ISE 311

VERİ BİLİMİ

DR. TUĞRUL TAŞCI

Veri Kaynakları

- Müşteri Alış-veriş Kayıtları (Fiziksel ve Sanal Mağazalar)
- Müşteri İşlem Kayıtları (Telekomünikasyon, Bankacılık ve Internet Bankacılığı)
- İşletme İşlem Kayıtları (Diğer işletmelerle yapılan alım-satımlar, banka işlemleri, borsa işlemleri)
- Bilimsel Veriler (uzay araştırmaları, ilaç araştırmaları, okyanus ve yer altı araştırmaları, deprem araştırmaları, canlılarla ilgili araştırmalar)
- Güvenlik ve Gözetleme Sistemleri (Şehir merkezleri, AVM'ler , Şehir giriş çıkışları, hava alanları, Otoparklar, Binalar)
- Uydu ve Haberleşme Sistemleri
- Olimpiyat Oyunlar, Ulusal ve uluslararası spor müsabakaları
- Dijital Medya: Dijital resim, müzik ve videolar (Filmler)
- Dijital Kütüphaneler
- Web Siteleri ve Mobil Uygulamalar
- E-Posta & Sosyal Medya: Youtube, Facebook, Twitter, Instagram, WhatsApp
- Tıbbi Kayıtlar ve Kişisel Veriler

Verilerle Ne Tür Uygulamalar Yapılabilir ?

Karakterizasyon: Belli bir sınıfa ait karakteristik özellikler

Ayrıştırma: Belli sınıfları birbirinden ayırmak

Evrilme ve Sapma: Zamanla değişen verilerle ilgili yapılan çalışmalardır.

Sınıflandırma: Verinin önceden belirlenmiş sınıflara bölümlenmesidir.

Kümeleme: Verilerin benzerliklerine göre gruplanması

Tahmin: Bilinmeyen değerlerin tahmin edilmesi

Aykırılık Analizi: Aykırı ya da sıra dışı değerlerin/durumların tespit edilmesi

Birliktelik Analizi: Veriler arasındaki karşılıklı ilişkilerin analizi.

Değişken Tespiti

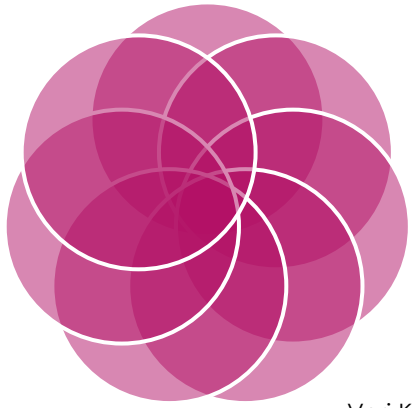
Görselleştirme

- Anlık ve grafiksel olarak keşfetme

Veri Analizi Sürecindeki Zorluklar / Veri Seti Türleri

Problemin doğru tespiti

Etkin Yöntemler
- Büyük miktarda verilerden anlamlı bilgi çıkarmak için etkin yöntemlere olan gereksinim



Yüksek Boyut

Veri Akışı
- Sürekli yeni veri gereksinimi

Karmaşık ve Düzensiz Veri
- Farklı ve uyumsuz veri kaynaklarından gelen verilerin birleştirilmesi

Veri Kalitesi
- Tekrarlı, geçersiz, eksik veri, veri tipinin yanlış seçimi, örneklemenin yanlış yapılması

Gizliliğin Korunması

Kayıtlar
- İlişkisel kayıtlar, Çapraz kayıtlar, Metin, İşlemler

Çoklu Ortam Verileri
- Ses, Resim, Video

Web ve Sosyal Ağlar
- Site İçerikleri, Sosyal Ağ Profil Bilgileri, Paylaşımlar

Sıralanmış Veri Setleri
- Zaman serileri, Sıralı işlem Verileri, Genetik Kod Dizileri

Konum Verileri
- Haritalar ve GPS verileri

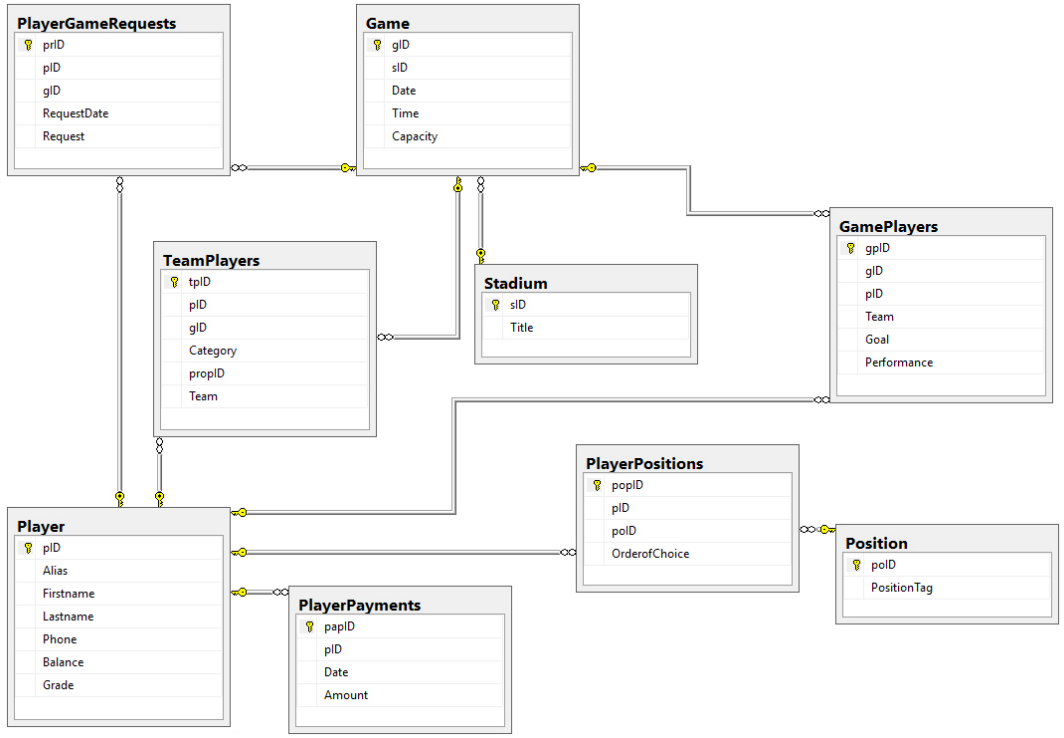
Alana Özel Veriler
- Dil, Kimya, Tıp, Jeoloji vb.

Veri Matrisi

- Matris çok boyutlu uzayda bir nokta olarak düşünülebilir.
- Veri çoğunlukla n×d boyutundaki bir matrisle temsil edilir.
  - n verinin miktarını, d ise boyutsallığını temsil eder.
- Satırlar veri setindeki kayıtları, sütunlar ise verinin kullanılabilecek özelliklerini gösterir.

Kişi	Yaş	Kilo	Boy	Cinsiyet
K01	34	90	165	Erkek
K02	23	65	178	Bayan
K03	45	73	167	Erkek
K04	26	58	159	Bayan
K05	19	75	189	Erkek
K06	21	49	175	Bayan
K07	56	78	163	Erkek
K08	33	57	161	Bayan
K09	29	62	165	Bayan

İlişkisel Veri



İşlem Verisi & Sıralı Veri

İşlem Verisi

Müşteri	İşlem
M01	Ekmek, peynir, süt
M02	Sigara, çakmak
M03	Ekmek, çikolata
M04	Yoğurt, Sucuk, Mısır
M05	Un, nişasta
M06	Yağ, şeker
M07	Çay

Sıralı Veri

(A B) (D) (C E)
(B D) (C) (E)
(C D) (B) (A E)
(D E) (A) (C)

Metin Verisi

Uluslararası hakemli dergilerde yayınlanan makaleler

Tasci T., Oz C. (2014), "A Closer Look to Probabilistic State Estimation – Case: Particle Filtering", Optoelectronics & Advanced Materials – Rapid Communications, Vol. 8(5-6), pp. 521 – 534.

Tasci T., Parlak Z., Kibar A., Tasbasi N. &, Cebeci H.I. (2014), " A Novel Agent-Supported Academic Online Examination System", Educational Technology & Society, Vol.17 (1), pp. 154 – 168.

Uluslararası Diğer Hakemli Dergilerde Yayınlanan Makaleler

Hiziroglu K., Tasci T. & Ozcelik T. O. (2012), "Analysis of Current Occupational Health and Safety Situation and Needs of SMEs in Turkey", Journal of Labor Relations, Vol. 3(2), pp. 66 – 89.

Uluslararası Bildiriler

Yolcu G., Oz C. & Tasci T., "Developing and Establishing a Painting Program Controlled by Hand Motions Using Kinect", 2nd International Symposium On Innovative Technologies In Engineering And Science (ISITES), Karabuk University, June 18-20, 2014, Karabuk, Turkey.

