EKO 469-Veri Madenciliği

Hafta 4 – Veri Önişleme

Arş. Gör. Dr. Muhammed Kotan

Veri Madenciliği

Süreç

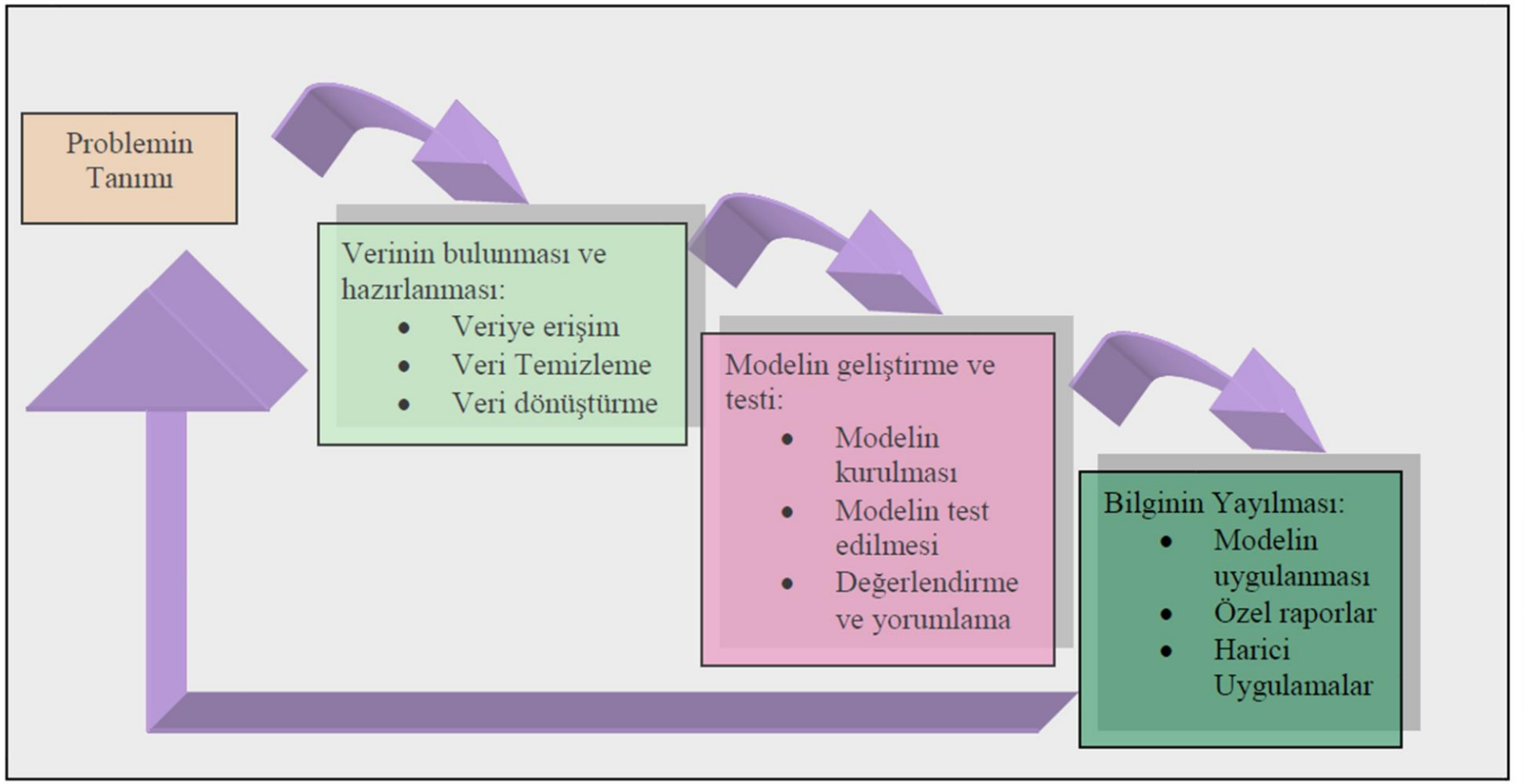
- Veri Madenciliğinin Süreci:
  - Problemin Tanımı
  - Veri seçilmesi ve hazırlanması
  - Verilerin bulunması ve analizi
  - Modelin oluşturulması
  - Modelin geçerliliğinin testi
  - Bilginin üretilmesi, eylem planına dönüştürülmesi ve sonuçların ölçülüp değerlendirilmesi

olarak değerlendirilebilir

Bir başka deyişle veri madenciliğini verinin ***temizlenmesi, bütünleştirilmesi, indirgenmesi, dönüştürülmesi, algoritmanın seçimi ve uygulanması , sunum ve değerlendirme*** biçiminde süreçlerini yazmak da mümkündür

Veri Madenciliği

Süreç



Veri Madenciliği

Veri Hazırlama

- Ham veriler her zaman veri madenciliği uygulamaya hazır halde en iyi veri seti değildir. Veri madenciliği yöntemlerini uygulamaktan çok veri hazırlamak için daha fazla çaba harcanır.

- Verinin hazırlanmasında 2 temel görev vardır:
  - Veriyi standart bir form halinde organize etmek:
    Genellikle standart form ilişkisel tablolardır.

  - Veri setlerinin:
    - Ön İşleme
    - Boyut azaltma

  süreçlerinden geçirilerek en iyi madencilik performansının elde edilmesini sağlamaktır.

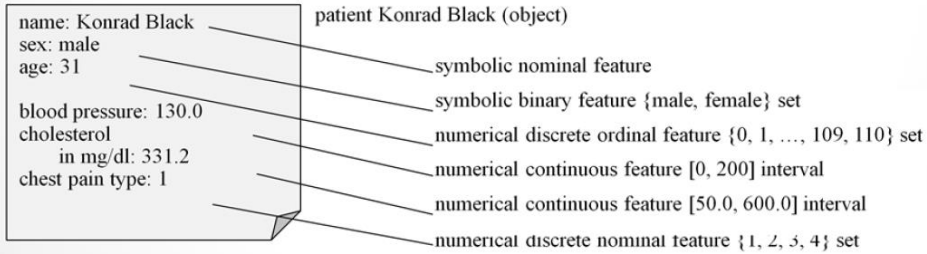
Veri Madenciliği

Veri Hazırlama

Values, Features, and Samples

Example that concerns patients at a heart disease clinic:

- patient is an object (*sample*) that can be described by a number of features, such as name, sex, age, diagnostic test results like blood pressure, cholesterol level, and qualitative evaluations like chest pain and its severity type …



Data Set

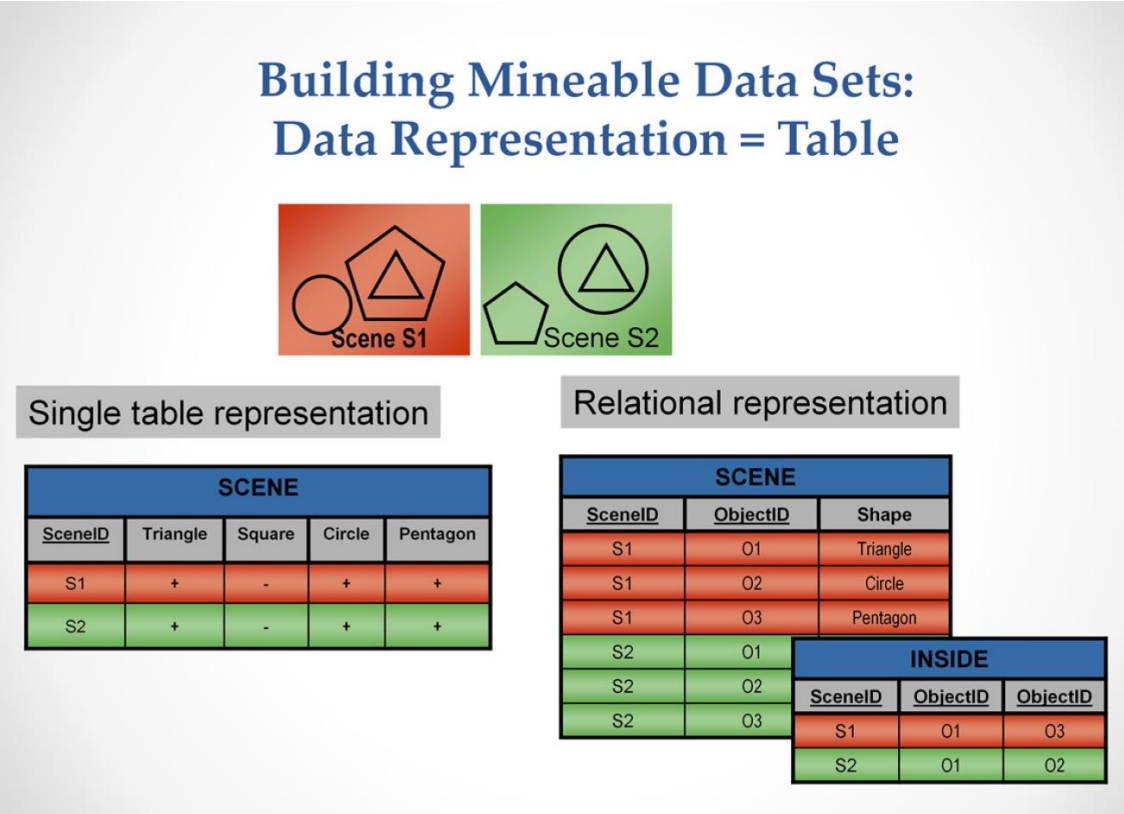
Example that concerns patients at a heart disease clinic:

Name	Age	Sex	Blood pressure	Blood pressure test date	Cholesterol in mg/dl	Cholesterol test date	Chest pain type	Defect type	Diagnosis
Konrad Black	31	male	130.0	05/05/2005	NULL	NULL	NULL	NULL	NULL
Konrad Black	31	male	130.0	05/05/2005	331.2	05/21/2005	1	normal	absent
Magda Doe	26	female	115.0	01/03/2002	NULL	NULL	4	fixed	present
Magda Doe	26	female	115.0	01/03/2002	407.5	06/22/2005	NULL	NULL	NULL
Anna White	56	female	120.0	12/30/1999	45.0	12/30/1999	2	normal	absent
…	…	…	…	…	…	…	…	…	…

Veri Madenciliği

Veri Hazırlama

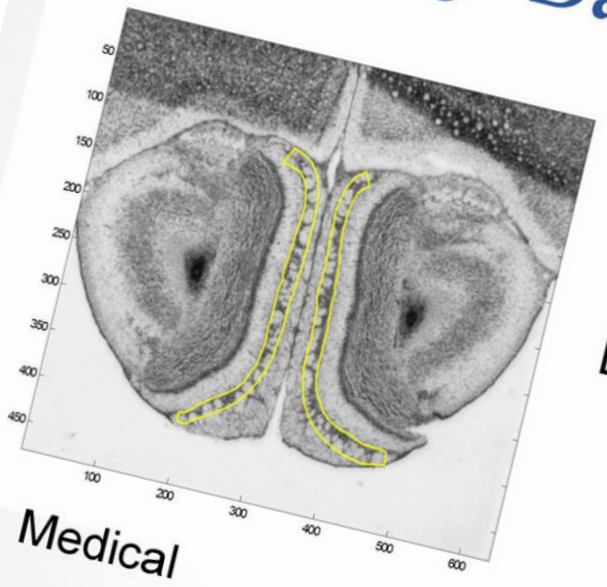
- Birçok veri seti başlangıçta bir tablo biçiminde değildir ve multimedya verilerinin dönüştürülmesi çok daha karmaşıktır.



Veri Madenciliği

Veri Hazırlama

Image Data -> Table?



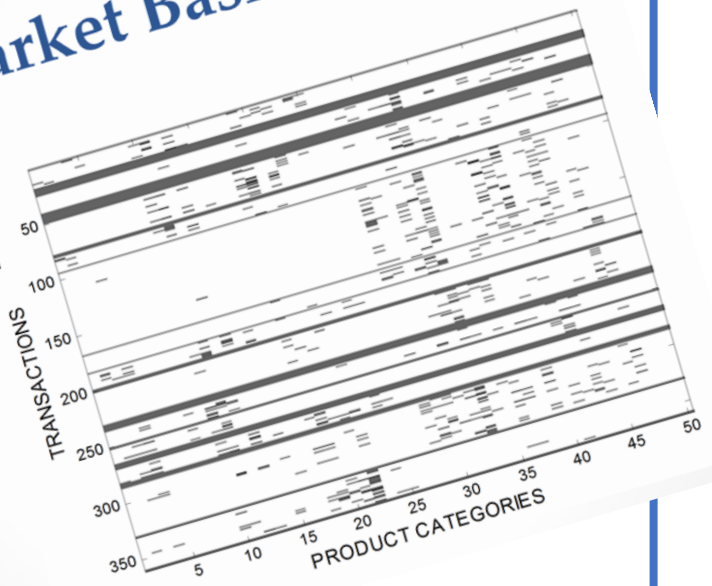
Medical

X coor.	Y coor.	RGB value
100	250	87
100	255	85
...	...	...

Geographical

Market Basket Data

Trans. ID	Products
01	01, 03, 44, 76
02	22, 37, 76
...	...



Verilerin Hazırlanması

- HAM VERİ = DAĞINIK VERİ
  - *Kayıp,
  - *Yanlış kaydedilmiş,
  - *Veri başka bir popülasyondan olabilir (heterojen)
  - * Farklı yapılar, biçimler
  - *Sıkıştırılmış veya sıkıştırılmamış
  - *Gereksiz
  - …

Ham (Gerçek dünya) verileri Ön İşlemeye ihtiyaç duyar.

Verilerin Hazırlanması

- Ham verileri bir tablo biçiminde dönüştürdüğümüzü varsayalım.

Veri madenciliğinde bir sonraki adım nedir?

- Başlangıçta veri madenciliği için hazırlanan tüm ham veri setleri genellikle büyüktür; birçoğu insanlarla ilgilidir ve dağınık olma potansiyeline sahiptir.

- Bir öncül olarak, bu ilk veri setlerinde eksik değerler, bozulmalar, yanlış kayıt, yetersiz örnekleme vb. bulunması beklenmelidir.

- Bu sorunlardan hiçbirini göstermeyen ham veriler şüphe uyandırması gerekebilir. Bu nedenle, ek ön işleme gereklidir.

Verilerin Hazırlanması

- Günümüzün veritabanları, çok büyük boyutlu olmaları ve farklı kaynaklardan beslenmeleri nedeni ile gürültülü, eksik ve tutarsız verilere karşı oldukça açıktır.
- Düşük kaliteli veriler ile yüksek kalitedeki veri madenciliği sonuçlarına ulaşmak imkansızdır.
- Madencilik sürecinin verimliliğini arttırmak ve işlemleri kolaylaştırmak için öncelikle veri kalitesi arttırılmalıdır, yani veriler ön işleme tabi tutulmalıdır.

Veri Kalitesi

- Veri kalitesini belirleyen faktörler:
  - Kesinlik (Accuracy): Tüm veriler doğru mu?
  - Tamamlık (Completeness): Eksik yada ulaşılamayan veri var mı?
  - Tutarlılık (Consistency): Farklı kaynaklardan alınan veriler uyumlu mu?
  - Güncellik (Timeliness): Veri girişi ve güncellemeler zamanında yapıldı mı?
  - İnandırıcılık (Believability): Veri ne kadar güvenilir?
  - Yorumlanabilirlik (Interpretability): Veri kolay anlaşılıyor mu?.

