değerini minimize eden K kümelerini bulmaktır. O halde k-ortalamalar algoritmasında E_k^2 değerinin bir önceki iterasyona göre azalması beklenir.

K-Means Algoritmasının Adımları

- K-Means algoritmasına başlamadan önce k küme sayısının belirlenmesi gerekir. Sonra aşağıdaki işlemler gerçekleştirilir.

1. Her bir kümenin merkezi belirlenir. Bu merkezler $M_1, M_2,..,M_k$ biçimindedir.
2. $e_1,e_2,..,e_k$ küme içi değişmeler hesaplanır. Bu değişmelerin toplamı olan $E_k{}^2$ değeri bulunur.
3. M_k merkez değerleri ile gözlem değerleri arasındaki uzaklıklar hesaplanır. Bir gözlem değeri hangi yakın ise o merkez ile ilgili küme içine dahil edilir.
4. Yukarıdaki 2. ve 3. adımlar kümelerde değişiklik olmayıncaya kadar devam ettirilir.

K-Means Algoritmasının Özellikleri

- Gerçeklemesi kolay
- Karmaşıklığı diğer kümeleme yöntemlerine göre az

- K-Means algoritması bazı durumlarda iyi sonuç vermeyebilir
  - Veri grupları farklı boyutlarda ise
  - Veri gruplarının yoğunlukları farklı ise
  - Veri gruplarının şekli küresel değilse
  - Veri içinde aykırılıklar varsa

Örnek 2.

- Aşağıdaki gözlem değerleri k-ortalamalar yöntemi ile kümelenmek isteniyor.

Gözlemler	Değişken1	Değişken2
X_1	4	2
X_2	6	4
X_3	5	1
X_4	10	6
X_5	11	8

- Kümelerin sayısı başlangıçta k=2 kabul edilir. Rasgele iki küme belirlenir.

$$C_1 = \{ X_1, X_2, X_4 \}$$
$$C_2 = \{ X_3, X_5 \}$$

Örnek 2.

Gözlemler	Değişken1	Değişken2	Küme Üyeliği
X_1	4	2	C_1
X_2	6	4	C_1
X_3	5	1	C_2
X_4	10	6	C_1
X_5	11	8	C_2

- Adım 1. a) Belirtilen iki kümenin merkezleri şu şekilde hesaplanır.

$$M_1 = \left\{\frac{4+6+10}{3}, \frac{2+4+6}{3}\right\} = \{6.67, 4.0\}$$

$$M_2 = \left\{\frac{5+11}{2}, \frac{1+8}{2}\right\} = \{8.0, 4.5\}$$

Örnek 2.

- b) Küme içi değişmeler şu şekilde hesaplanır.

$$e_1^2 = [(4-6{,}67)^2 + (2-4{,}0^2)] + [(6-6{,}67)^2 + (4-4{,}0^2)] + [(10-6{,}67)^2 + (6-4{,}0^2)] = 26{,}67$$

$$e_2^2 = [(5-8^2) + (1-4{,}5^2)] + [(11-8)^2 + (8-4{,}5)^2] = 42{,}50$$

- Bu durumda toplam kare-hata şu şekilde hesaplanır.

$$E^2 = e_1^2 + e_2^2 = 26{,}67 + 42{,}50 = 69{,}17$$

Örnek 2.

- C) M_1 ve M_2 merkezlerinden olan uzaklıkların minimum olması istendiğinden aşağıdaki hesaplamalar yapılır. Öklid uzaklık formülü kullanılarak söz konusu mesafeler hesaplanır. Örneğin (M_1, X_1) noktaları arasındaki uzaklık M_1={6.67, 4.00} ve X_1={4, 2} olduğuna göre şu şekilde hesaplanır.

$$d(M_1, X_1) = \sqrt{(6{,}67-4)^2 + (4-2)^2} = 3{,}33$$

$$d(M_2, X_1) = \sqrt{(8-4)^2 + (4{,}5-2)^2} = 4{,}72$$

- Bu işlemler sonucunda X_1 gözlem değerinin M_1 ve M_2 merkezlerine olan uzaklıkları göz önüne alındığında $d(M_1, X_1) < d(M_2, X_1)$ olduğu görülür. Bu durumda M_1 merkezinin X_1 gözlem değerine daha yakın olduğu anlaşılır. O halde $X_1 \in C_1$ olarak kabul edilir. Benzer biçimde tüm gözlem değerleri için tablo oluşturulur.

Örnek 2.