Tasci T., Tasbasi N., Velichkov A., Kloos U. & Tullius G., "A Comparative Evaluation of Two 3D Optical Tracking Systems" ,JVRC 2012 - Joint Virtual Reality Conference of ICAT - EGVE - EuroVR, October 17-19,2012, Madrid, Spain

Ulusal hakemli dergilerde yayınlanan makaleler

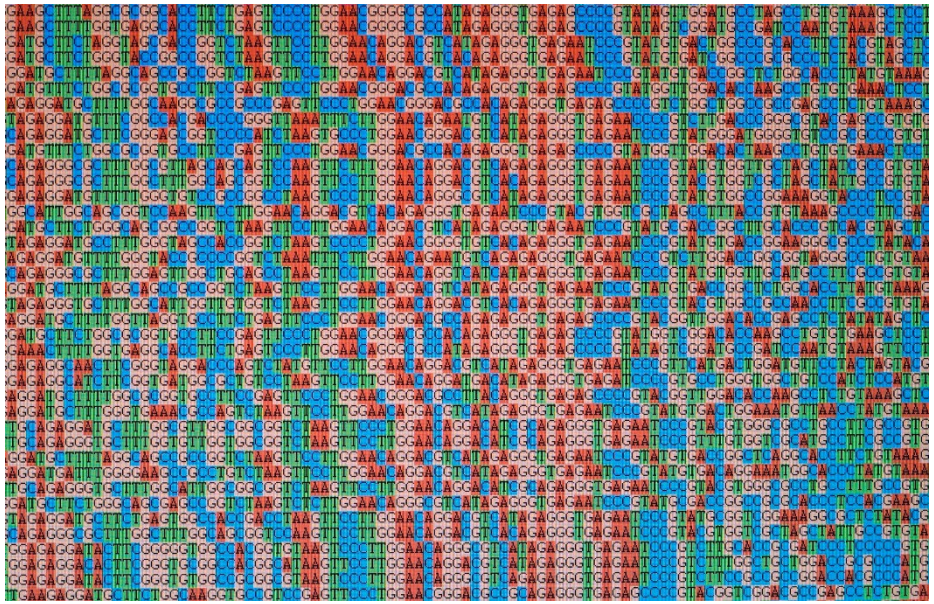
Ulusal bilimsel toplantılarda sunulan ve bildiri kitabında basılan bildiriler

Tasci, T., Goksu A. & Kantoglu B., "E-Dönüşümde Bilgi ve İletişim Teknolojilerinin Kullanımı", Akademik Bilişim Konferansı, February 11-13, 2004, Trabzon, Turkey

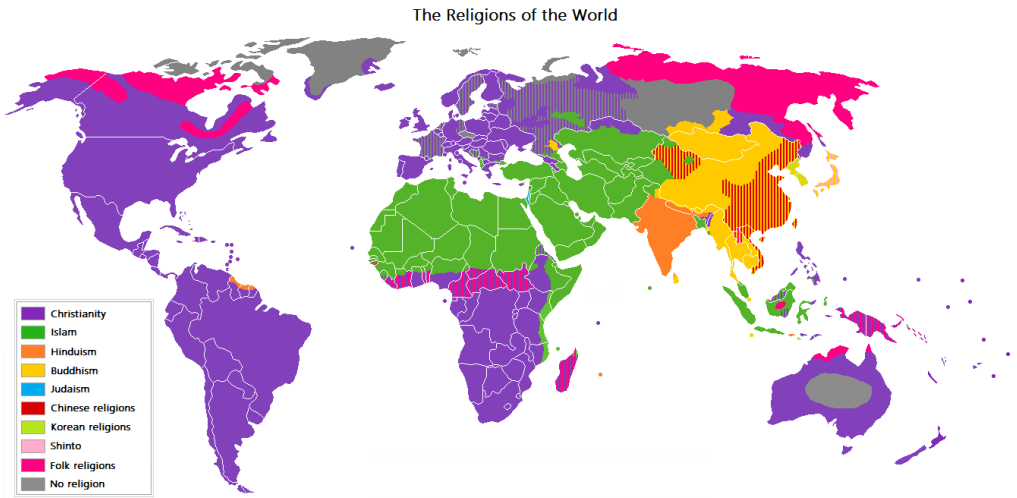
Diğer Yayınlar

Tuğrul TAŞCI "Temel Bilgi Teknolojisi Kullanımı - İşletim Sistemleri" ,Sakarya Üniversitesi, 978-605-4735-03-7, 2012.

Gen Dizisi Verisi



Infografik Verisi

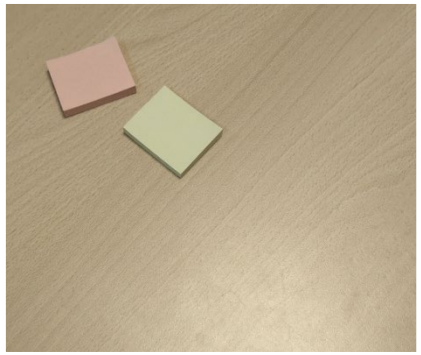


Harita ve Konum Verisi

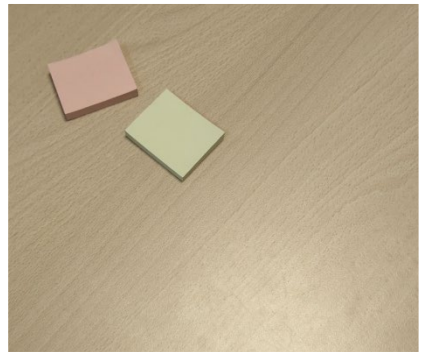
Resim & Video Verisi

Resim Verisi

Video Verisi



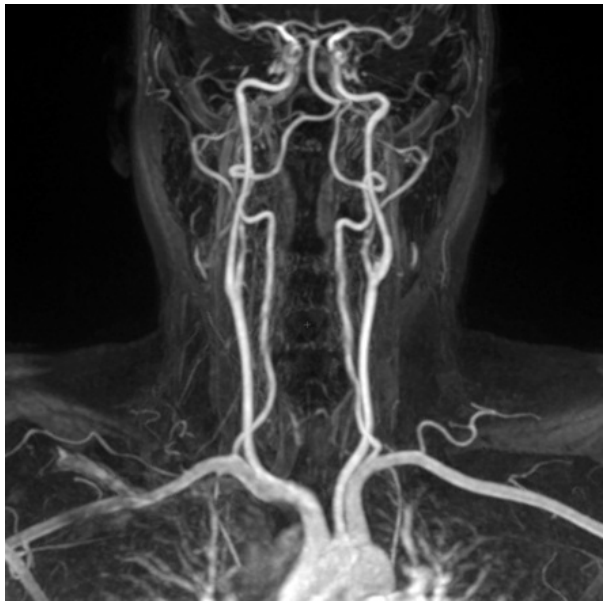
Kare 65



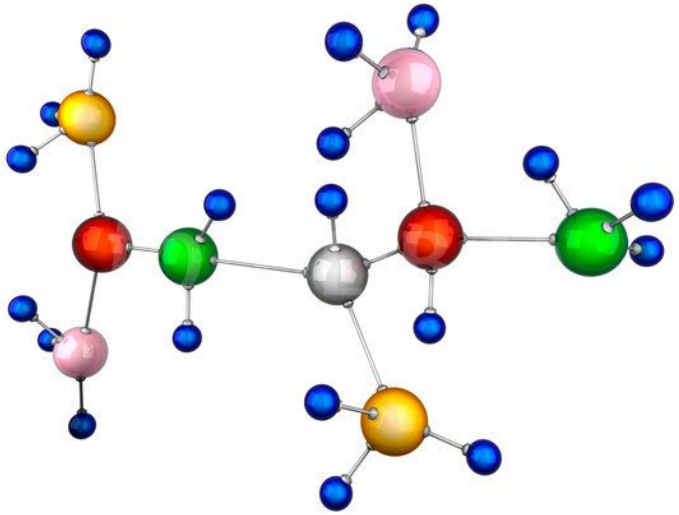
Kare 70

Tıbbi & Kimyasal Veri

Tıbbi Veri (MR)



Kimyasal Veri



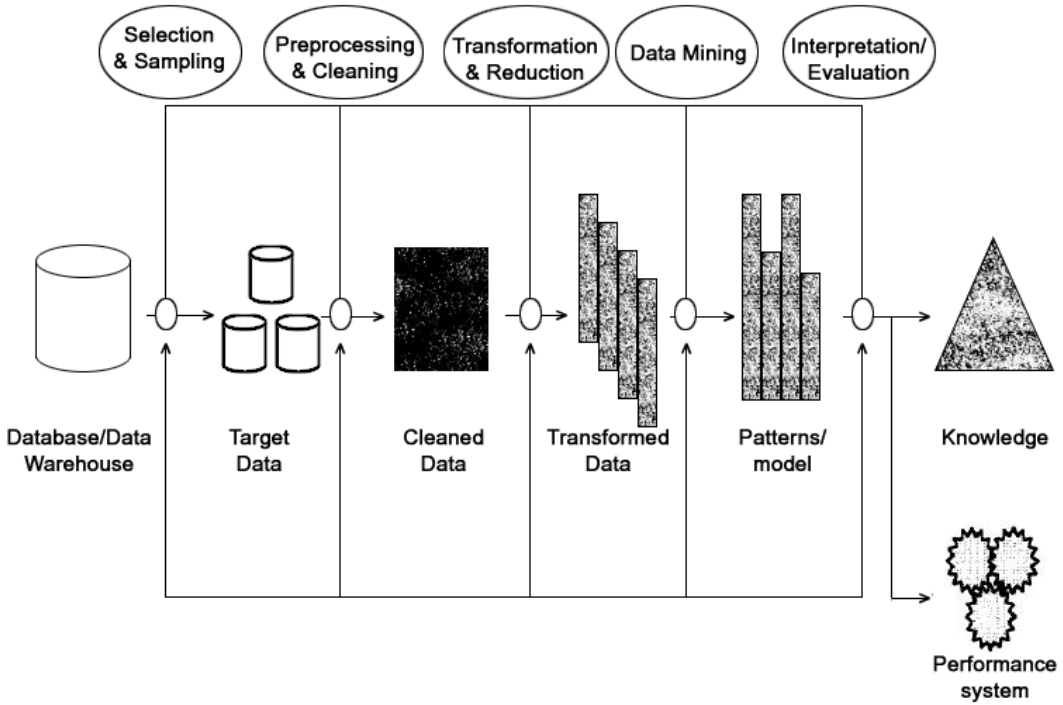
Elde Edilen Kirli Veri & Nedenler

- Eksik veri kayıtlarının nedenleri
  - Veri toplandığı sırada bir nitelik değerinin elde edilememesi, bilinmemesi
  - Veri toplandığı sırada bazı niteliklerin gerekliliğinin görülememesi
  - İnsan, yazılım ya da donanım problemleri
- Hatalı veri kayıtlarının nedenleri
  - Hatalı veri toplama gereçleri
  - İnsan, yazılım ya da donanım problemleri
  - Veri iletimi sırasında problemler
- Tutarsız veri kayıtlarının nedenleri
  - Verinin farklı veri kaynaklarında tutulması
  - İşlevsel bağımlılık kurallarına uyulmaması

