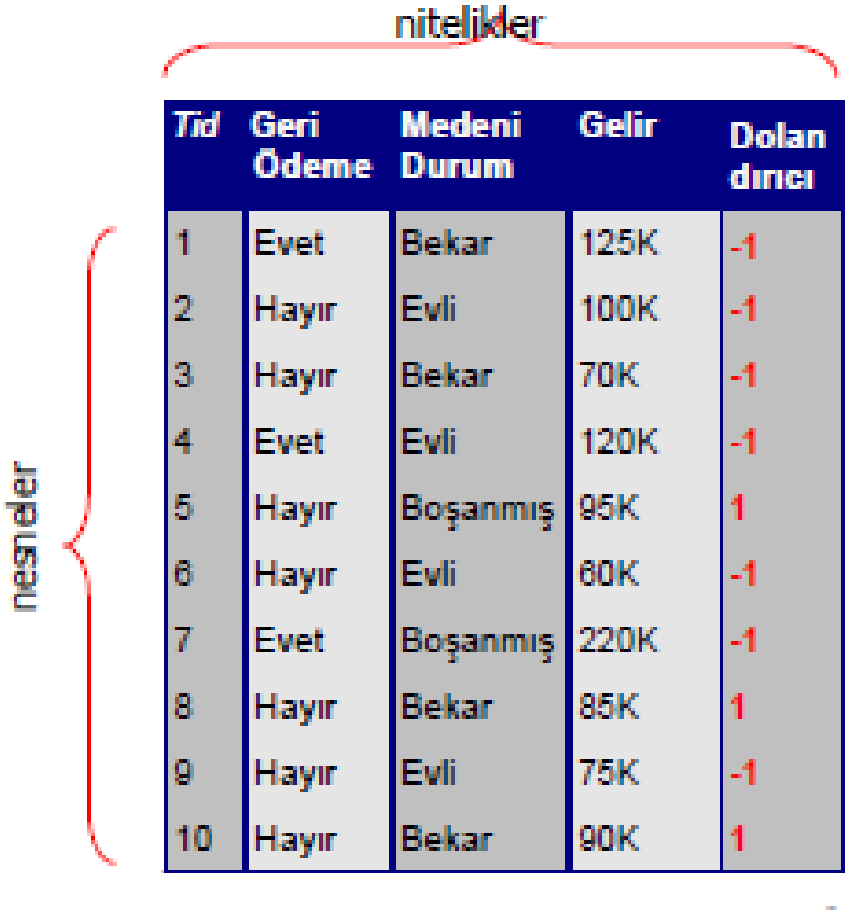
VERİ MADENCİLİĞİ DERSİ SUNU-2

Doç.Dr.Mehmet Akif ŞAHMAN

VERİ NEDİR?

- Nesneler ve nesnelerden oluşan küme
  - Kayıt(record),varlık(entity), örnek(instance) nesne için kullanılır.
- Nitelik (attribute) bir nesnenin (object) bir özelliğidir.
  - Bir insanın yaşı, ortam sıcaklığı vb.
  - Boyut(dimension), özellik (feature, characteristic) olarak da kullanılır.
- Nitelikler ve bu niteliklere ait değerler bir nesneyi oluşturur.



Tid	Geri Ödeme	Medeni Durum	Gelir	Dolandırıcı
1	Evet	Bekar	125K	-1
2	Hayır	Evli	100K	-1
3	Hayır	Bekar	70K	-1
4	Evet	Evli	120K	-1
5	Hayır	Boşanmış	95K	1
6	Hayır	Evli	60K	-1
7	Evet	Boşanmış	220K	-1
8	Hayır	Bekar	85K	1
9	Hayır	Evli	75K	-1
10	Hayır	Bekar	90K	1

DEĞER KÜMELERİ

- Nitelik için saptanmış sayılar veya semboller,
- Nitelik & Değer Kümeleri
  - Aynı nitelik farklı değer kümelerinden değer alabilir. Örn. Ağırlık : kg, lb (libre)
  - Farklı nitelikler aynı değer kümesinden değer alabilirler. Örn. ID, yaş : her ikisi de sayısal değer alırlar.