Gözlemler	M_1'den uzaklık	M_2'den uzaklık	Küme Üyeliği
X_1	$d(M_1, X_1) = 3{,}33$	$d(M_2, X_1) = 4{,}72$	C_1
X_2	$d(M_1, X_2) = 0{,}67$	$d(M_2, X_2) = 2{,}06$	C_1
X_3	$d(M_1, X_3) = 3{,}43$	$d(M_2, X_3) = 4{,}61$	C_1
X_4	$d(M_1, X_4) = 3{,}89$	$d(M_2, X_4) = 2{,}50$	C_2
X_5	$d(M_1, X_4) = 5{,}90$	$d(M_2, X_4) = 4{,}61$	C_2

Örnek 2.

- Bu durumda yeni kümeler şu şekilde olacaktır.

$$C_1 = \{ X_1, X_2, X_3 \}$$
$$C_2 = \{ X_4, X_5 \}$$

- Adım 2. Yukarıda belirtilen iki kümenin merkezleri şu şekilde hesaplanır.

$$M_1 = \left\{ \frac{4+6+52+4+1}{3}, \frac{}{3} \right\} = \{5, 2.33 \}$$

$$M_2 = \left\{ \frac{10+116+8}{2}, \frac{}{2} \right\} = \{10.5, 7 \}$$

Örnek 2.

- b) Küme içi değişmeler şu şekilde hesaplanır.

$$e_1^2 = [(4-5^2) + (2-2{,}33^2)] + [(6-5)^2 + (4-2{,}33)^2] + [(5-5)^2 + (1-2{,}33)^2] = 9{,}33$$

$$e_2^2 = [(10-10{,}5)^2 + (6-7)^2] + [(11-10{,}5^2) + (8-7^2)] = 2{,}50$$

- Bu durumda toplam kare-hata şu şekilde hesaplanır.

$$E^2 = e_1^2 + e_2^2 = 9{,}33 + 2{,}50 = 11{,}83$$

- Bu değerin bir önceki iterasyonda elde edilen $E^2 = 69{,}17$ değerinden daha küçük olduğu anlaşılmaktadır.

Örnek 2.

- M_1 ve M_2 merkezlerinden gözlem değerlerine olan uzaklıklar hesaplanır. Bunun sonucunda $d(M_1, X_1) < d(M_2, X_1)$ olduğu görülür. Bu durumda M_1 merkezinin X_1 gözlem değerine daha yakın olduğu anlaşılır. O halde $X_1 \in C_1$ olarak kabul edilir. Benzer biçimde tüm gözlem değerleri için tablo oluşturulur.

Gözlemler	M_1'den uzaklık	M_2'den uzaklık	Küme Üyeliği
X_1	$d(M_1,X_1) = 1{,}05$	$d(M_2,X_1) = 8{,}20$	C_1
X_2	$d(M_1,X_2) = 1{,}94$	$d(M_2,X_2) = 5{,}41$	C_1
X_3	$d(M_1,X_3) = 1{,}33$	$d(M_2,X_3) = 8{,}14$	C_1
X_4	$d(M_1,X_4) = 6{,}20$	$d(M_2,X_4) = 1{,}12$	C_2
X_5	$d(M_1,X_4) = 8{,}25$	$d(M_2,X_4) = 1{,}12$	C_2

Örnek 2.

- Bu durumda yeni kümeler şu şekilde oluşacaktır.

$$C_1 = \{ X_1, X_2, X_3 \}$$
$$C_2 = \{ X_4, X_5 \}$$

- Kümelerde önceki adıma göre herhangi bir değişme olmadığı için iterasyona son verilir.

VERİ MADENCİLİĞİ

(Birliktelik Kuralları)

İçerik

- Birliktelik Kurallarının Tanımı
- Destek ve Güven Ölçütleri
- Apriori Algoritması

Birliktelik Kuralları (Association Rules)

- Birliktelik kuralları
  - Veri kümesi içindeki yaygın örüntülerin, nesneleri oluşturan nitelikler arasındaki ilişkilerin bulunması ☐
- Birliktelik kurallarını kullanma: veri içindeki kuralları belirleme ☐
  - Hangi ürünler çoğunlukla birlikte satılıyor? ☐
  - Kişisel bilgisayar satın alan bir kişinin bir sonraki satın alacağı ürün ne olabilir? ☐
  - Yeni bir ilaca duyarlı olan DNA tipleri hangileridir? ☐
  - Web dokümanları otomatik olarak sınıflandırılabilir mi?