Veri Kalitesini Belirleyen Ölçütler

- Doğruluk (Accuracy)
- Eksiksizlik (Completeness)
- Uyumluluk (Consistency)
- Zamanlılık (Timeliness)
- İnanılabilirlik (Believability)
- Katma Değerlilik (Value Added)
- Yorumlanabilirlik (Interpretability)
- Erişebilirlik (Accessibility)
- Özgün (Intrinsic)
- Bağlamsal (Contextual)
- Temsi Edebilirlik (Representational)

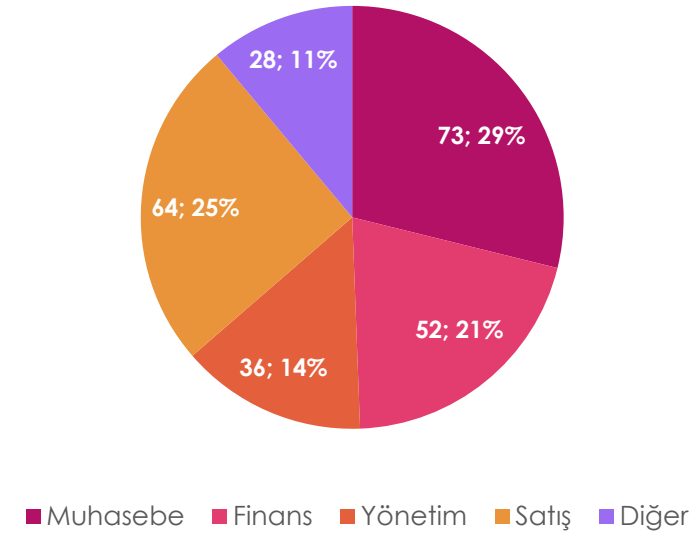
Bilgi Keşfi Süreci ve Veri Analizi

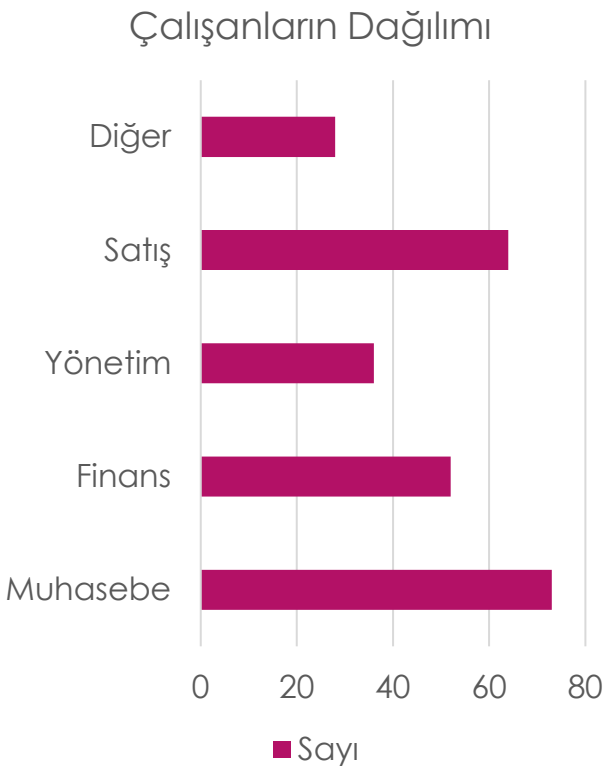


Veriyi Anlama – Görsel Teknikler

- Nominal Veri
  - Frekans Dağılımı
  - Sütun Grafikleri
  - Pasta diyagramı
  - Pareto diyagramı
- Nümerik Veri
  - Çizgi Grafikleri
  - Frekans Dağılımı
  - Histogram ve Ogive
  - Stemplot Diyagramı
  - Serpilme Diyagramı

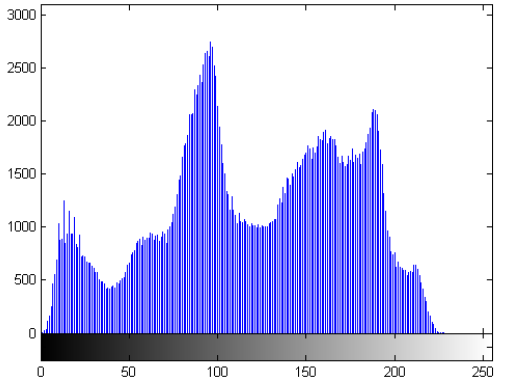
Departman	Sayı	Oran
Muhasebe	73	28.9
Finans	52	20.6
Yönetim	36	14.2
Satış	64	25.3
Diğer	28	11.1
Toplam	253	100



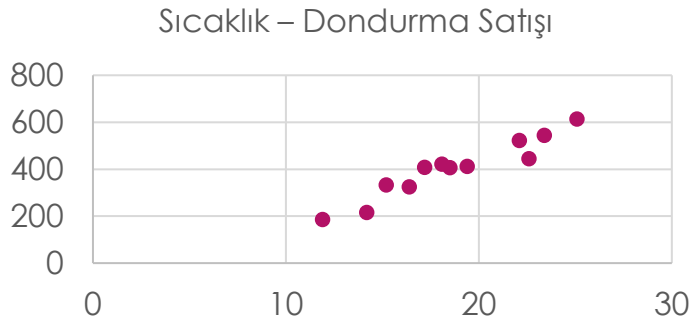


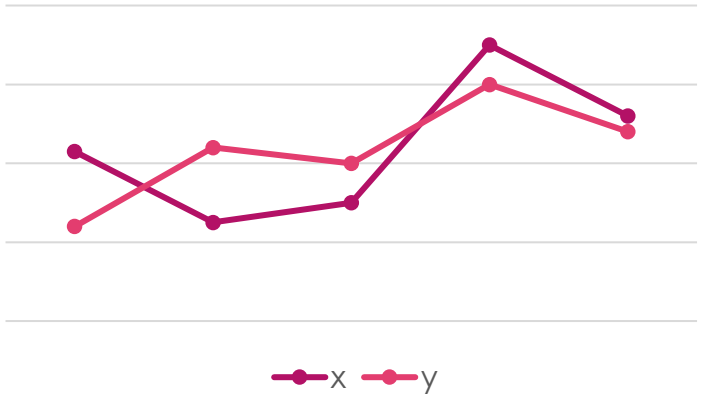
Veriyi Anlama – Görsel Teknikler

- Nominal Veri
  - Frekans Dağılımı
  - Sütun Grafikleri
  - Pasta diyagramı
  - Pareto diyagramı
- Nümerik Veri
  - Çizgi Grafikleri
  - Frekans Dağılımı
  - Histogram ve Ogive
  - Stemplot Diyagramı
  - Serpilme Diyagramı



Stem	Leaf
0	5
1	6, 7
2	3, 6, 8
3	4, 5, 5, 5, 5, 8, 9
4	7, 7, 7, 8
5	4, 5
6	0





Veriyi Anlama – Sayısal Teknikler

Merkezi Eğilim Ölçütleri:

Ortalama:

- Basit şekli ile bütün gözlem değerlerinin toplam gözlem adedine bölünmesi ile hesaplanır. Aykırı değerlerden çok etkilenir.

Medyan:

- Sıralı veri setlerindeki ortanca değerdir. Aykırı değerlerden daha az etkilenir.

Mod:

- Bir veri setinde en çok tekrarlanan değerdir. Ancak veri seti çok büyükse anlamlıdır.

Değişkenlik Ölçütleri:

Değişim Aralığı:

- Veri setinin dağıldığı aralıktır. (Max-Min).

Varyans:

- Gözlenen değer ile beklenen değer arasındaki farktır.

- **Standart Sapma:**

- Varyansın ortalama düzeyine normalleştirilmiş halidir.

Doğrusal İlişki Ölçütleri:

Kovaryans:

- İki değişkenin birlikte değişme derecesini gösterir.

Korelasyon Katsayısı:

- Kovaryans değerinin -1 ile +1 arasında normalleştirilmiş halidir.

Kesikli & Sürekli Değişken

Kesikli (Discrete) Değişken

- Sadece sayılabilir değerler alan değişkenler.
- Çok fazla olası değer alan değişkenler:
  - Bir gündeki şikayet sayısı
  - Hane halkını sahip oldukları telefon sayısı
  - Telefon açılmadan önce çalma sayısı
- İki değer alan değişkenler:
  - Cinsiyet: Kız veya Erkek
  - Sorunlu Parça: Evet veya Hayır

Sürekli (Continuous) Değişken

- Sürekli (sayılamayan) değerler alan değişkenler.
  - Bir parçanın kalınlığı
  - Bir işi tamamlamak için geçen süre
  - Solüsyonun ısısı
  - Ağırlık
- Ölçümlerin doğruluk ve hassasiyetlerine bağlı olarak herhangi bir değer alabilirler.