Veri ön işlemede başlıca görevler

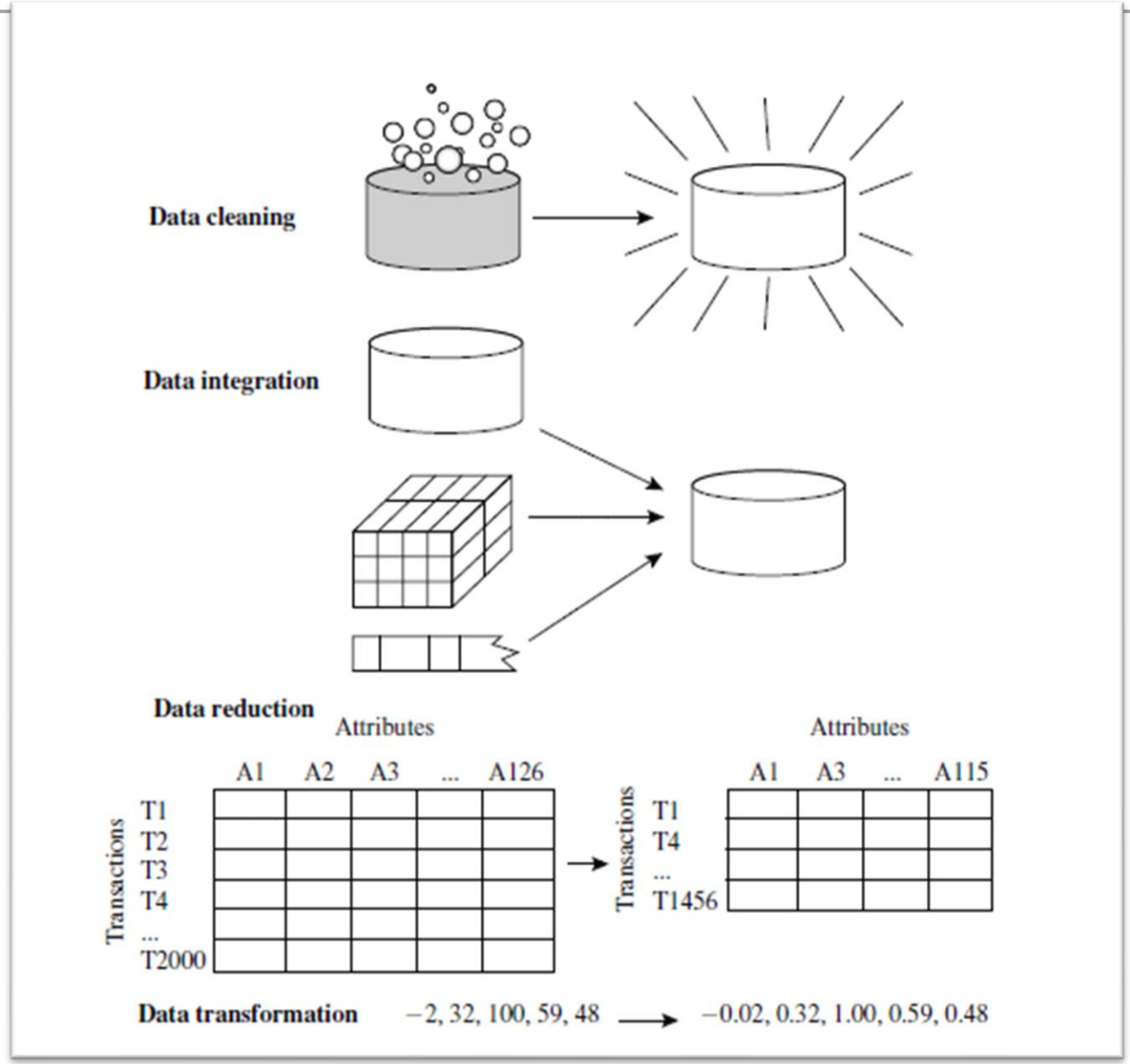
- **Veri Temizleme (Data Cleaning)**
  Eksik verilerin doldurulması, gürültülü verilerin düzeltilmesi, aykırı verilerin (outlier) belirlenmesi ve temizlenmesi, uyuşmazlıkların (inconsistencies) çözümlenmesi

- **Veri Bütünleştirme (Data Integration)**
  Farklı veri kaynaklarının, Veri Küplerinin veya Dosyaların entegre olması

- **Veri İndirgeme (Data Reduction)**
  Boyut Küçültme (Dimensionality reduction)
  Sayısal Küçültme (Numerosity reduction)
  Verinin Sıkıştırılması (Data compression)

- **Veri Dönüşümü ve Veri Ayrıştırma (Data Transformation and Data Discretization)**
  Normalleştirme (Normalization)
  Kavram Hiyerarşisi (Concept hierarchy generation)

Veri ön işlemede başlıca görevler



1. Veri Temizleme

- Verinin temizlenmesi, kirli veri olarak da adlandırılan kayıp ve gürültünün ortadan kaldırılmasıdır.
- Ayrıca yanlış ve aşırı uçta bulunan verilerin ortadan kaldırılması da verilerin temizlenmesi konusuna girer.
  - Meslek= “ ” (girilmemiş)
  - Maaş= “−10” (hata)
  - Yaş= “42” , Doğum Tarihi= “03/07/2010” (tutarsız veri)
  - Eski puanlama “1, 2, 3”, yeni puanlama “A, B, C” (tutarsız veri)
  - Bir veri kaynağında öznitelik “ Ad”, diğer kaynakta “ İsim”
  - Doğum tarihi bilinmeyen herkese 1 Ocak yazılması (gizlenmiş eksik veri)

- Bazı durumlarda, eksik bir değerin verilerde bir hataya yol açmayabileceğini not etmek önemlidir!
- Örneğin, bir kredi kartı için başvururken, adaylardan sürücü ehliyet numarası vermeleri istenebilir. Ehliyet sahibi olmayan adaylar doğal olarak bu alanı boş bırakabilir.

1. Veri Temizleme

Kayıp Veriler

1. **Kayıp verinin bulunduğu kaydı veritabanından -veri kümesinden- çıkarmak yada bu gibi kayıtları iptal etmek**

   Eğer kayıp verili kayıt sayısı, toplam kayıt sayısına oranlandığında, sonuçlar hassasiyetini etkilemeyecek kadar küçükse bu yöntem kullanılabilir. Aksi takdirde bu yöntem yarardan çok zarar getirecektir.

2. **Kayıp verileri elle teker teker doldurmak**

   Kullanılan veri kümesi küçük ve gerçek hayattan kayıp verilere ulaşmak mümkün ise, yeteri kadar zaman varsa ve bu verilere kesinlikle ihtiyaç varsa bu yöntem kullanılabilir. Aksi takdirde zaman kaybı olacaktır.

1. Veri Temizleme

Kayıp Veriler

3. Tüm kayıp verilere aynı bilgiyi girmek

•Tüm kayıp verilere ‘‘bilinmiyor’’ gibi ortak bir değer girilebilir. Ancak bu yöntem ilginç ve farklı sonuçlar doğurabilir.

•Medeni hali boş olan tüm kayıtlara medeni hal anlamına gelen MH girilebilir. Yapılan çalışma sonunda medeni halin MH olması anlamlı bir sonuçmuş gibi çıkabilir, yani kullanılan bu tür veri girişleri algoritmayı yanıltabilir.

•Bunun yanında medeni halini boş bırakan tüketicilerin alışveriş alışkanlıklarının incelenmesi işletme için yararlı bilgilerin keşfedilmesini sağlayabilir. Bu kişiler medeni hal bilgisini özellikle boş bırakmış olabilirler, bu durum alışveriş tercihleri ile birlikte değerlendirildiğinde faydalı bilgiler elde edilebilir.

1. Veri Temizleme

Kayıp Veriler

4. Kayıp olan verilere tüm verilerin ortalama değerinin verilmesi

•Gelir bilgisi eksik olan kayıtlara diğer gelir değerlerinin ortalaması verilebilir. Ortalama hesabı yapılırken tüm kayıtların ortalaması yerine sadece ilgili sınıfın ortalamasının dikkate alınması daha uygun bir yöntemdir.

•Aşağıdaki tabloda 4. kayıt için eksik veri yerine veri girişi yaparken tüm kayıtların ortalaması alınırsa yada sınıf olarak branş veya yaş seçimi yapılırsa farklı değerler elde edilecektir.

Oyuncu ID	Yaş	Gelir (TL)	Branş
1	21	12000	Voleybol
2	35	10000	Futbol
3	44	39000	Futbol
4	32	???	Futbol
5	33	25000	Basketbol
6	52	17500	Yüzme

1. Veri Temizleme

Kayıp Veriler

5. **Diğer değişkenlerin yardımı ile kayıp verilerin tahmini**

Eldeki verilerden yararlanılarak

- Doğrusal interpolasyon,
- Bayesyen sınıflandırma,
- Karar ağaçları,
- Maksimum beklenti,
- Jackknife

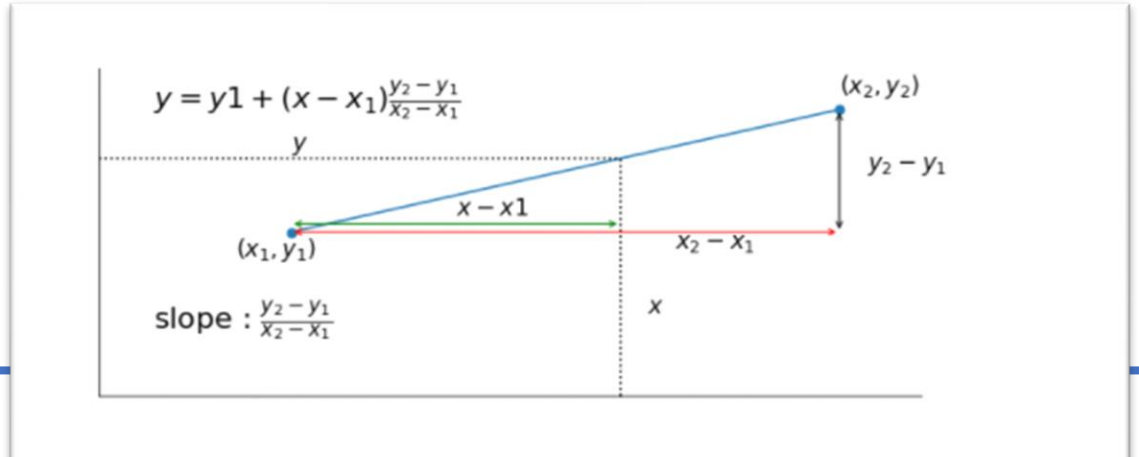
gibi veri madenciliği teknikleri ile kayıp veriler tahmin edilebilir.