Destek ve Güven Ölçütleri

- Birliktelik çözümlemelerinin en yaygın uygulaması perakende satışlarda müşterilerin satın alma eğilimlerini belirlemek amacıyla yapılmaktadır. Müşterilerin bir anda satın aldığı tüm ürünleri ele alarak satın alma eğilimini ortaya koyan uygulamalara «pazar sepet çözümlemesi» denilir.
- Pazar sepet çözümlemelerinde satılan ürünler arasındaki ilişkileri ortaya koymak için «destek» ve «güven» gibi iki ölçütten yararlanılır. Bu ölçütlerin hesaplanmasında destek sayısı adı verilen bir değer kullanılır. Kural destek ölçütü tüm alışverişler içinde hangi oranda tekrarlandığını belirler.

Destek ve Güven Ölçütleri

- Kural güven ölçütü A ürün grubunu alan müşterilerin B ürün grubunu da alma olasılığını ortaya koyar. A ürün grubunu alanların B ürün grubunu da alma durumu yani birliktelik kuralı A → B biçiminde gösterilir. Bu durumda kural destek ölçütü şu şekilde ifade edilir.

$$destek\ A(\rightarrow B\ ) = \frac{sayı(A,B)}{?}$$

- Burada sayı(A,B) destek sayısı A ve B ürün guruplarını birlikte içeren alışveriş sayısını göstermektedir. N ise tüm alışverişlerin sayısını göstermektedir. A ve B ürün gruplarının birlikte satın alınması olasılığını ifade eden kural güven ölçütü şu şekilde hesaplanır.

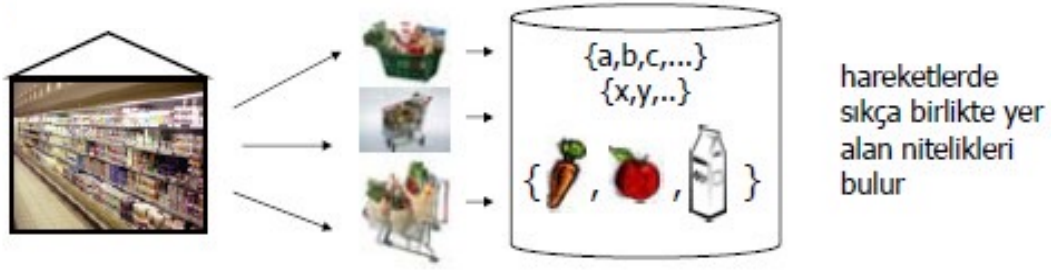
$$güven(A \rightarrow B) = \frac{sayı(A,B)}{sayı(A)}$$

Destek ve Güven Ölçütleri

- Birliktelik kuralları belirlenirken destek ve güven ölçütleri yanı sır bu değerleri karşılaştırmak üzere eşik değere gereksinim vardır. Hesaplanan destek veya güven ölçütlerinin destek(eşik) ve güven(eşik) değerlerinden büyük olması beklenir. Hesaplanan destek veya güven ölçütleri ne kadar büyük ise birliktelik kurallarının da o derce güçlü olduğuna karar verilir.

Birliktelik Kuralları Bulma

- Bir niteliğin (veya nitelikler kümesinin) varlığını harekette bulunan başka niteliklerin varlıklarına dayanarak öngörme



- Kural şekli: **"Gövde ⟶ Baş [destek, güven] "**

satın alma(x, "ekmek") ⟶ satın alma (x, "süt") [%0.6, %65]

öğrenci(x, "BLG"), kayıt(x, "VTYS") ⟶ not(x, "A") [%1, %75]

Birliktelik Kuralları Bulma

- Bütün niteliklerden oluşan küme $I=\{i_1, i_2,\ldots,i_d\}$
  - I={ekmek, süt, bira, kola, yumurta, bez}
- Hareket $T_j \subseteq I$, $Tj=\{i_{j1}, i_{j2},\ldots,i_{jk}\}$
  - T1={ekmek, süt}
- Hareketlerden oluşan veri kümesi $D=\{T_1, T_2,\ldots, T_N\}$

Market Alışveriş verisi

Hareket	*Öğeler*
T1	Ekmek, Süt
T2	Ekmek, Bez, Bira, Yumurta
T3	Süt, Bez, Bira, Kola
T4	Ekmek, Süt, Bez, Bira
T5	Ekmek, Süt, Bez, Kola

Yaygın nitelikler:

Bez, bira
Süt, ekmek, yumurta, kola
Bira, ekmek, süt

Bulunan İlişkilendirme Kuralları

{Bez} → {Bira},
{Süt, Ekmek} → {Yumurta, Kola},
{Bira, Ekmek} → {Süt}

Apriori Algoritması

- Birliktelik kurallarının üretilmesi için birçok yöntem kullanılmaktadır. Bunlardan en yaygın kullanılanı Apriori Algoritmasıdır.
- Apriori algoritması, özellikle çok büyük ölçekli veri tabanları üzerindeki veri madenciliği çalışmalarında geliştirilmiştir. Genel anlamda ilişki kuralı (association rule, birliktelik kuralı) çıkarımında kullanılan bir algoritmadır. Algoritmanın amacı, veri tabanında bulunan satırlar arasındaki bağlantıyı ortaya çıkarmaktır.
- Algoritmanın ismi, kendinden önceki çıkarımlara bağlı olduğu için, latince, önce anlamına gelen "prior" kelimesinden gelmektedir.
- Algoritma yapı olarak, aşağıdan yukarıya (bottom-up) yaklaşımı kullanmakta olup her seferinde tek bir elemanı incelemekte ve bu elemanla diğer adayların ilişkisini ortaya çıkarmaya çalışmaktadır.
- Ayrıca algoritmanın her eleman için çalışmasını, bir arama algoritmasına benzetmek mümkündür. Algoritma, bu anlamda genişlik öncelikli arama (breadth first search) yapısında olup adayları birer ağaç (tree) gibi düşünerek bu ağaç üzerinde arıyor kabul edilebilir.

Apriori Algoritmasının Adımları

- 1. Minimum destek sayısı ve minimum güven değerinin belirlenmesi
- 2. Öğe kümeler içerisindeki her bir öğenin destek değerinin bulunması
- 3. Minimum destek değerinden daha düşük desteğe sahip olan öğelerin devre dışı bırakılması
- 4. Elde edilen tekli birliktelikler dikkate alınarak ikili birlikteliklerin
- oluşturulması
- 5. Minimum destek değerinden düşük olan öğe kümelerin çıkartılması
- 6. Üçlü birlikteliklerin oluşturulması
- 7. Üçlü birlikteliklerden minimum destek değerini geçenlerin dışındakilerin çıkarılması
- 8. Üçlü birlikteliklerden birliktelik kurallarının çıkarılması

Örnek 1.

- Bir mağazada alışveriş yapan müşterilere ilişkin olarak kayıtlar tutulmuş ve beş müşterinin yaptığı alışveriş göz önüne alınmıştır. Müşterilerin bir defada yaptığı alışverişler bir satırda yer almaktadır ve aşağıdaki tabloda verilmiştir. Bu tablodaki veriler kullanılarak müşteri davranışları Apriori Algoritmasıyla ortaya konmak isteniyor.

Müşteri	Aldığı Ürünler
1	Şeker, Çay, Ekmek
2	Ekmek, Peynir, Zeytin, Makarna
3	Şeker, Peynir, Deterjan, Ekmek, Makarna
4	Ekmek, Peynir, Çay, Makarna
5	Peynir, Makarna, Şeker,Bira

Örnek 1.

- a) Çözümlemeye bazı varsayımlarla başlanır. Destek ve güven ölçütleri için eşik değerleri belirlenir.

  destek(eşik)=%60

  güven(eşik)=%75

- Burada destek(eşik)=%60 olduğuna ve tüm müşteri sayısı 5 olduğuna göre **eşik destek sayısının** (0,60)*5=3 olduğu anlaşılır.
- b) Beş müşterinin alışveriş yaptığı ürünlerin kümesi {şeker, çay, ekmek, makarna, peynir, deterjan, bira, zeytin} biçimindedir. Nu ürünlerin her biri için destek değerleri hesaplanır.

sayı(Şeker)=3	sayı(Deterjan)=1
sayı(Çay)=2	sayı(Bira)=1
sayı(Ekmek)=4	sayı(Zeytin)=1
sayı(Makarna)=4	

Örnek 1.

- Destek değerlerinin hesaplanması

Ürün	Sayı
Şeker	3
Çay	2
Ekmek	4
Makarna	4
Peynir	4
Deterjan	1
Bira	1
Zeytin	1

Örnek 1.

- c) Bu tablo üzerinde bazı ürünler eşik değere göre çıkarılır. Eşik destek sayısı 3 olduğuna göre bu eşik değerden küçük desteğe sahip olan ürünler çözümlemeden çıkarılır. Buna göre oluşan yeni tablo aşağıdadır.