Kategorik & Sıralı Değişken

Kategorik (Nominal) Değişken

- Sayısal büyüklük ifade etmeyen kategorik veri. Nominal değişkenler sadece niteliksel sınıflandırmalarda kullanılırlar. Bu değişkenlerin ölçümü ve sıralanması mümkün değildir.
  - İnsanların medeni hali, cinsiyeti, mesleği, göz rengi buna örnek olarak gösterilebilir.

Sıralı (Ordinal) Değişken

- Bu değişken ölçülen değerlerin birbirlerine göre büyüklüklerini belirleyen ancak bir değişkenin diğerinden ne kadar büyük ya da küçük olduğunu ifade edemeyen değişkenlerdir.
  - Rütbe, derece, yükseklik (uzun, orta, kısa) gibi sıralı verileri içerir.

Merkezi Eğilim Ölçüleri – Ortalama, Ağırlıklı Ortalama

$$\mu_A = \frac{\sum_{i=1}^{n}(a_i)}{n}$$

$$w\mu_A = \frac{\sum_{i=1}^{n}(a_i \times w_i)}{\sum_{i=1}^{n}(w_i)}$$

- Aritmetik ortalama bir merkezi eğilim ölçüsüdür.
- Bir değişkenin beklenen değeri olarak ya da bir veri dizini temsil eden tek bir orta değer olarak düşünülebilir.
- Aykırı değerlerin olduğu bir veri dizisinde ortalama anlamlı bir bilgi vermeyebilir.

Dizi A (1-20)	Ağırlık A (1-20)	Dizi B (1-20)	Ağırlık A (1-20)
1,3	84	1	47
2	92	1,6	41
2	21	1,6	78
2,8	93	2,2	81
3,1	67	2,4	22
3,6	18	2,8	52
3,6	35	2,8	47
5,6	59	4,4	67
5,6	97	4,4	73
5,6	97	4,4	77
6,9	24	5,4	31
7,1	98	5,6	70
7,9	97	6,2	67
7,9	54	6,2	20
8,4	82	6,6	16
9,9	22	7,8	52
9,9	48	7,8	97
11	93	8,6	37
11,5	82	9	61
11,5	97	9	26

Dizi A (21-40)	Ağırlık A (1-20)	Dizi B (21-40)	Ağırlık A (1-20)
11,5	69	9	77
11,7	13	9,2	29
12,2	87	9,6	53
12,2	94	9,6	72
12,2	71	9,6	90
13,5	78	10,6	97
15,8	77	12,4	57
16,8	45	13,2	18
16,8	69	13,2	19
17,6	25	13,8	29
17,8	74	14	85
18,1	12	14,2	29
18,1	35	14,2	83
18,9	14	14,8	28
19,1	18	15	94
22,2	84	17,4	38
22,4	73	17,6	23
22,4	38	17,6	29
24	96	18,8	64
25,7	13	20,2	50

- Veri dizisindeki elemanların ilişkisini tanımlayan bir ağırlık değeri bulunuyorsa, ağırlıklı aritmetik ortalama almak daha makul olacaktır.
- Veri setinde ağırlık değerleri eksik ise muhtemel değerleriyle doldurulabilir.

Merkezi Eğilim Ölçüleri – Medyan & Mod

Medyan (Ortanca)

Aritmetik ortalamanın kullanılamadığı durumlarda, veri setinde en çok tekrar eden eleman da (mod) kullanılabilir.

Mod veri setinin geri kalan kısmından uzak ise anlamlı bir bilgi vermez.

Kategorik bir veri setinde mod en çok tercih edilen seçeneği belirleyebilir.

Numerik bir veri setinde birden çok mod olabilir. Bu durumda modlardan birisini seçmek gerekebilir.

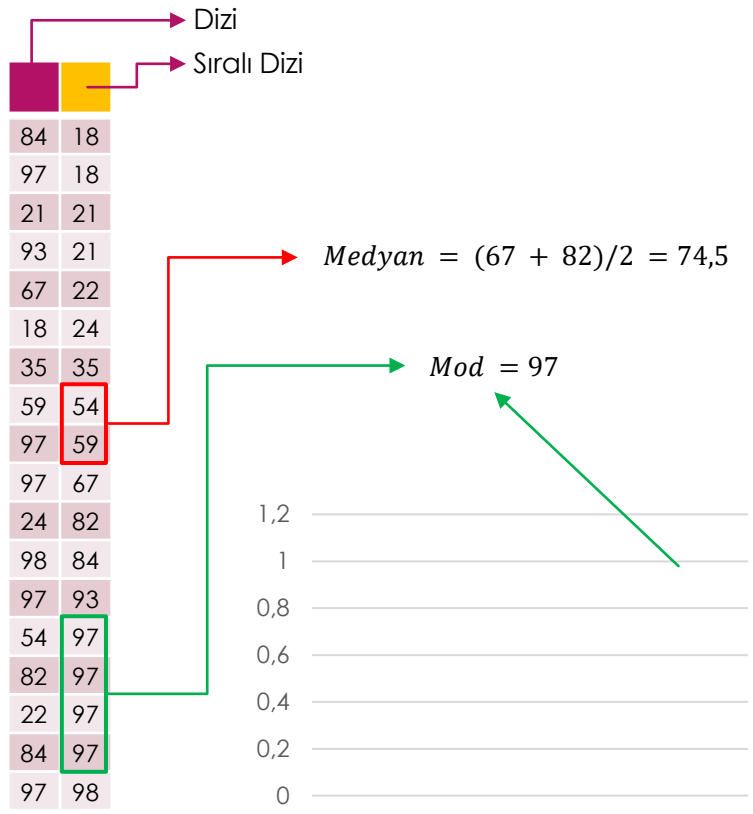
Mod

Veri seti tam anlamıyla normal dağılmış ise ortalama, medyan ve mod aynı değeri alacaktır.

Veri setinde aykırı değerler olduğunda aritmetik ortalama uygun sonuçlar vermeyebilir.

Bu gibi durumlarda veri setindeki medyan (ortanca, sıralamada ortada bir yerde bulunan) değer kullanılabilir.

Veri seti sayısı çift ise medyan ortadaki değerlerden birisi ya da ikisinin ortalaması olarak alınabilir.



Değişkenlik Ölçüleri – Varyans, Standart Sapma

$$\sigma^2{}_A = \frac{\sum_{i=1}^{n}(a_i-\mu)^2}{(n-1)}$$

Değişim Aralığı: Veri setinin dağıldığı aralıktır. (Min-Max).

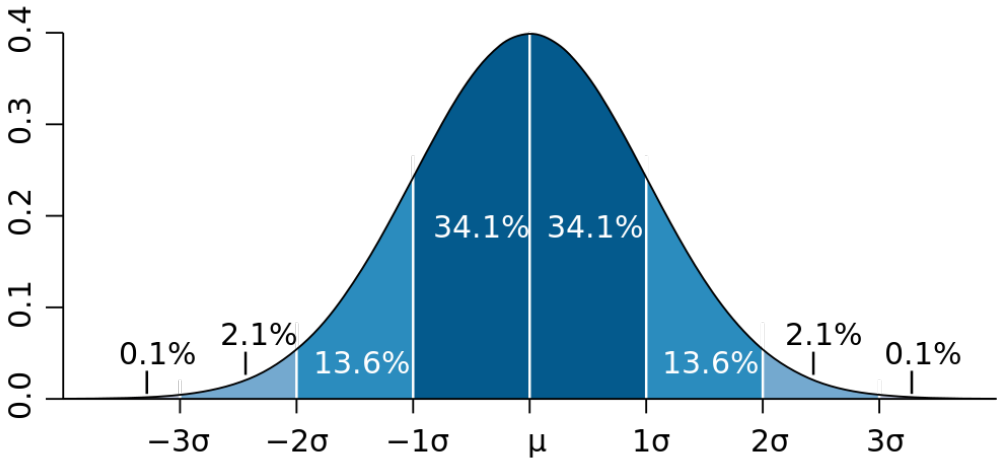
Varyans (Variance): Gözlenen değer ile beklenen değer arasındaki farktır.

Standart Sapma (Standart Deviation): Varyansın ortalama düzeyine normalleştirilmiş halidir. (Varyansın karekökü olarak hesaplanır.)

Merkezi eğilim ölçüleri dağılım hakkında bilgi vermez. Bir veri setinin ortalamasının ne olduğu kadar, verilerin bu ortalama etrafında nasıl değişkenlik gösterdiğinin de bilinmesi önemlidir.