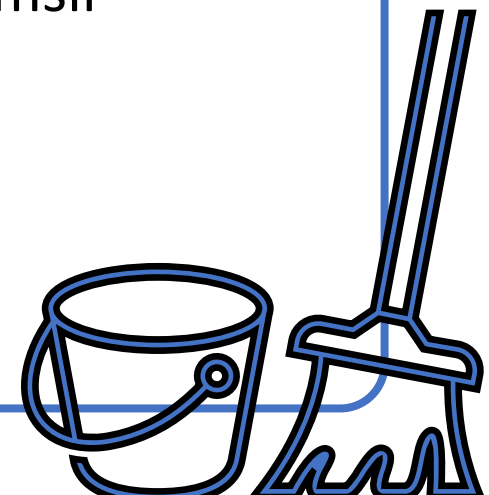
1. Veri Temizleme

Diğer değişkenlerin yardımı ile kayıp verilerin tahmini

Doğrusal interpolasyon

- Bu yöntem doğrusal eğri birleştirme mantığına dayanır.
- Yöntemin uygulanabilmesi için verilerin belli bir sınıf değerine bağlı olarak doğrusal bir şekilde değişiyor olması gerekmektedir.
- Örneğin öğrencilerin yaşlarının okudukları sınıfa göre değişmesi, kandaki potasyum miktarı ile sodyum miktarının doğrusal ilişkisi gibi
- Verilen denklemde y tahmin edilen eksik değer, x doğrusal ilişkili sütunda y ye karşılık gelen değer, x_1, y_1, x_2, y_2 ise y ye en yakın komşu noktaları temsil etmektedir.





1. Veri Temizleme

Diğer değişkenlerin yardımı ile kayıp verilerin tahmini

Maksimum beklenti

- Bu yöntemle algoritma birden fazla veriyi aynı anda tahmin ederek tüm kayıp verilere aynı değeri atar.
- Tüm veri dizisi ele alınabileceği gibi, belli sınıflara göre parçalama da yapılabilir. Örneğin maaş tahmin edilecekse deneyim yada yaşa göre filtreleme yapılabilir.
- Tüm kayıp veriler için keyfi bir değer seçilir
- Her iterasyonda yeni kayıp veri hesabı ve önceki atanan değer ile karşılaştırma yapılır. Aradaki fark başlangıçta belirlenen eşik değerden fazla ise yeni iterasyona geçilir, işlemler tekrarlanır.
- Örneğin bir üretim bandındaki ürünleri inceleyerek örnek olarak aldığımız 10 ürünün 5'inin kötü üretim olduğunu gözlemleyelim. Bu durumda bundan sonraki ürünlerin kötü üretilme oranının %50 olabileceğini söyleyebiliriz.

1. Veri Temizleme

Diğer değişkenlerin yardımı ile kayıp verilerin tahmini

Maksimum beklenti

n: Veri Kümesi
k: Bilinen veri sayısı
n – k: Kayıp veri sayısı
e: Eşik değer (durma kriteri)

1. Tüm kayıp verilere rastgele bir değer ata x^p
2. Kayıp verileri hesapla

$$y' = \frac{\sum_{i=1}^{k} x_i}{n} + \frac{\sum_{j=k+1}^{n} x_j^p}{n}$$

3. if ($(y' - x^p) < e$)
Dur
Else
$x^p = y'$ ve 1. adıma git

1. Veri Temizleme

Diğer değişkenlerin yardımı ile kayıp verilerin tahmini

Jackknife

- Bu yöntem maksimum beklenti yöntemine benzer fakat her bir kayıp veri için ayrı değer tahmini yapılır.
- Kayıp veriler maksimum beklenti yöntemindeki gibi hep birlikte değil teker teker tahmin edilir.

n: Veri kümesi
k: Bilinen veri sayısı
j: Kayıp veri sayısı
e: Eşik değeri

for(i=1 to j)

1. Kayıp veri için rastgele bir değer ata: x^p
2. Yeni kayıp veriyi hesapla

$$y' = \frac{\sum_{i=1}^{k} x_i + x^p}{k+1}$$

3. If $|x^p - y'| \leq e$
   k++
   $x^p = y'$
   Dur
   Else
   $x^p = y'$

Gürültülü Veri

- ***Gürültü***, ölçülen bir değişkende rastgele bir hata veya değişimdir.
  Gürültü aşağıdaki durumlarda oluşabilir:
    - •Veri toplama araçlarındaki hatalar
    - •Veri giriş problemleri
    - •Veri iletim problemleri
    - •Teknoloji sınırları
    - •Öznitelik isimlendirme tutarsızlıkları
- Veri temizlemesini gerektiren diğer durumlar
    - •Tekrarlı kayıtlar
    - •Eksik veriler
    - •Tutarsız veriler

Gürültünün Temizlenmesi

- Paketleme/kutulama (Binning)
  - •Veri sıralanır ve eşit frekanslarda paketlere bölünür.
  - •Eksik veriler farklı yöntemlerle doldurulur:
    Ortalama, medyan, sınır değerler yardımıyla
- Regresyon (Regression)
  - •Regresyon fonksiyonlarına tabi tutularak eksik verilerin girilmesi
- Bölütleme (Kümeleme , Clustering)
  - •Aykırı verilerin bulunması ve temizlenmesi
- Bilgisayar ve insan bilgisinin ortaklaşa kullanılması
  - •Şüpheli değerlerin tespit edilmesi ve manuel olarak kontrol edilmesi

Gürültünün Temizlenmesi

Pakletleme

➢ Verileri sıraya diz ve sıralı verileri frekans yada mesafeye göre eşit olarak böl

Sıralı ücret verisi (dolar): 4, 8, 9, 15, 21, 21, 24, 25, 26, 28, 29, 34

* Veriyi eşit frekanslı paketlere böl :
  - **Paket 1:** 4, 8, 9, 15 (ortalama 9, sınır değerler 4 ve 15)
  - **Paket 2:** 21, 21, 24, 25 (ortalama 22.75, sınır değerler 21 ve 25)
  - **Paket 3:** 26, 28, 29, 34 (ortalama 29.25, sınır değerler 26 ve 34)

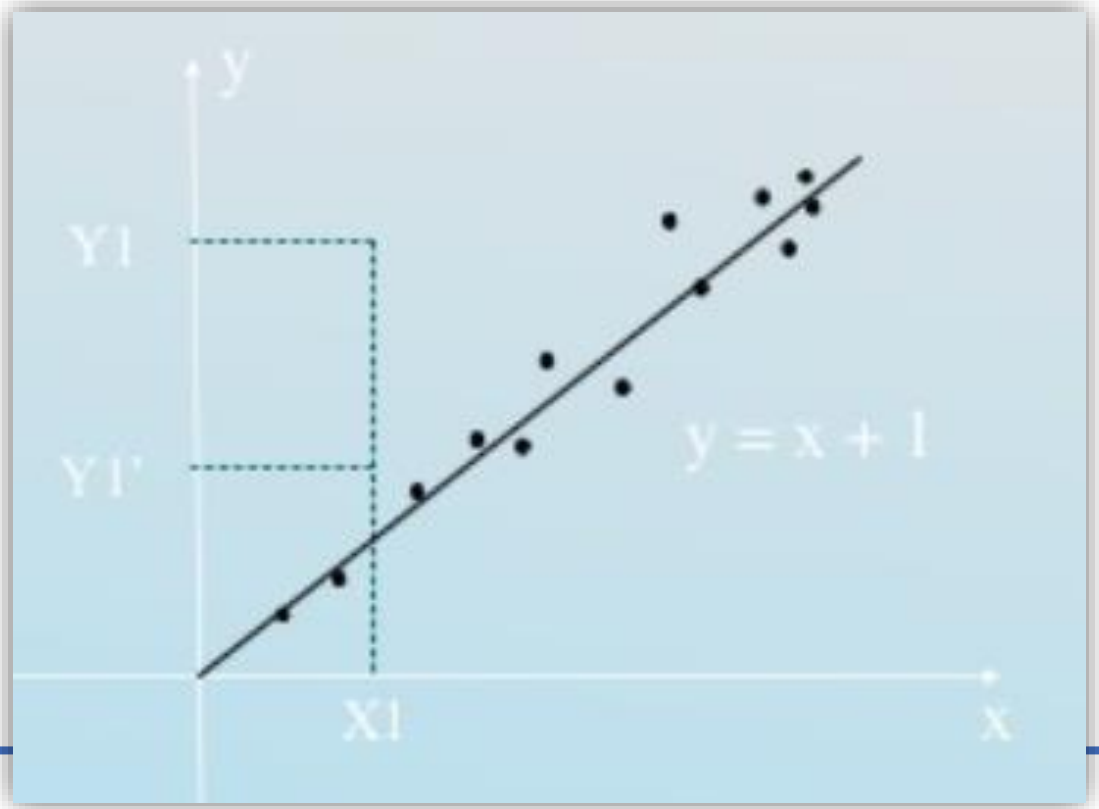
* Ortalama ile gürültü temizleme:
  - Paket 1: 9, 9, 9, 9
  - Paket 2: 23, 23, 23, 23
  - Paket 3: 29, 29, 29, 29