Ürün	Sayı
Şeker	3
Ekmek	4
Makarna	4
Peynir	4

Örnek 1.

- d) Çözümlemeye katılacak ürünler bu şekilde belirlendikten sonra
  - ikili gruplar oluşturarak bu grupların destek sayıları hesapla
  - sayı(şeker,ekmek)=2 sayı(şeker,makarna)=2
    sayı(şeker,ekmek)=2 sayı(şeker,peynir)=2
    sayı(ekmek,makarna)=3 sayı(ekmek,peynir)=3
    sayı(makarna,peynir)=4

Örnek 1.

- İkili ürün gruplarının destek değerleri

Ürün	Sayı
Şeker,Ekmek	2
Şeker,Makarna	2
Şeker,Peynir	2
Ekmek,Makarna	3
Ekmek,Peynir	3
Makarna,Peynir	4

Örnek 1.

- e) tablodan bazı eşik değerine göre çıkarılır. B ürünler Buna göre,

Ürün	Sayı
Ekmek,Makarna	3
Ekmek,Peynir	3
Makarna,Peynir	4

Örnek 1.

- f) Çözümlemeye katılacak ürünler bu şekilde belirlendiğine göre bu ürünlerin üçlü gruplar oluşturulur.

sayı(ekmek,makarna,şeker)=1

sayı(ekmek,makarna,çay)=1

sayı(ekmek,makarna,peynir)=3

...

sayı(ekmek,peynir,şeker)=1

sayı(ekmek,peynir,deterjan)=1

...

sayı(makarna,peynir,şeker)=2

sayı(makarna,peynir,çay)=1

...

Örnek 1.

- Eşik destek sayısına göre kalan üçlü ürün grupları aşağıdadır.

Ürün	Sayı
Ekmek,Makarna, Peynir	3

- Bu aşamadan sonra birliktelik kuralları elde edilebilir.

Örnek 1.

sayı(A,B)=sayı(ekmek,makarna,peynir)=3

$$destek(A \rightarrow B) = \frac{sayı(ekmek, makarna, peynir)}{N}$$

$$= \frac{3}{5} = 0.6$$

biçiminde kural destek ölçütü elde edilir. Bu destek ölçütü koşul olarak verdiğimiz eşik değerden küçük değildir. O halde bu nesne kümesini kullanabileceğimiz anlaşılır. Kural destek sayılarına bağlı olarak birliktelik kuralları türeterek bu kurallar için güven ölçütlerini elde edeceğiz.

Sonuç 1:

$$güven(Ekmek, makarna \rightarrow peynir) = \frac{sayı(Ekmek, makarna, peynir)}{sayı(Ekmek, makarna)}$$

$$= \frac{3}{3} = \%100$$

Örnek 1.

Sonuç 2: $güven(Ekmek \rightarrow peynir, makarna) = \frac{sayı(Ekmek, makarna, peynir)}{sayı(Ekmek)}$

$= \frac{3}{4} = \%75$

Sonuç 3: $güven(peynir \rightarrow ekmek, makarna) = \frac{sayı(Ekmek, makarna, peynir)}{sayı(peynir)}$

$= \frac{3}{4} = \%75$

Sonuç 4: $güven(makarna \rightarrow ekmek, peynir) = \frac{sayı(Ekmek, makarna, peynir)}{sayı(makarna)}$

$= \frac{3}{4} = \%75$

Örnek 1.'e ait Birliktelik Kuralları

Birliktelik kuralı	Anlamı	Güven
Ekmek&Makarna→Peynir	Ekmek ve Makarnanın bulunduğu ürün kümesinde Peynirin olma olasılığı	%100
Ekmek→Peynir&Makarna	Ekmeğin yer aldığı bir ürün kümesinde peynir ve makarnanın olma olasılığı	%75
Peynir→Ekmek&Makarna	Peynirin yer aldığı bir ürün kümesinde ekmek ve makarnanın olma olasılığı	%75
Makarna→Ekmek&Peynir	Makarnanın yer aldığı bir ürün kümesinde ekmek ve peynirin olma olasılığı	%75

VERİ MADENCİLİĞİ

(Web Madenciliği)

Kaynak: M. Ali Akcayol, Gazi Üniversitesi, Bilgisayar Mühendisliği Bölümü Ders Notları