Excel Uygulaması

- Normal Dağılım Grafiği



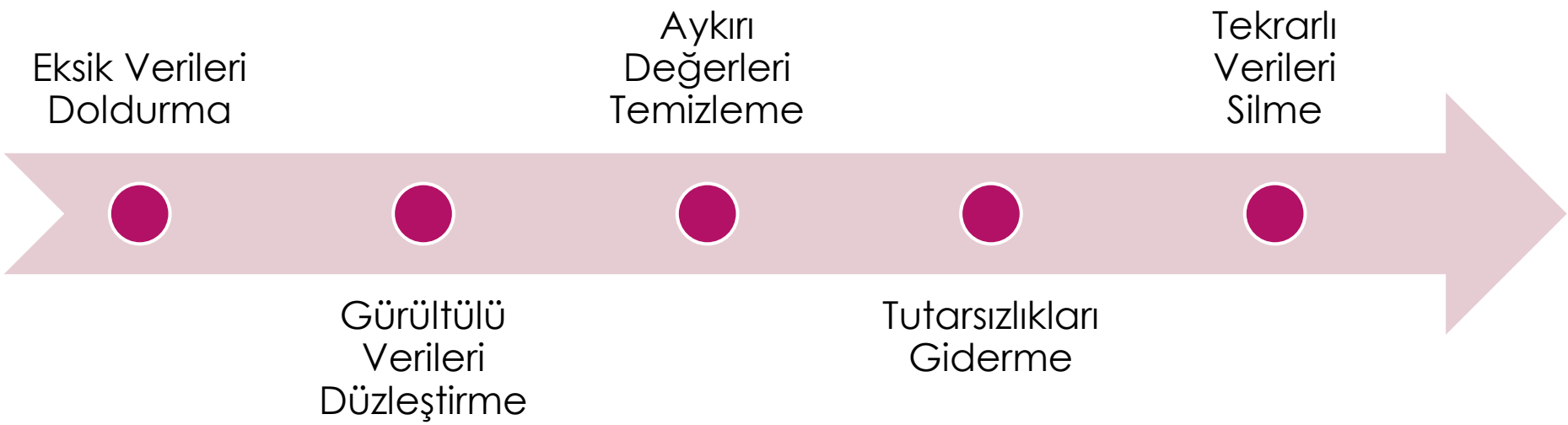
Veri Seçimi & Örnekleme

- Farklı veri kaynaklarından, belli bir problemin çözümü için analiz yapmak amacıyla bir alt veri kümesi belirlenmesi işlemidir.
  - Son 5 yılın verileri
  - Verilerin % 30' u

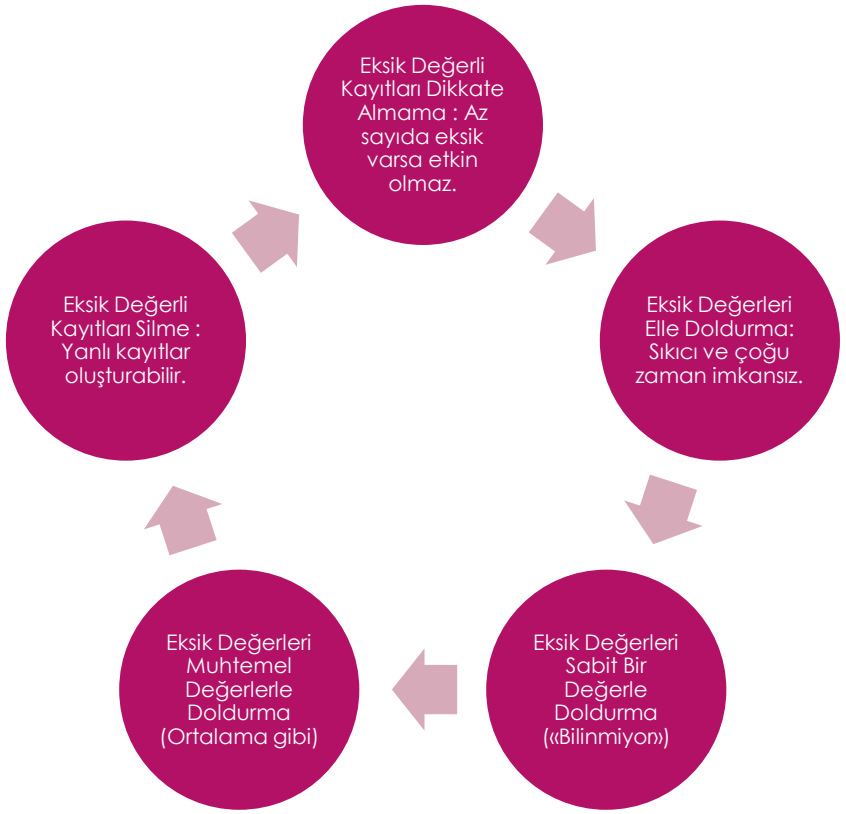
Veri Ön İşleme

- Veri Temizleme (Data Cleaning)
  - Eksik Değerleri Doldurma
  - Gürültülü Veri Düzleştirme (Smoothing)
  - Aykırı Değerleri Tespit Etme ve Ortadan Kaldırma
  - Tutarsızlıkları Giderme
  - Tekrarlı Verileri Silme
- Veri Bütünleştirme (Data Integration)
  - Farklı veri kaynaklarından alınan verileri tek bir ortamda toplama işlemidir.
  - Veri ön-işleme adımlarından birisidir.
- Veri Dönüştürme (Data Transformation)
  - Normalizasyon (Normalization)
  - Birleştirme (Aggregation)
- Veri Azaltma (Data Reduction)
  - Veri Azaltma
  - Boyut Azaltma
- Veriyi Kesikli Hale Getirme (Data Discretization)
  - Eşit aralıklı bölümleme
  - Eşit frekanslı bölümleme
  - Küme tabanlı bölümleme

Veri Temizleme İşlemleri



Eksik Verileri Doldurma



- Eksik Verileri Muhtemel Değerlerle Doldurma Yöntemleri
  - Ortalama, medyan, ya da mod değerleriyle doldurma
  - Eksik değerli kayda en çok benzeyen bir kaydın ilgili değerleriyle doldurma
  - Eksik olmayan değerli kayıtların olasılık dağılımına bağlı bir değerler ile doldurma
  - Regresyon uygulayarak doldurma
  - Tahmin edici başka bir yöntemle eksik değeri doldur

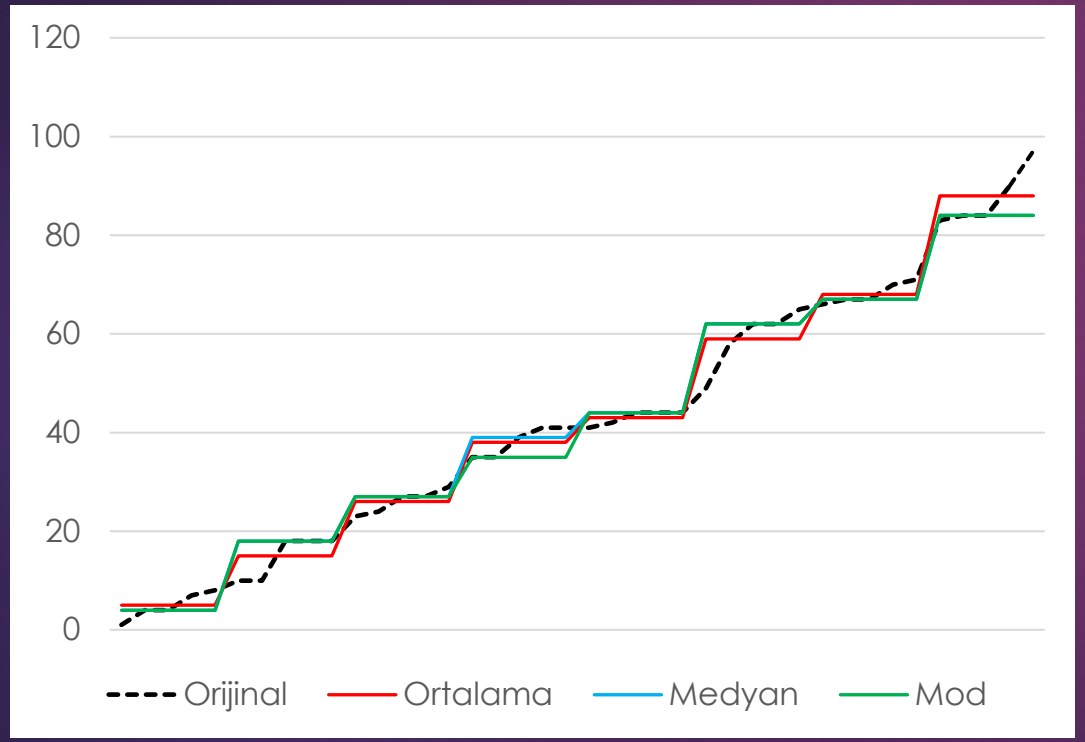
Gürültülü Verileri Düzleştirme (Smothing)

Gürültü, ölçülen değişkendeki rastgele hata ya da değişkenlik olarak tanımlanır.