* Sınır değerler ile gürültü temizleme:
  - Paket 1: 4, 4, 4, 15
  - Paket 2: 21, 21, 25, 25
  - Paket 3: 26, 26, 26, 34

Gürültünün Temizlenmesi

Regresyon

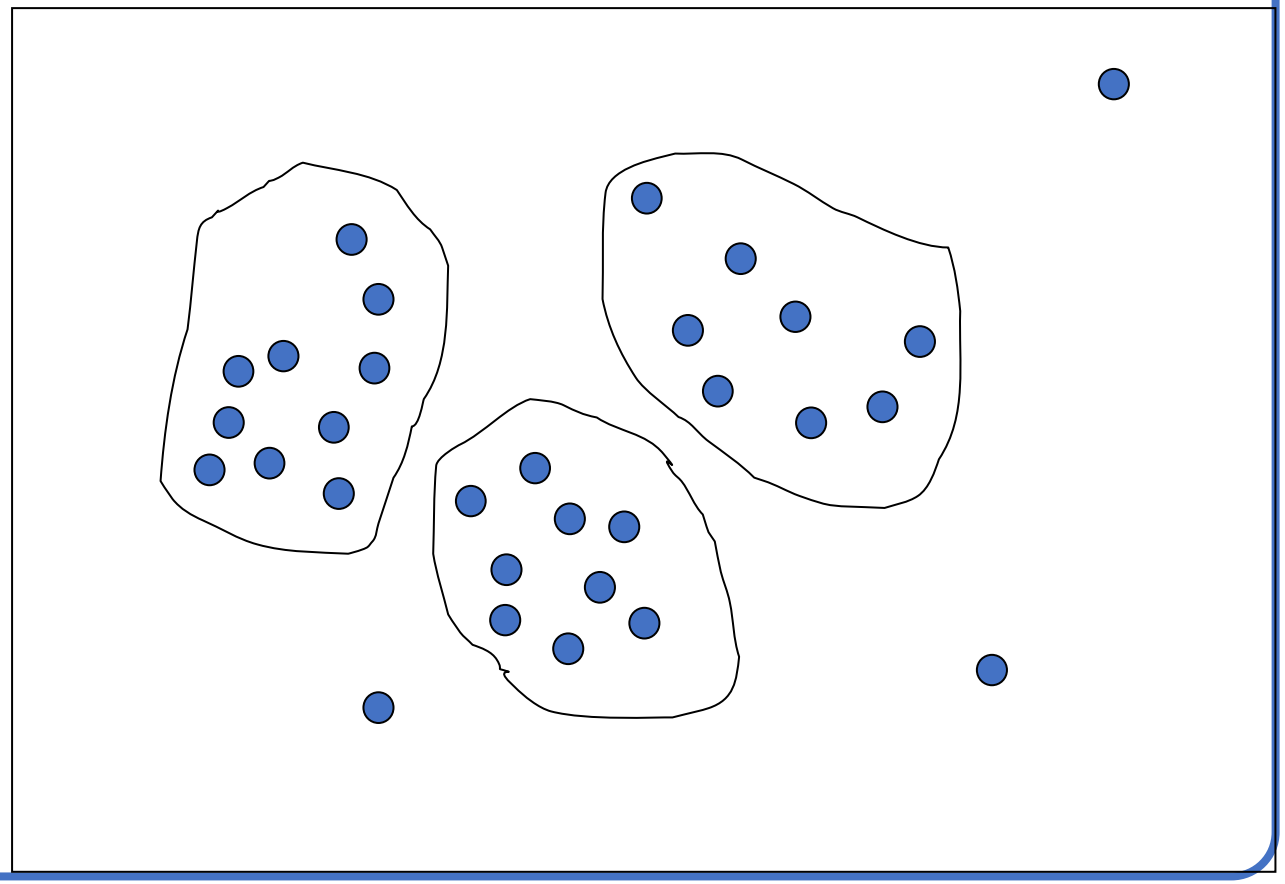
- Tespit edilen uç değerlerin yerine regresyon eğrilerine uygun yeni değerler girilebilir.



Gürültünün Temizlenmesi

Kümeleme

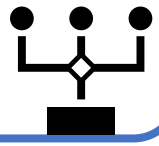
- Kümeleme ile uç (aykırı) değerler belirlenerek bunlara yeni değerler atanabilir.



2.

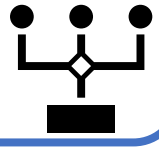
Veri Bütünleştirme

- Veri bütünleştirme, farklı veri kaynaklarında bulunan verilerin tek bir veri kaynağında toplanması işlemidir. Farklı kaynaklardaki verilerin birleştirilmesi bir takım sorunları da beraberinde getirmektedir.
- Örneğin, bir veri tabanında girişler “öğrenci-ID” şeklinde yapılmışken, bir diğerinde “ogrenci-numarası” şeklinde olabilir.
- Bu tip hatalara şema birleştirme hataları denir. Bu tip hatalardan kaçınmak için meta veriler kullanılır.
- Nitelik değerlerinde tutarsızlıklar olabilir, aynı veri farklı kaynaklarda farklı metrikler yada gösterim şekli ile kayıt altına alınmış olabilir; (inch/cm/metre), (18 Ekim 2019 / 18.10.2019 / 10/18/2019).



2. Veri Bütünleştirme

- Veri bütünleştirme sırasında tekrarlı ve artık (gereksiz) verilerin oluşması da bir başka problemdir.
- Aynı nitelik farklı kaynaklarda farklı isimlerle yer alabilir yada bir nitelik başka bir nitelik kullanılarak hesaplanabilir (yıllık gelir).
- Bazı gereksiz veriler korelasyon analizi ile tespit edilebilir. Kategorik veriler için ki-kare testi, nümerik veriler için korelasyon katsayısı yada kovaryans kullanılabilir.
- Birbiri ile yüksek ilişkili öznitelikler belirlenerek tek bir değişken altında birleştirilebilir yada özniteliklerden bir tanesi seçilebilir.



2. Veri Bütünleştirme

➢ Korelasyon

Herhangi iki olay arasında pozitif ya da negatif bir ilişki söz konusudur. Bu ilişki tam ilişkinin bir yüzdesi olarak belirtilir. İki ya da daha fazla değişkenin arasındaki ilişkinin yönünün ve derecesinin araştırılması korelasyon analizinin konusudur. Korelasyon katsayısı da bu ilişkinin derecesinin tayininde kullanılan bir ölçüdür. Aşağıdaki formül yardımıyla korelasyon katsayısı hesaplanabilir:

$$r = \frac{\Sigma[(X - \bar{X})(Y - \bar{Y})]}{\sqrt{\Sigma(X - \bar{X})^2 \Sigma(Y - \bar{Y})^2}}$$

Burada $X - \bar{X} = x \; ve \; Y - \bar{Y} = y$ alınırsa

$$r = \frac{\Sigma xy}{\sqrt{\Sigma x^2 \Sigma y^2}}$$ ya da $$r = \frac{\Sigma xy}{n\sigma_x \sigma_y.}$$ elde edilir.

Korelasyon katsayısının değeri -1 ile +1 arasındadır. +1 olduğunda değişkenler arasında ikişki tam ve pozitifdir denir. -1 için de tersi söz konusudur. Korelasyon değerinin +1'e yakın olması ilişkinin kuvvetli olduğuna, 0'a yakın olması ise ilişkinin zayıf olduğuna işaret eder.

2. Veri Bütünleştirme

➢ **Korelasyon örneği**

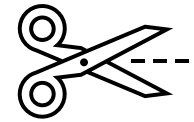
1	Veri Tabanı 1	Veri Tabanı 2	Veri Tabanı 3
2	Aylık Gelir	Aylık Harcama	Yıllık Toplam Kazanç
3	2269	1880	27479
4	3554	2916	44369
5	2554	2097	31325
6	3483	2815	43383
7	2573	2134	30388
8	2817	2338	32400
9	2684	2219	31099
10	3030	2509	34962
11	2520	2088	29899
12	2636	2160	31717
13	3141	2578	36409
14	2352	1950	27456
15	2507	2067	29187
16	3084	2506	37667
17	3341	2725	38794

	Aylık Gelir	*Aylık Harcama*	*Yıllık Toplam Kazanç*
Aylık Gelir	1,00		
Aylık Harcama	0,99	1,00	
Yıllık Toplam Kazanç	0,98	0,98	1,00

Üç farklı tablodan gelen üç farklı değişken için korelasyon matrisi hesaplanmıştır. Değişkenler için 1'e yakın korelasyon katsayısı elde edildiği için bu değişkenlerden biri tercih edilerek veri bütünleştirme işlemi yapılabilir. Üç sütunu birden kullanmak gereksizdir.