İçerik

- İnternet
- World Wide Web
- Web'in Oluşumu
- Web Tarayıcılar
- Web Arama Motorları
- Web Madenciliği
  - Web yapı madenciliği **(Web structure mining)**
  - Web içerik madenciliği **(Web content mining)**
  - Web kullanım madenciliği **(Web usage mining)**

İnternet

- Günümüzde World Wide Web (Kısaca Web) hayatımızın her alanında giderek yaygın bir şekilde kullanılmaktadır.
- **Web, en büyük ve yaygın kullanılan bilgi kaynağı olup arama ve bilgiye erişim hızlı ve kolay bir şekilde yapılabilmektedir.**
- Web üzerinde milyarlarca doküman (Web sayfası) bulunmakta ve milyonlarca kişi sürekli yeni dokümanlar eklemektedir.
- Web, veriye erişimi ve hızlı aramayı sağlamakla birlikte diğer kişilerle bilgi paylaşımını da sağlamaktadır.
- İnternet **diğer kişilerle sesli ve görüntülü görüşme için de kullanılmaktadır.** Bu açıdan **İnternet'in sanal bir topluluk olduğu söylenebilir.**

İnternet

- İnternet günümüzde alışveriş şeklini de değiştirmiştir.
- Mağazaya giderek alışveriş yapmak yerine bilgisayar başında ürünleri almakta ve ödemelerini yapmaktayız.
- Bankacılık, rezervasyon, ödeme başta olmak üzere tüm işlemler elektronik olarak yapılabilmektedir.
- Bu hem maliyet hem de konfor yönünden daha çok tercih edilmektedir.
- **İnternet yaşam kalitesini ve iş yapış şeklimizi de değiştirmiştir.**

World Wide Web

- **Web, kullanıcıların bir bilgisayardan diğer bilgisayarda bulunan veriye ulaşmasını sağlayan İnternet tabanlı bilgisayar ağıdır.**
- Web standart istemci-sunucu (client-server) modelini kullanmaktadır.
- Bu modelde kullanıcılar kendi bilgisayarlarındaki program ile uzaktaki bilgisayar bağlanırlar.
- Web üzerinde gezinti için tarayıcı (browser) denilen programlar kullanılır.
- Browser'lar uzaktaki bilgisayardan istekte bulunurlar ve HTML (HyperText Markup Language) biçiminde gelen bilgiyi yorumlayarak istemci taraftaki kullanıcının ekranında görüntülerler.
- Web üzerinde gezinti yapılırken dokümanlar arasındaki bağlantılar (hyperlink) kullanılır.
- Bu şekilde oluşturulan dokümanlar hypertext olarak adlandırılırlar.

Web'in Oluşumu

- **Web 1989 yılında Tim Berners-Lee tarafından bulunmuştur.** World Wide Web terimini ilk kullanan ve ilk istemci programını yazan kendisidir.
- **Tim Berners-Lee "Information Management: A Proposal" adlı bir öneriyi çalışmakta olduğu CERN laboratuarında 1989 yılında sunmuştur.**
- Bu önerisinde hiyerarşik doküman yapısının avantajlarını ve dezavantajlarını ortaya koymuştur.
- Önerilen doküman yapısıyla bağlantılar (hypertext) aracılığıyla dokümanlar arasında geçiş yapılabilmektedir.
- **Bu öneri dağıtık hypertext sistem olarak adlandırılmıştır ve günümüz Web mimarisinin temelini oluşturmaktadır.**

Web'in Oluşumu

- Başlangıçta destek bulamamış olsa da 1990 yılında Tim-Berners Lee tarafından tekrar önerilmiştir.
- Aynı yıl desteklenen proje ile günümüz Web mimarisi geliştirilmeye başlanmıştır.
- İstemci ve sunucu arasında geliştirilen protokol ile iletişim sağlanmıştır.
- Bu çalışmayla **HyperText Trasfer Protocol (HTTP), HyperText Markup Language (HTML)**ve **Universal Resource Locator (URL)**tanımlanmıştır.