- Kova Yöntemi (Binning): Sıralanmış değerleri komşuluğundaki değerlerle değiştirerek daha makul veriler elde etme işlemi.
- Kümeleme (Clustering): Aykırı değerleri tespit ederek ortadan kaldırma
- Regresyon (Regression): Verileri bir fonksiyona uydurarak düzleştirme

- Kova Yöntemi
  - Veriler sıralanır.
  - Veriler eşit aralıklı ya da eşit frekanslı olarak bölünür.
  - Veriler uygun bir değerlerle değiştirilir.
  - **Örnek veri seti:**
    - {1,9,4,21,3,13,2,7,18,23,27,6}
  - **Eşit aralıklı:**
    - {1,2,3,4,6,7,9},{13,18},{21,23,27}
  - **Eşit Frekanslı:**
    - {1,2,3,4},{6,7,9,13},{18,21,23,27}

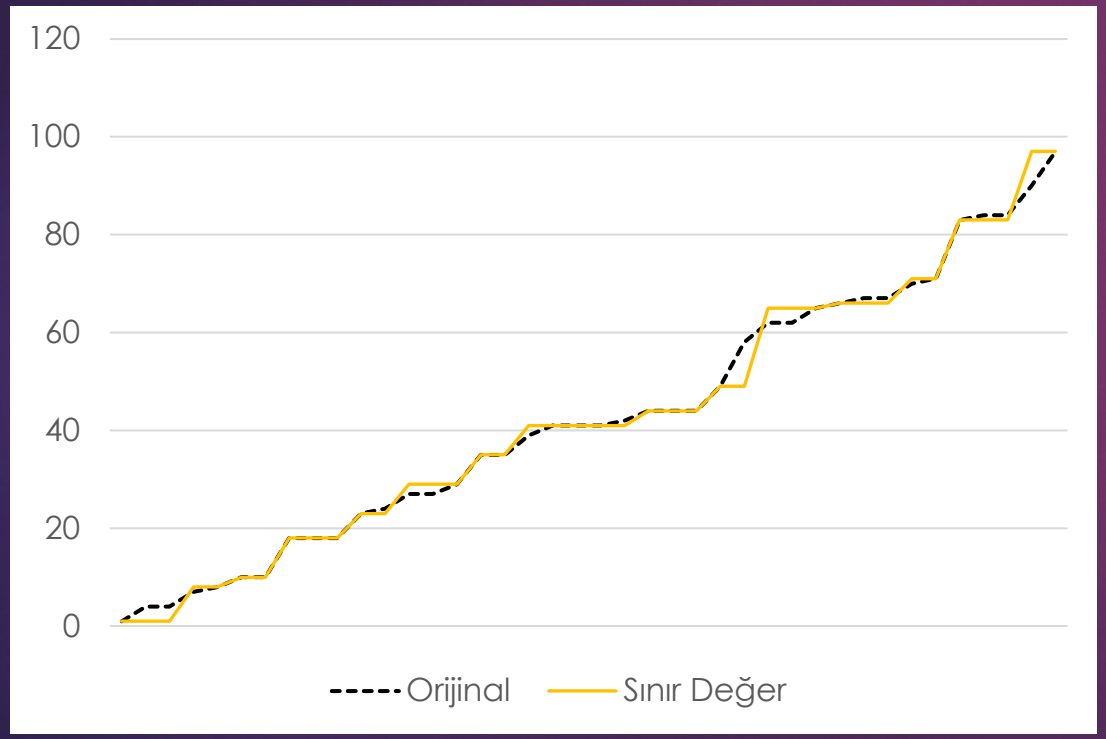
Ortalama, Medyan, Mod ile Düzleştirme



Orj.	Ort.	Med.	Mod
1	**5**	**4**	**4**
4	5	4	4
4	5	4	4
7	5	4	4
8	5	4	4
10	**15**	**18**	**18**
10	15	18	18
18	15	18	18
18	15	18	18
18	15	18	18
23	**26**	**27**	**27**
24	26	27	27
27	26	27	27
27	26	27	27
29	26	27	27
35	**38**	**39**	**35**
35	38	39	35
39	38	39	35
41	38	39	35
41	38	39	35

Orj.	Ort.	Med.	Mod
41	**43**	**44**	**44**
42	43	44	44
44	43	44	44
44	43	44	44
44	43	44	44
49	**59**	**62**	**62**
58	59	62	62
62	59	62	62
62	59	62	62
65	59	62	62
66	**68**	**67**	**67**
67	68	67	67
67	68	67	67
70	68	67	67
71	68	67	67
83	**88**	**84**	**84**
84	88	84	84
84	88	84	84
90	88	84	84
97	88	84	84

Sınır Değerler ile Düzleştirme



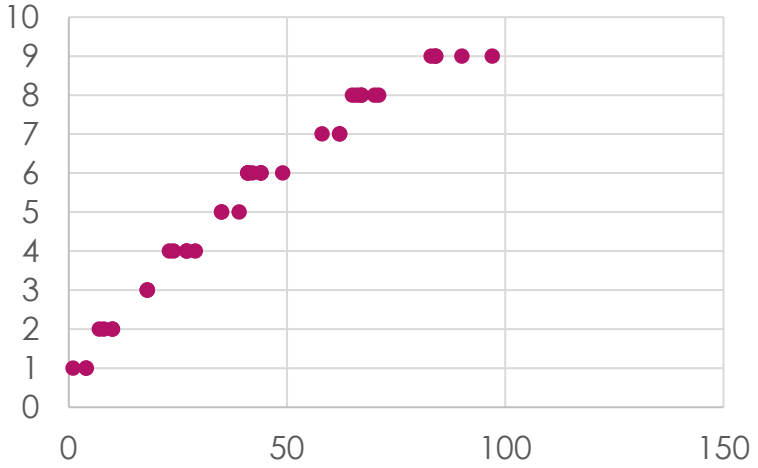
Orijinal	Sınır Değer
1	1
4	1
4	1
7	8
8	8
10	10
10	10
18	18
18	18
18	18
23	23
24	23
27	29
27	29
29	29
35	35
35	35
39	41
41	41
41	41

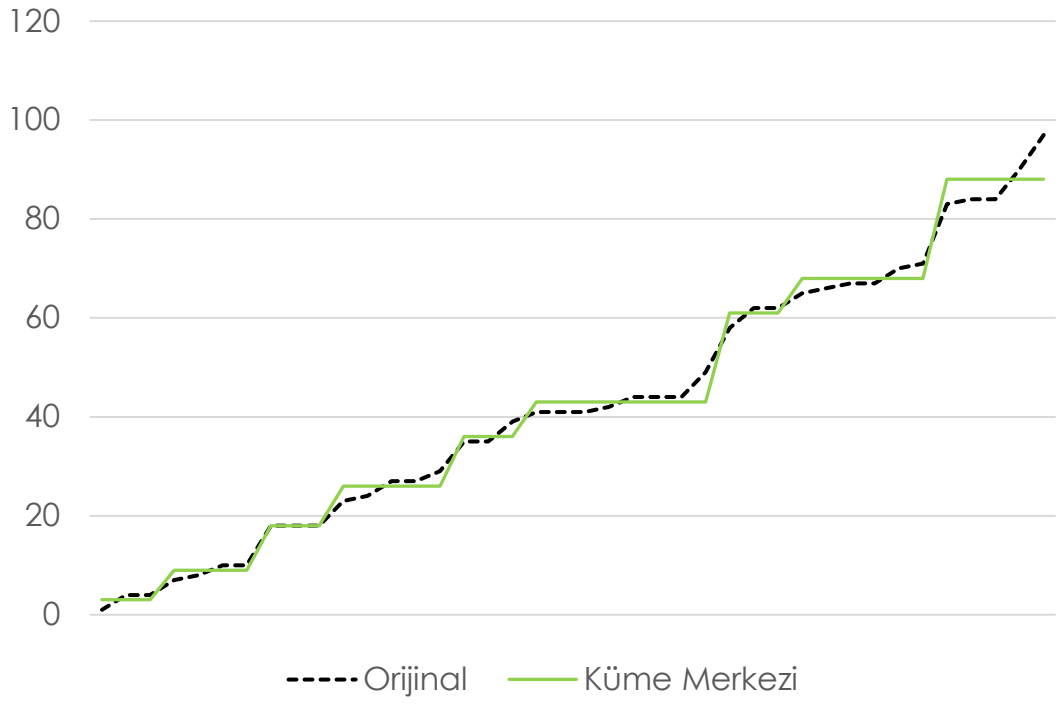
Orijinal	Sınır Değer
41	41
42	41
44	44
44	44
44	44
49	49
58	65
62	65
62	65
65	65
66	66
67	66
67	66
70	71
71	71
83	83
84	83
84	83
90	97
97	97

K-Means Kümeleme ile Düzleştirme

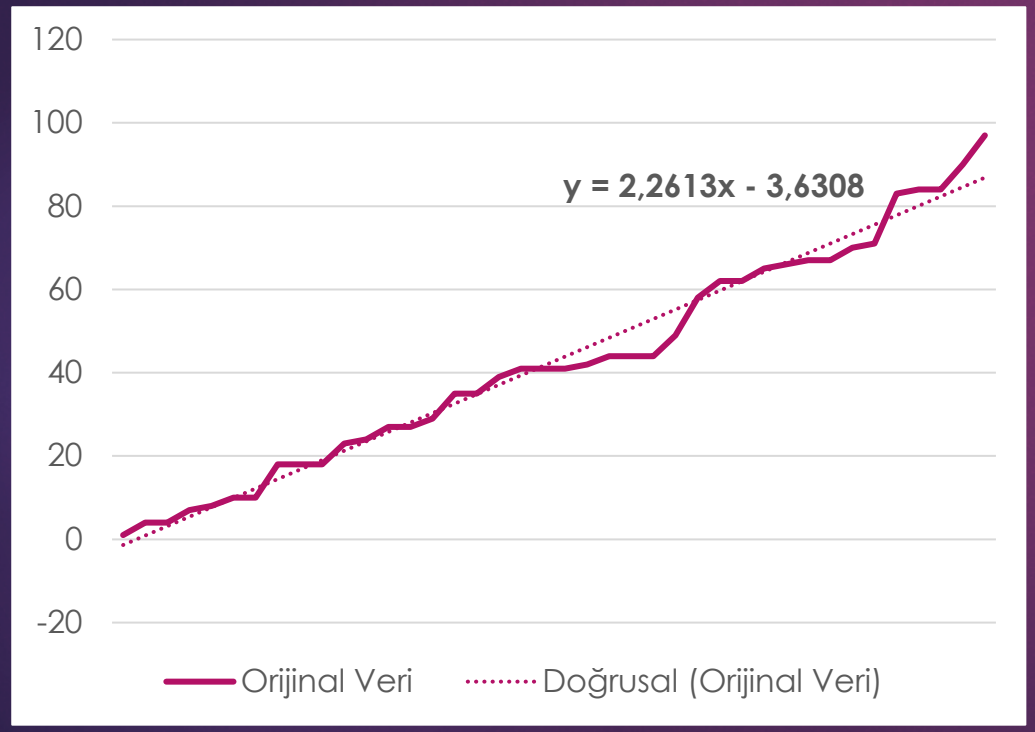
Orijinal	Küme	Merkez
1	1	3
4	1	3
4	1	3
7	2	9
8	2	9
10	2	9
10	2	9
18	3	18
18	3	18
18	3	18
23	4	26
24	4	26
27	4	26
27	4	26
29	4	26
35	5	36
35	5	36
39	5	36
41	6	43
41	6	43