3. Veri İndirgeme

- Büyük veri setlerinde karmaşık veri analizi ve veri madenciliği çalışmaları yapmak uzun zaman alabilir.
- Veri indirgeme teknikleri, hacim olarak daha küçük olan, ancak orijinal verilerin bütünlüğünü yakından koruyan veri kümesinin azaltılmış bir temsilini elde etmek için uygulanabilir.
- Yani, azaltılmış veri setindeki madencilik daha verimli olmalı, ancak aynı (veya hemen hemen aynı) analitik sonuçları vermelidir.



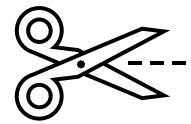
3. Veri İndirgeme

Veri İndirgeme Stratejileri

- **Boyut İndirgeme (Dimensionality reduction): Önemsiz öznitelikleri kaldırma**
  - Dalgacık Dönüşümü (Discrete Wavelet transforms - DWT)
  - Temel Bileşen Analizi (Principal Components Analysis - PCA)
  - Öznitelik altkümesi seçimi, öznitelik oluşturma (Feature subset selection, feature creation)

- **Sayısal İndirgeme (Numerosity reduction): veri azaltma**
  - Regresyon Modelleri
  - Histogramlar, kümeleme, örnekleme
  - Veri küpü

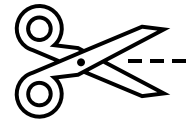
- **Veri sıkıştırma (Data compression)**



3. Veri İndirgeme > Veri İndirgeme Stratejileri

Dalgacık Dönüşümü (Discrete Wavelet transforms - DWT)

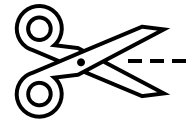
- Veriyi, orijinal veriyle aynı uzunlukta farklı bir vektöre çevirir. Bu çevrilen yeni vektör de kısaltılabilir – kesilebilir olduğundan veri, işimize yarayacak kadar azaltılmış olur.
- Ayrıca smoothing (düzeltme) işlemi yapmadan verinin gereksiz bilgilerden temizlenmesi görevini de yerine getirir.
- DWT küp yapıdaki çok boyutlu verilere uygulanabilir. DWT nin ne kadar kompleks olduğu küpteki hücre sayısına bağlıdır.
- DWT ayrık ve çarpık verilerle daha iyi sonuçlar verir ve gerçek hayatta da sıklıkla kullanılmaktadır.
- (parmak izi resimlerinin sıkıştırılması, zaman serisi veri analizi, veri temizleme)



3. Veri İndirgeme > Veri İndirgeme Stratejileri

Temel Bileşen Analizi (PCA - principal component analysis)

- Verinin yapısını bozmadan veriden belirli sayıdaki kaydı belli boyuta göre alır ve analiz eder. Hem kayıt sayısında azalma olur, hem de kolon – boyut indirgenir.
- PCA da veri azaltmak için orijinal veriden n tane kayıt k tane boyut ortagonal olarak bulunur. İşleyişinde veri normalize edilir. PCA orijinal veriyle aynı yapında n tane ortagonal vektör hesaplar. Bu vektörlere temel bileşen denir. Vektörler sıralanır. Sıralanan bu vektörlere göre en zayıf bileşen çıkartılarak data azaltılır.
- PCA nın uygulanması ucuzdur. Yüksek performans beklenmez.
- Sıralı – sırasız kolonlara, veri özelliklerine uygulanabilir.
- Boyut sayısı arttığında DWT yi kullanmak daha çok tercih edilebilir.

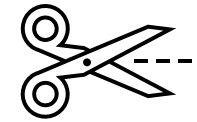


3. Veri İndirgeme > Veri İndirgeme Stratejileri

Öznitelik altkümesi seçimi, öznitelik oluşturma

- Veri setindeki tüm öznitelikler veri madenciliği süreci için aynı öneme sahip olmayabilir. Veriyi ifade edebilecek en küçük öznitelik altkümesinin seçilmesi gereklidir.

- Örneğin alışveriş eğiliminin belirlenmesinde telefon numarası gereksiz bir özniteliktir, alışveriş alışkanlıklarında etkin bir rolü yoktur.



3. Veri İndirgeme > Veri İndirgeme Stratejileri

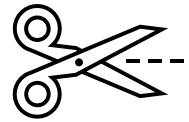
Öznitelik altkümesi seçimi, öznitelik oluşturma

1.İleriye doğru kademeli seçim: Prosedür, azaltılmış set olarak boş bir öznitelik seti ile başlar. Orijinal özelliklerin en iyileri belirlenir ve azaltılmış sete eklenir. Sonraki her bir yineleme veya adımda, kalan orijinal özniteliklerin en iyisi sete eklenir.

2. Kademeli olarak geriye doğru eleme: Prosedür tüm özniteliklerle başlar. Her adımda, sette kalan en kötü öznitelik kaldırır.

3. İleri seçim ve geriye doğru elemenin kombinasyonu: Her aşamada en iyi öznitelik seçilir ve kalan öznitelikler arasında en kötü kaldırılarak işlemler devam eder.

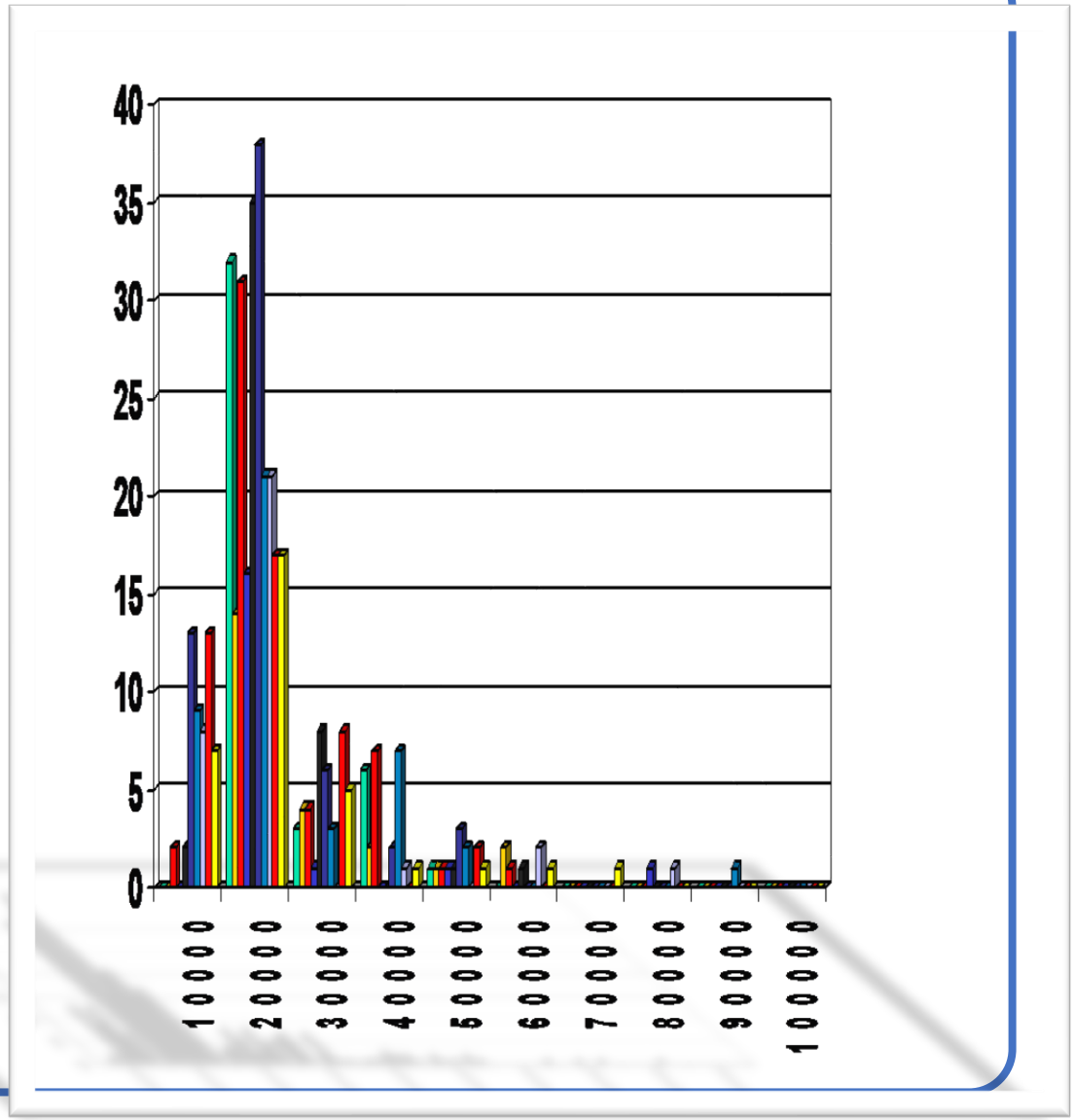
4. Karar ağacı: Verilerden bir ağaç oluşturulur. Ağaçta görünmeyen tüm özniteliklerin alakasız olduğu kabul edilir. Ağaçta görünen öznitelikler kümesi, özniteliklerin azaltılmış alt kümesini oluşturur.