Web Tarayıcılar

Mosaic ve Netscape Browser'lar

- **Web'in önemli gelişmelerinden birisi de 1993 yılında mosaic tarayıcının geliştirilmesidir.**
- Mosaic grafik arayüze sahiptir ve Unix işletim sistemi için geliştirilmiştir. Kısa süre sonra mosaic tarayıcının Windows ve Macintosh versiyonları geliştirilmiştir.
- **1994 yılının ortalarında Netscape tarayıcı geliştirilmiştir.**
- Microsoft tarafından geliştirilen **Internet Explorer tarayıcı 1995 yılında geliştirilmiştir.**
- **Web'in popüler ve başarılı olmasında en önemli aşamalardan birisi Mosaic tarayıcının geliştirilmesidir.**

Web Arama Motorları

- Internet
- Internet, Web'in iletişim ağını sağlar.
- Internet'e ilişkin çalışmalar ARPA (Advanced Research Projects Agency) tarafından desteklenmiştir.
- İlk ARPANET bağlantısı 4 node ile 1969 yılında yapılmıştır.
- 1972 yılında ise 40 node ile bağlantı yapılmıştır.
- 1973 yılında Vinton Cerf ve Bob Kahn tarafından TCP/IP (Transmission Control Protocol / Internet Protocol) protokolünün ilk versiyonu geliştirilmiştir.
- Geliştirilen TCP/IP protokol yığını ile birbirinden uzakta farklı ağlar içinde yer alan bilgisayarlar birbirine bağlanmıştır.
- 1982 yılında TCP/IP protokolünü kullanan Internet doğmuştur.

Web Arama Motorları

Search Engines

- Bilginin Dünya üzerinde dağıtık ve çok büyük boyutlarda bulunmasından dolayı bilgiyi bulmak ve erişmek daha önemli hale gelmeye başladı.
- **Çok büyük bir alanda ve dağıtık bulunan bilginin bulunması için arama motorları geliştirilmeye başlanmıştır.**
- **Excite arama motoru 1993 yılında** 6 Stanford Üniversitesi öğrencisi tarafından **geliştirilmiştir.**
- 1994 yılında EINET Galaxy geliştirilmiştir ve **1994 yılında Yahoo! geliştirilmiştir.**
- Yahoo! diğer alternatiflerine göre favoriler listesi ve öneriler dizini sunmaktaydı.
- Ardından Lycos, Infoseek, Alta Vista, Inktomi, Ask Jeeves, Northernlight gibi arama motorları geliştirilmiştir.

Web Madenciliği

- **Son on yılda Web'in gelişimi sonucunda Dünya'nın en büyük veri kaynağı ortaya çıkmıştır.**
- **Web kendine özgü çok sayıda karakteristik özelliğe sahiptir ve çok büyük veri üzerinde veri madenciliği önemli ve zor bir iş haline gelmiştir.**
- Web üzerindeki veri miktarı çok büyüktür ve gün geçtikçe hızla artmaktadır. Aranan her türlü bilgi Web üzerinde bulunabilmektedir.
- **Web üzerinde yapılandırılmış tablolar, yapılandırılmış Web sayfaları, düz metinler** ve **multimedia dosyaları gibi çok farklı dosyalar bulunmaktadır.**
- **Web üzerindeki veri heterojendir.**

Web Madenciliği

- **Aynı bilgiye sahip Web sayfaları çok farklı biçimlerde ve içeriğe sahip** şekilde Web üzerinde bulunabilmektedir.
- **Bu farklılık Web sayfalarındaki bilgilerin entegrasyonunu çok zor hale getirmektedir.**
- **Web üzerindeki bilginin çok önemli bir kısmı bağlantılara**sahiptir.
- **Hyperlink'ler aynı site üzerindeki Web sayfaları arasında** veya **çok farklı sitelerdeki Web sayfaları arasında olabilmektedir.**
- Hyperlink'ler Web sayfaları için çok önemlidir.
- **Çok sayıda** Web sayfası tarafından **link verilen sayfalar otorite sayfalar**

Web Üzerindeki Verilerin Özellikleri

- **Web üzerindeki bilgi gürültüye sahiptir. Gürültü iki farklı kaynaktan dolayı oluşmaktadır.**
- **Bunlardan birincisi,** Web sayfası **gezinti linkleri**, **reklamlar**, **copyright bilgileri**, **privacy bilgileri**, v.b. gibi çok farklı türde veriye sahiptir.
- İyi bir **Web**bilgisi **analizi için gürültüleri ortadan kaldırmak gereklidir**.
- **İkincisi**, **Web üzerindeki bilginin kalite kontrolü bulunmamaktadır** ve herhangi birisi istediği bilgiyi bir link üzerindeki Web sayfasına yazabilir.
- **Web üzerindeki verinin büyük bir kısmı düşük kalitede, hatalı ve eksiktir.**
- Web üzerinde ticari uygulamalar bulunmaktadır ve insanlar çok sayıda