Orijinal	Küme	Merkez
41	6	43
42	6	43
44	6	43
44	6	43
44	6	43
49	6	43
58	7	61
62	7	61
62	7	61
65	8	68
66	8	68
67	8	68
67	8	68
70	8	68
71	8	68
83	9	88
84	9	88
84	9	88
90	9	88
97	9	88





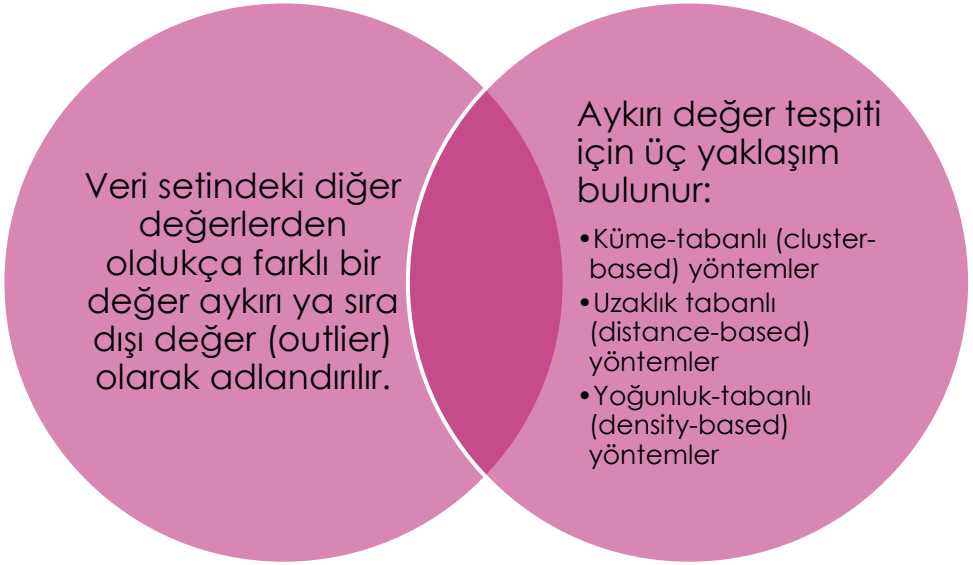
Regresyon ile Düzleştirme



Orijinal	Regresyon Doğrusu
1	-1
4	1
4	3
7	5
8	8
10	10
10	12
18	14
18	17
18	19
23	21
24	24
27	26
27	28
29	30
35	33
35	35
39	37
41	39
41	42

Orijinal	Regresyon Doğrusu
41	44
42	46
44	48
44	51
44	53
49	55
58	57
62	60
62	62
65	64
66	66
67	69
67	71
70	73
71	76
83	78
84	80
84	82
90	85
97	87

Aykırı / Sıra Dışı Değerleri Tespit Etme ve Ortadan Kaldırma



- **Küme tabanlı yöntemlerde**, aykırı değer bir veri kaydının herhangi bir kümeye ait olup olmaması, diğer küme merkezlerine olan uzaklıkları, en yakın kümenin boyutu gibi ölçülere bağlı olarak tespit edilir.
- **Uzaklık tabanlı yöntemlerde**, k-en yakın komşu yöntemine göre en büyük uzaklığa sahip veri kayıtları aykırı değerler olarak kabul edilir.
- **Yoğunluk tabanlı yöntemlerde**, aykırı değerler belli bir bölge içindeki veri kayıtlarının sayısına bağlı olarak tespit edilir.

Tutarsız Verileri Düzeltme

- Bir veri setindeki «Meslek» alanı **Meslek = « »** şeklinde girildiğinde bu değer boş (null) olarak kabul edilmez. Bunun manuel olarak uygun şekilde düzeltilmesi ya da eksik veri alanı olarak kabul edilip muhtemel değerlerle doldurulması gerekir.
- Bazı veri alanlarında mantıksal hatalar olabilir.
  - Maaş = «-2000» | Kilo = <<1398>> | Göz Rengi = «Beyaz»
- Veri setleri içinde bazı alanlardaki değerler ya da bu alanların isimleri birbiriyle tutarsız olabilir.
  - Yaş= «35», Doğum Tarihi: «03/10/2004»
  - Önceki oylama değerleri: «1,2,3», yeni oylama değerleri: «A,B,C»
  - Bir kaynakta veri alanı adı «Ad», diğerinde «İsim» şeklinde olabilir.

Tekrarlı Verileri Silme / Tekil Hale Getirme

- Veri setindeki tekrar eden kayıtlar iki probleme yol açar.
- Birden fazla kayıt gerçekte tek bir kayıt olabilir ancak bu kayıtlardaki öznitelikler farklılık gösterebilir. Bu durumda bu kayıtların uygun şekilde birleştirilmesi gerekir.
- Birbiriyle çok benzerlik gösteren ancak gerçekte ayrı kayıtlar tek bir kayda indirilebilir.
  - **Örnek**: Aynı ad ve soyada sahip iki kişiye ait kayıt

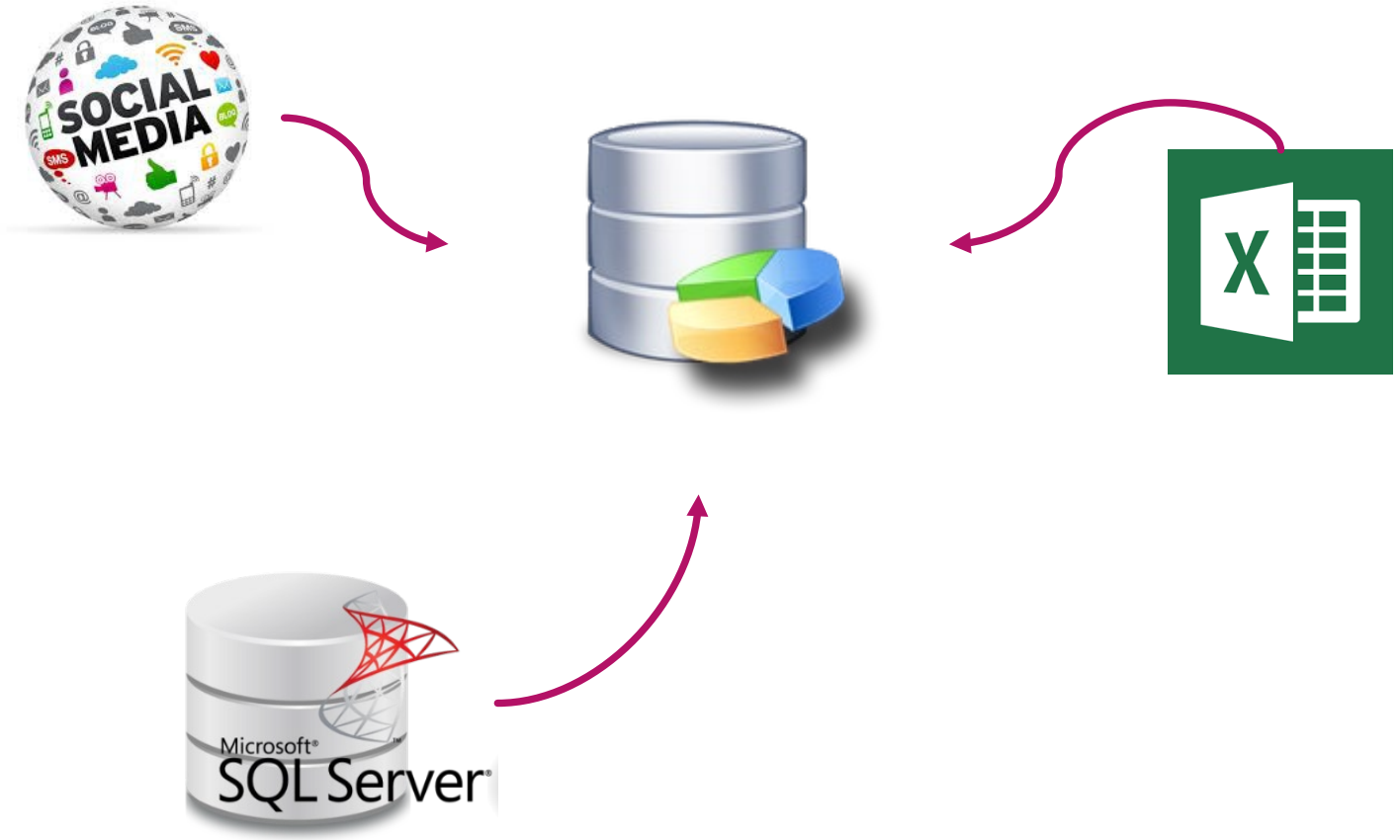
Id	Ad	Soyad	Tip	Telefon	Şehir
1	Zafer	TAŞCI	Akraba	05046543085	Ankara
2	Mevlüt	DİKMEN	Arkadaş	03184567328	Denizli
3	Murat	TAŞCI	Akraba	05148906521	Denizli