3. Veri İndirgeme > Veri İndirgeme Stratejileri

Histogram

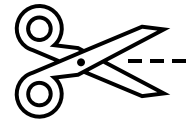
- Histogramlar veri dağılımlarını gösterdikleri için tercih edilen bir indirgeme aracıdır.
- Birbirine benzer öznitelikleri elimize etme (boyut indirgeme)
- Çok boyutlu histogramlar, nitelikler arasındaki bağımlılıkları yakalayabilir.



3. Veri İndirgeme > Veri İndirgeme Stratejileri

Örnekleme

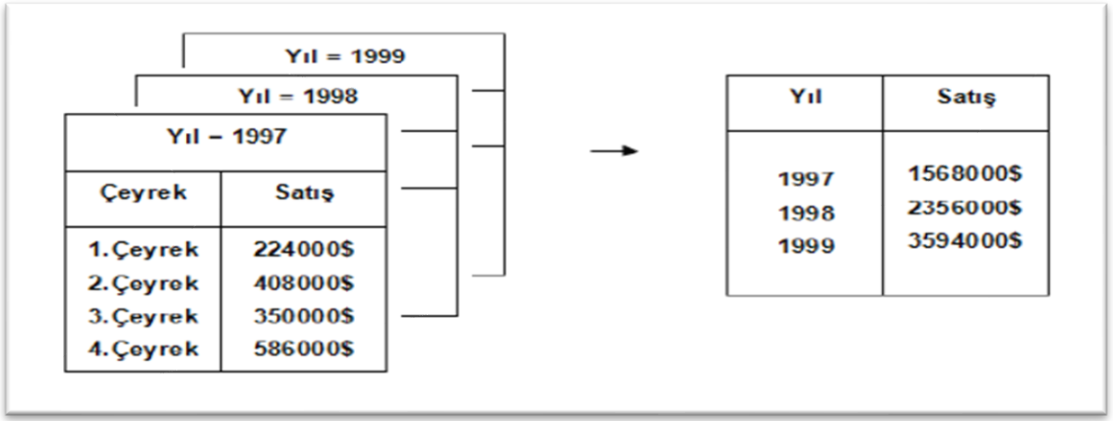
- Veri setinden bütünü en iyi temsil edecek şekilde örneklerin çekilmesi örnekleme olarak adlandırılır.
- Çok büyük bir veri seti ile çalışılıyorsa bu yöntem tercih edilebilir.
- Basit rastgele örnekleme: Veri setindeki her bir örneğin örnekleme girme şansı eşittir.
- Tabakalı örnekleme: Her tabakadan seçilecek kişi sayısı tabaka büyüklüğü ile orantılı olacak şekilde örneklem seçme işlemi yapılır. Veri setinde %40 erkek, %60 kadın dağılımı varsa örneklemde de aynı oranın olması istenir.



3. Veri İndirgeme > Veri İndirgeme Stratejileri

Veri Küpü

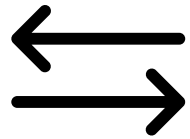
- Bir ürünün yıllık satış rakamları yeterli ve anlamlı ise bu bilgiyi 4 çeyrek dönemlik periyotlarda tutmaya gerek yoktur.



- Veri küpleri çok değişkenli birleştirilmiş bilginin saklandığı küplerdir. Örneğin bir firmanın satış tutarları yıllar, satışı yapılan ürünler ve firmanın farklı satış yerleri için aynı küp üzerinde gösterilebilir. Veri küpleri özet bilgiye herhangi bir hesaplama yapmadan hızlı bir biçimde erişilmesini sağlarlar.

4. Veri Dönüşümü

- Veriyi kesikli hale getirme
- Nominal veriler için kavram hiyerarşisi
- Normalizasyon
  - Min-Max
  - Z Skor
  - Ondalık Ölçekleme



4. Veri Dönüşümü

Veriyi kesikli hale getirme

- Bazı veri madenciliği algoritmaları kategorik veri ile çalışır. Bu durumda sayısal verilerin nominal veriye dönüştürülmesi (kesikli hale getirilmesi) gerekmektedir.

Yıllık Gelir
3169
4031
4516
2909
2907
4312
2360
4120
3521
4680
4306
2427
2459
2432

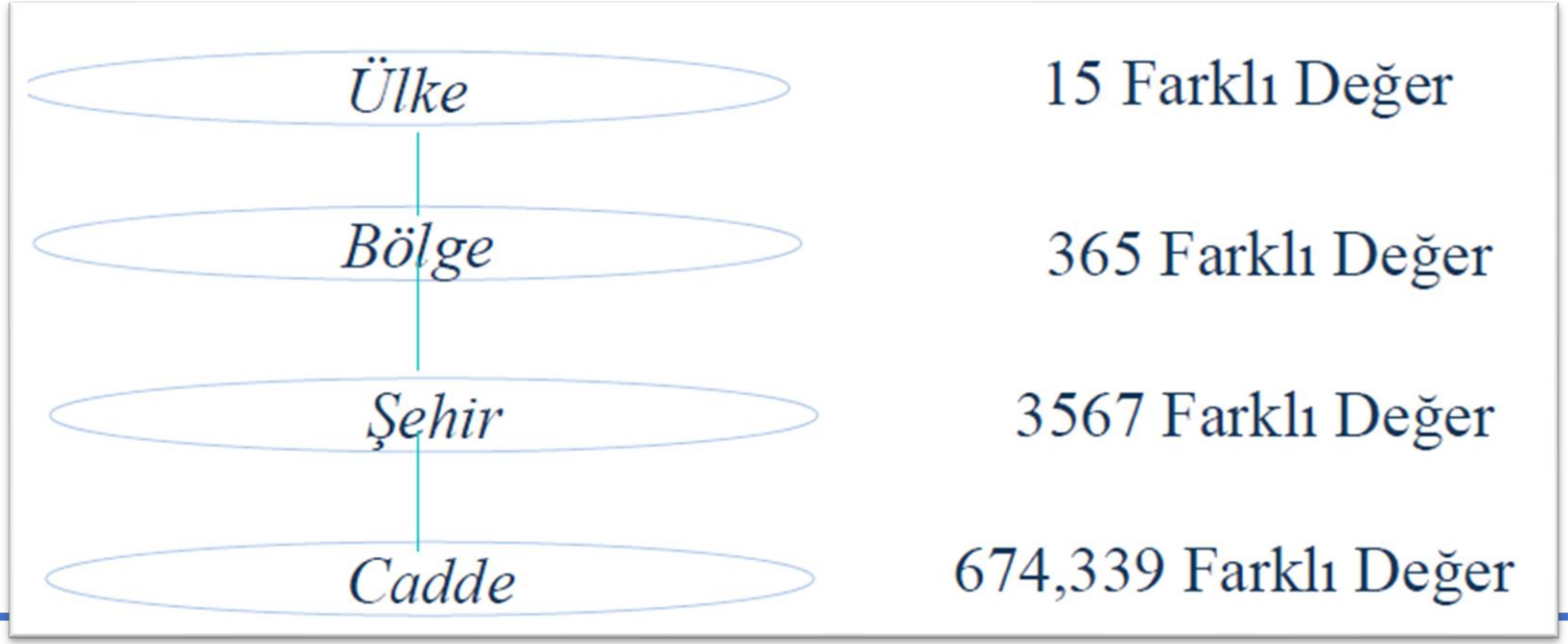
Gelir	
2000-2700	Düşük
2701-3600	Orta
3601-4200	Yüksek
4200 den fazla	Çok yüksek

Yıllık Gelir
Orta
Yüksek
Çok Yüksek
Orta
Orta
Çok Yüksek
Düşük
Yüksek
Orta
Çok Yüksek
Çok Yüksek
Düşük
Düşük
Düşük

4. Veri Dönüşümü

Nominal veriler için kavram hiyerarşisi

- Farklı hiyerarşik sevilerde düzenlenmiş verilerde alt seviyedeki değişken (en fazla sayıda farklı değer bulunan) yerine üst seviyedekinin (en az sayıda farklı değer bulunan) tercih edilmesi.
- Cadde yerine ilçe yada şehir bilgisinin kullanılması



4. Veri Dönüşümü

Min-Max Normalizasyonu

Veri Dönüştürme :

Min-Max Normalleştirilmesi

Örnek :

$$x^* = \frac{x - x_{min}}{x_{max} - x_{min}}$$

x^*	Dönüştürülmüş değer
x	Gözlem değerleri
x_{max}	En büyük gözlem değeri
x_{min}	En küçük gözlem değeri

$x_{max} = 84$
$x_{min} = 40$

İlk eleman için

$$x^* = \frac{x - x_{min}}{x_{max} - x_{min}} = \frac{40 - 40}{84 - 40} = 0$$

x	x*
40	0
52	0,2727
64	0,5455
72	0,7273
84	1,0000

Benzer biçimde diğer gözlemler için aynı hesaplamalar yapılır.