Web Üzerindeki Verilerin Özellikleri

- **Web üzerindeki bilgi dinamiktir ve sürekli değişmektedir.**
- **Değişiklikleri anlık izlemek bazı uygulamalar için çok önemlidir.**
- **Web sanal bir topluluktur.** Web sadece insanlar arasında veri iletişimini değil **insanlar arasındaki etkileşimi de sağlamaktadır.**
- Yukarıdaki özelliklerin hepsi Web üzerindeki bilginin elde edilmesi için kullanılacak yöntemler için hem fırsatları hem de zorlukları beraberinde getirmektedir.
- **Web madenciliği, veri madenciliğinde kullanılan tüm tekniklerin uygulanmasını içermez.**
- Çok zengin ve farklı özelliklere sahip veriyi bulundurmasından dolayı **Web madenciliği kendine özgü algoritmalara sahiptir.**

Web Madenciliği

- **Web madenciliği kullanılabilir bilgiyi** Web **bağlantılarından**, **sayfa içeriklerinden** ve **kullanılan veriden** elde eder.
- Web madenciliği çok sayıda veri madenciliği tekniğini kullanır ancak **sahip olduğu verinin heterojen olması**, **yarı yapılandırılmış** veya **yapılandırılmamış** olmasından dolayı **sadece veri madenciliği uygulaması olarak görmek doğru değildir.**
- Çok sayıda veri madenciliği yöntemi son on yılda geliştirilmiştir.
- Web mining üç türde ele alınmaktadır. Bunlar;
  - **Web yapı madenciliği**
  - **Web içerik madenciliği**
  - **Web kullanım madenciliği**

  yöntemleridir.

Web yapı madenciliği

- Web yapısı madenciliği **faydalı ve kullanılabilir bilgiyi** Web sayfalarında bulunan **bağlantılardan çıkarır.**
- **Bağlantılar kullanılarak hangi sayfanın daha önemli olduğu gibi bilgiler elde edilebilir.**
- **Ayrıca aynı ortak ilgilere sahip olan benzer kullanıcıları belirleyebiliriz.**
- Klasik **veri madenciliğinde bu tür bilgiler bulunmaz.**

Web içerik madenciliği

- **Web içerik madenciliğinde** faydalı ve kullanılabilir bilgiler **Web sayfalarının içeriğinden elde edilir.**
- Örneğin **Web sayfaları içeriklerine göre sınıflandırılabilir.**
- **Bu özellikler klasik veri madenciliğinde de kullanılmaktadır.**
- Web sayfalarında **kullanıcıların forum bilgilerine müşteri görüşlerine dayanarak çıkarımlar yapılabilmektedir.**

Web kullanım madenciliği

- **Web kullanım madenciliği,** kullanıcıların **Web sayfalarına erişim bilgilerini kullanır.**
- **Kullanıcıların tıklama bilgileri**, **sayfalarda gezinme bilgileri**, **sayfalar üzerindeki etkileşim bilgileri** gibi veriler kullanılır.
- Yukarıdaki işlerin yanı sıra Web üzerindeki verilerin zengin ve çok çeşitli oluşu Web madenciliğinde çok farklı uygulama alanları oluşturmaktadır.
- **Web madenciliği süreci ile veri madenciliği süreci birbirine benzemektedir.** Sadece **veri toplama aşaması farklıdır.**
- **Klasik veri madenciliğinde veriler bir veri ambarında tutulur.**
- **Web madenciliğinde** ise veriler **dağıtık bulunan Web üzerinde bulunur** ve **toplanması çok önemli ve zor bir iştir.**
- Veriler elde edildikten sonra **ön işleme, Web madenciliği ve post- processing** işlemleri gerçekleştirilir.

Web madenciliğinin kullanım alanları

- Web madenciliğinin günümüzde birçok alanda kullanılmasının en önemli sebebi, kişilerin web sayfalarında göstermiş oldukları davranışların, hareketlerin ve yapmış oldukları işlem bilgilerinin var olan iş süreçlerine entegrasyonunu sağlayarak müşterinin en iyi şekilde anlaşılmasını sağlayan müşteri odaklı bir sistem oluşturmasıdır.
- Web madenciliği kullanım alanları aşağıdaki gibidir.
- Web üzerinden ürün satışı gerçekleştiren şirketler web verilerini analiz ederek müşteri profili ve kümeleri oluşturmaktadırlar.
- Google vd. arama motorları web içerik madenciliği uygulayarak aranan anahtar kelimeyi içeren web sitelerini belirlemektedirler.
- Web madenciliği uygulanarak web sitelerinin iyileştirilmesi ve güncel kalması sağlanmaktadır.