TT-PC\TT.TTENTITY - dbo.ContactInfo

piID	Name	Surname	Category	MobilePhone	cID
1	Ahmet	TAŞCI	Akraba	05397022029	İstanbul
2	Numan	ŞENEL	Tanıdık	05422627387	Sakarya
3	Ercan	YAKUT	Arkadaş	05446530740	Gümüşhane
4	Zafer	TAŞCI	Akraba	02166321278	Ankara
5	Zafer	TAŞCI	Akraba	05052883085	Ankara
6	Mehmet	GÜÇLÜ	Arkadaş	05366351666	Elazığ
7	Sinan	ORMAN	Tanıdık	05326871677	Antalya
8	Recep	SAZ	Arkadaş	05449098866	Mersin
9	Ali	ERDİNÇ	Arkadaş	02642955496	Erzurum
10	Mahmut	SELİM	Tanıdık	02642955492	Ankara

Veri Bütünleştirme

- Farklı kaynaklardaki verileri tek bir ortamda birleştirme işlemidir.
- Elde edilen tek veri kaynağı veri ambarı (data warehouse) olarak adlandırılır.
- Sıkça güncellenen veri kaynakları için bütünleştirme işlemi zorluklar içerir.



Veri Bütünleştirme İşlemleri

Farklı Kaynaklar Aynı Değerler - Birimler

Brand	Model	Display	Weight	Width	Height	Depth
Dell	XPS 13 9350	13.3	2.86	11.97	7.87	0.35

1 inç = 2,54 cm

1 kg = 2,02 lbs

Marka	Model	Ekran	Agirlik	En	Boy	Kalinlik
Dell	XPS 13 9350	29,8	1,29	30,4	20	0,9

Gereksiz Veriler

- Veri kaynakları birleştirildiğinde gereksiz veri ortaya çıkabilir. Örneğin, yıllık gelir bir kaynakta aylık değerler halinde iken diğerinde hesaplanmış olabilir.
- Gereksiz değerler korelasyon analizi ile tespit edilebilir.
- Kategorik değişkenlerin birbirinden bağımsız olup olmadığını tespit etmek için ki-kare (Chi-square) analizi yapılabilir.

Veri Dönüştürme

Verilerin uygun şekilde analiz edilebilmesi için farklı kaynaklardan gelen, şemaları, saklanma biçimleri, ölçekleri farklı olabilecek veri setlerinin tek bir veri seti haline getirilmesi işlemleri olarak düşünülebilir.

- Veri birleştirme, normalizasyon, öznitelik belirleme gibi alt adımlardan oluşur.
- Normalizasyon, veri setlerinin karşılaştırabilmesi ve birleştirilmesi için gerekli bir işlemdir.
- Farklı veri setleri birleştirilirken, bazı özniteliklerin aynı/çok benzer olup olmadığı kontrol edilmelidir. Bu durumdaki öznitelikler analiz dışı bırakılabilir.
- Gerektiği durumlarda yeni öznitelik oluşturulabilir.

Korelasyon Analizi

$$r_{A,B} = \frac{\sum_{i=1}^{n}(a_i - \bar{A}) \times (b_i - \bar{B})}{(n-1) \times \sigma_A \times \sigma_B}$$

- Değişkenler arasındaki ilişkiyi tespit etmek amacıyla kullanılan bir analiz yöntemidir.
- *r* korelasyon katsayısı -1 ile 1 arasında bir değer alır.
  - Korelasyon katsayısının pozitif çıkması beklenen durum: Reklam harcaması – Satış miktarı
  - Korelasyon katsayısının negatif çıkması beklenen durum: İçilen sigara sayısı – Koşulabilen mesafe

Dizi A (1-20)	Dizi B (1-20)
1,3	1
2	1,6
2	1,6
2,8	2,2
3,1	2,4
3,6	2,8
3,6	2,8
5,6	4,4
5,6	4,4
5,6	4,4
6,9	5,4
7,1	5,6
7,9	6,2
7,9	6,2
8,4	6,6
9,9	7,8
9,9	7,8
11	8,6
11,5	9
11,5	9

Dizi A (21-40)	Dizi B (21-40)
11,5	9
11,7	9,2
12,2	9,6
12,2	9,6
12,2	9,6
13,5	10,6
15,8	12,4
16,8	13,2
16,8	13,2
17,6	13,8
17,8	14
18,1	14,2
18,1	14,2
18,9	14,8
19,1	15
22,2	17,4
22,4	17,6
22,4	17,6
24	18,8
25,7	20,2

- $r_{A,B} = 0{,}999991$
- A ve B değişkenleri birbiriyle yüksek dereceden ilişkilidir.
- Veri bütünleştirmede bir tanesi analiz dışı bırakılabilir.

Normalizasyon

$$\bar{x}_i = \frac{x_i - \min(X)}{\max(X) - \min(X)}$$

Normalizasyona birkaç faklı nedenden dolayı ihtiyaç duyulabilir:

- Farklı ölçeklerle ölçülen değerlerin tek bir ölçeğe indirgenip (ortalama gibi) merkezi eğilimleri tespit etmek için
- Olasılık dağılımlarını, tek bir dağılıma uydurmak için
  - Örnek: öğrenci notlarını normal dağılıma uydurmak
- Farklı veri setlerindeki normalize edilmiş verileri kıyaslamak için
- Çok büyük ya da çok küçük değerler yerine birbirine daha yakın değerlerle çalışmak için

Dizi A (1-20)	Dizi B (1-20)
84	0,84
92	0,93
21	0,1
93	0,94
67	0,64
18	0,07
35	0,27
59	0,55
97	0,99
97	0,99
24	0,14
98	1
97	0,99
54	0,49
82	0,81
22	0,12
48	0,42
93	0,94
82	0,81
97	0,99

Dizi A (21-40)	Dizi B (21-40)
69	0,66
13	0,01
87	0,87
94	0,95
71	0,69
78	0,77
77	0,76
45	0,38
69	0,66
25	0,15
74	0,72
12	0
35	0,27
14	0,02
18	0,07
84	0,84
73	0,71
38	0,3
96	0,98
13	0,01

Veri / Veri Boyutu Azaltma İşlemleri

- Veri analizi ya da madenciliği için öncelikle problem tespiti yapılmalıdır.
- Çoğu durumda elde edilmek istenen sonuçlar eldeki tüm verilerin işlenmesini gerektirmez.
- Tüm verilerin veri analizi / madenciliği ile işlenmesi çok uzun zaman alabilir.
- Veri azaltma, verilerin analiz için uygun hale getirmek amacıyla veri miktarının ya da boyutunun daha küçük hale getirilmesi işlemidir.
- Veri sıkıştırma (data compression) ile analiz yapılacak veri boyutları azaltılabilir.
- DWT, PCA ile boyutsal, histogram, regresyon ile sayısal veri azatma yapılabilir.
- Veri azaltmada aşağıdaki yöntemler kullanılır.
  - **Veri seçimi:** Verinin belli bir kısmı ile çalışma
  - **Örnekleme (Sampling):** Tüm veri setini temsil edebilen alt bir veri seti ile çalışma
  - **Öznitelik Seçimi:** Analiz için ihtiyaç duyulabilecek sayıda öznitelik ile çalışma
  - **Boyut Azaltma:** Belli öznitelikleri birleştirerek daha az sayıda öznitelik ile çalışma

Veri Örnekleme (Data Sampling)

- Veri setinin analizde kullanılacak olan bir alt kümesinin belirlenmesi işlemidir.
- Tüm veri setinin işletilmesi pahalı ve zaman alıcı bir işlem olduğundan çoğu durumda örneklemeye gereksinim duyulur.
- Temel fikir tüm veri setini temsil edebilecek daha az sayıda kayıt içeren bir veri seti elde etmektir.

- Seçilen kayıt veri setine dahil edilerek (with replacement) ya da edilmeyerek (without replacement) örnekleme yapılabilir.

- Sistematik (Systematic) Örnekleme
- Basit Rassal(Random) Örnekleme
- Tabakalı (Stratified) Rassal Örnekleme
- Küme (Cluster) Örnekleme

Veriyi Kesikli Hale Getirme (Data Discretization)

- Öznitelikler çoğunlukla 3 tiptir: Kategorik (Nominal), Sıralı (Ordinal), Sürekli (Continuous)
- Veriyi kesikli hale getirme, sürekli bir özniteliği aralıklara bölerek ayrık öznitelik oluşturma olarak tanımlanır.
- Örnekler:
  - Başarı notları düşük, orta, yüksek gibi sıralı hale getirilebilir.
  - Ağırlık değerleri aşağı ya da yuvarlanabilir.
- Ayrık öznitelikler sürekli özniteliklere göre daha küçük boyut gerektirirler.
- Veri analizi/ madenciliği işlemleri kesikli özniteliklerle daha hızlı ve etkin olarak yapılabilir.
- Veriyi kesikli hale getirme ile birlikte değerlendirilen diğer bir husus kategorik veri için kavramsal hiyerarşi oluşturma (Concept Hierarchy Generation) işlemidir.
  - Örnek : {Sokak, Cadde, İlçe, İl, Ülke}