Min-Max normalleştirme dönüşümü için elde edilen sonuçlar.

4. Veri Dönüşümü

Z-score normalizyonu

Veri Dönüştürme :

Z-score Normalleştirilmesi

Örnek :

$$x^* = \frac{x - \bar{x}}{\sigma_x}$$

- x^* Dönüştürülmüş değer
- x Gözlem değerleri
- $\bar{x}$ Verilerin arit.ortalamasını
- σ_x Değerlerin Standart sapması

$\bar{X}$ airtmetik ortalamanın bulunması için ;

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = 62,4$$

Z-score standartlaştırma işlemi için X serisinin standart hatasının bulunması gerekmektedir.

x
40
52
64
72
84

Verilerine göre standart hatanın bulunması gerekmektedir.

$$\sigma_x = \sqrt{\frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1}} = 17,11$$

4. Veri Dönüşümü

Z-score normalizyonu

Veri Dönüştürme :

Z-score Normalleştirilmesi

Örnek :

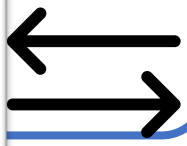
$$x^* = \frac{x - \bar{x}}{\sigma_x} = \frac{40 - 62,4}{17,11} = -1,3091$$

İlk gözlem değeri için hesaplanan bu işlemi diğer veriler içinde

İşlemi tekrarlarsak;

Xi	X*
40	-1,3091
52	-0,6078
64	0,0935
72	0,5610
84	1,2623

Z-score dönüşümü sonucu için elde edilen değerlerdir.



4. Veri Dönüşümü

Ondalık ölçekleme

- Ondalık ölçekleme ile normalleştirmede, ele alınan değişkenin değerlerinin ondalık kısmı hareket ettirilerek (değişken 10 un katlarına bölünerek) normalleştirme gerçekleştirilir.

- Veri setindeki en büyük değer 0-1 aralığına çekilir.

- Ondalık ölçeklemenin formülü: $v'(i) = V(i)/10^k$

- Örneğin 900 maksimum değer ise, n=3 olacağından 900 sayısı 0,9 olarak normalleştirilir.

Algoritma Seçimi

Veri madenciliği yöntemlerini uygulayabilmek için veriler üzerinde yapılan işlemlerin gerekli olanları uygulanır.
Veri hazır hale getirildikten sonra konuyla ilgili veri madenciliği algoritmaları uygulanır.
Söz konusu kullanılan yöntemler ve algoritmalar :

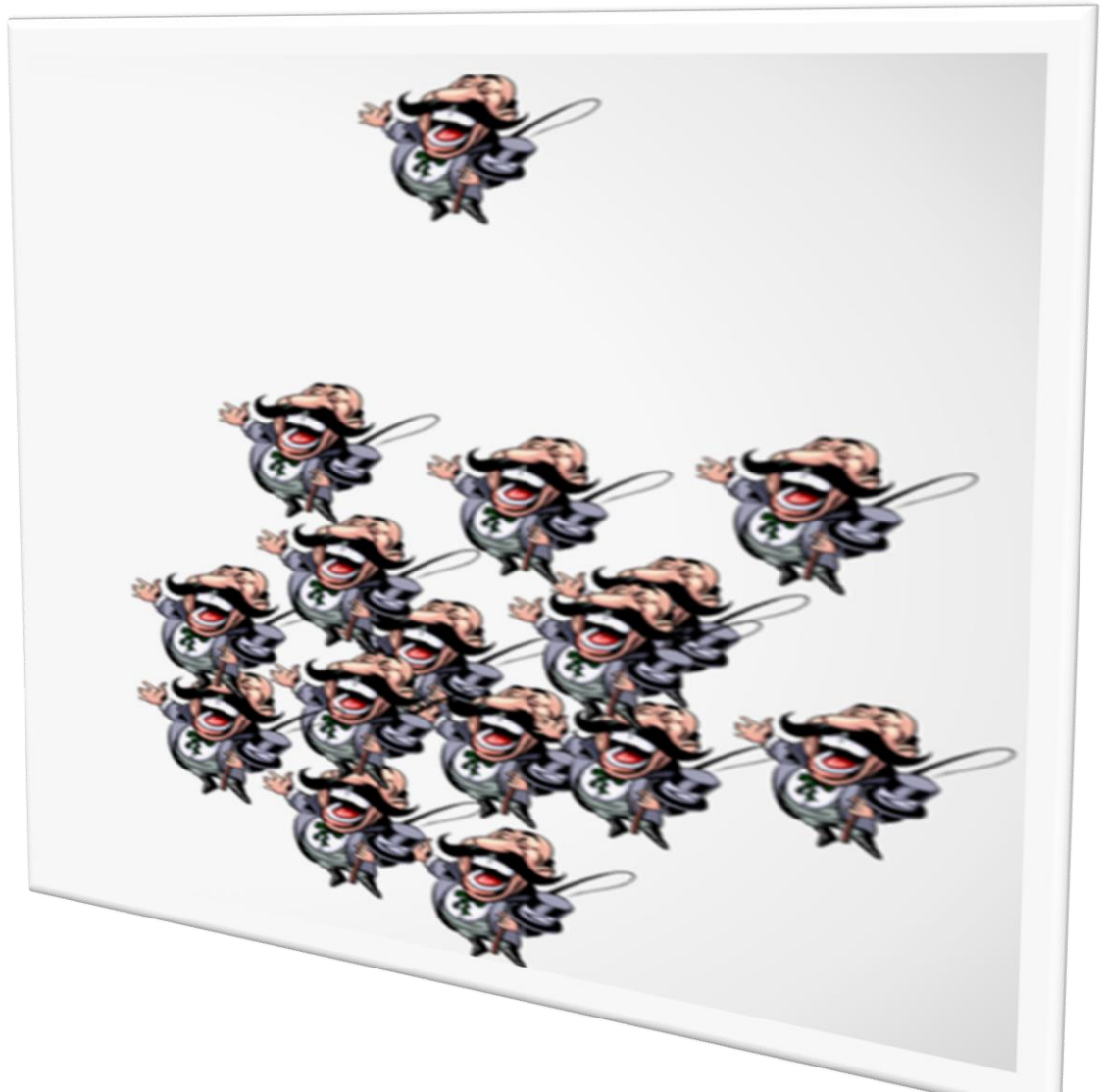
Kullanılan Yöntemler :

- ❖Sınıflandırma
- ❖Kümeleme
- ❖Görselleştirme
- ❖İlişki kurma
- ❖Tahmin modelleri

Kullanılan Algoritmalar :

- ❖Sinir Ağları (neural networks)
- ❖Karar Ağaçları (decision trees)
- ❖Genetik Algoritmalar
- ❖İstatistiksel Analiz

Verilerin konuşmasına müsaade edin…



Söyleyeceği çok şeyi olabilir…fakat sadece biraz **gürültülü** olabilir.

Referanslar

- Veri madenciliği ders notları, Dr. Burcu Çarklı YAVUZ, SAÜ
- Data Mining: Concepts and Techniques, J. Han, M. Kamber, J. Pei
- Veri madenciliği Kavram ve Algoritmaları, Gökhan Silahtaroğlu, Papatya Yayıncılık
- Veri Madenciliği Ders Notları, Doç. Dr. Nilüfer YURTAY, SAÜ
- https://matthew-brett.github.io/teaching/linear_interpolation.html
- Veri Madenciliği Yöntemleri, Yalçın Özkan, Papatya Yayıncılık
- DATA MINING: Concepts, Models, Methods & Algorithms, Second Edition, John Wiley, 2011, M. Kantardzic