İstatiksel Veri Türleri

- **Nümerik Veriler :** Sayısal-Nümerik-Nicel veriler de denmektedir. Boy, yaş gibi süreklilik arz eden değerler Nümerik verileridir. 'Daha fazla' ifadesi ile kullanılabilirler. Sürekli ve süreksiz olarak iki başlıkta ele alınabilir:
  - Sürekli Nümerik Veriler : Yaş, Sıcaklık
  - Aralıklı Nümerik Veriler : Çocuk Sayısı, Kaza sayısı
- **Nominal Veriler :** Kategorik bir veri çeşididir. 'Daha fazla' ifadesi ile kullanılamazlar. İkiye ayrılır:
  - İkili (Binary) Veriler : Var-Yok, Kadın-Erkek, Hasta-Sağlıklı …
  - İkiden Çok Kategorili : Medeni Durum, Renk, Irk, Şehir…

İstatiksel Veri Türleri

- **Ordinal Veriler :** Ordinal veriler de yine kategorik veri türündendir. Fakat değerleri arasında sıralı bir ilişki bulunmaktadır. 'Daha fazla' ifadesi ile kullanılabilirler ancak ne kadar daha fazla olduğunun ölçüsünü veremezler. Örn. Eğitim düzeyi, sosyoekonomik ölçek skorları gibi. Nominal veriler, ordinal verilere göre daha az bilgi taşırlar.
- **Ratio Veriler :** Nümerik verilere benzerler. 100 santigrat derece, 50 santigrat derecenin iki katı denilemez ama derece kelvine çevirilirse 60 kelvin 30 kelvinin iki misli sıcak denilebilir. Oran verilebilir veri türlerine Ratio veriler denir. Burada kelvin derece ratio türünden bir değişken iken, santigrat ise nümerik veri türüne örnek olarak verilebilir.

Özet Karşılaştırma

Interval (Nümerik): Sayısal, aradaki farklar anlamlı; ancak sıfır gerçek anlam taşımaz.
Örnek: Sıcaklık değerleri (°C).
Nominal: Sadece isim veya kategori.
Örnek: Favori meyveler.
Ordinal: Sıralı, fakat aralarındaki fark ölçülemez.
Örnek: Müşteri memnuniyeti seviyeleri.
Ratio: Sayısal, aradaki fark ve oranlar anlamlı, gerçek sıfır noktası vardır.
Örnek: Ağırlık, boy, yaş.

VERİ AMBARI

Temel olarak veri madenciliği çalışmalarının yapılabilmesi için veriler ve yapısal veri tabanları kullanılması gerekmektedir. Bu veriler veya veri tabanları, veri madenciliği uygulamalarında mevcut halleri ile kullanılamazlar dolayısıyla **kullanılabilir** hale getirilmesi gerekmektedir.

İşte belirli bir döneme ait, yapılacak çalışmaya göre konu odaklı olarak düzenlenmiş, birleştirilmiş ve sabitlenmiş işletmelere ait veri tabanlarına **veri ambarları** denir.

VERİ AMBARI

Veri Ambarları

Konu Odaklıdır : Birbiri ile alakalı verilerin ilişkilendirilir. Örn. Satış, sipariş, tedarik verileri gibi,

Bütünleşiktir : Birden çok veri tabanı ve kaynaklar tekrarlar ile gereksiz verilerden ayrıştırılarak birleştirilmiştir.

Belirli Bir Döneme veya Zaman Dilimine Aittir : Her veri belirli bir zaman aralığında ve zaman ile ilişkilendirilmiştir.

Uçucu Değildir : Yukarıda bahsedilen işlemlerden geçirilerek oluşturulan veriler kaydedilir, silinmez ve veri eklenmez.

VERİ AMBARI

Müşteri ID	Adı	Soyadı	Doğ.Tarihi	…
123	Hasan	Selçuk	11/10/1980	
234				

Ürün ID	Marka	Tip	Miktar	…
45600	Torku	Tip 1	300 g	
45601	Ülker	Tip 2	400 g	

Müşteri ID	Ürün ID	İşlem No	Adet	Tarih
123	45600	1	1	10.05.2016
123	45601	1	1	11.11.2016
234	45[illegible]	[illegible]	[illegible]	[illegible]

Veri Tabanına Ait Tablolar

VERİ AMBARI →

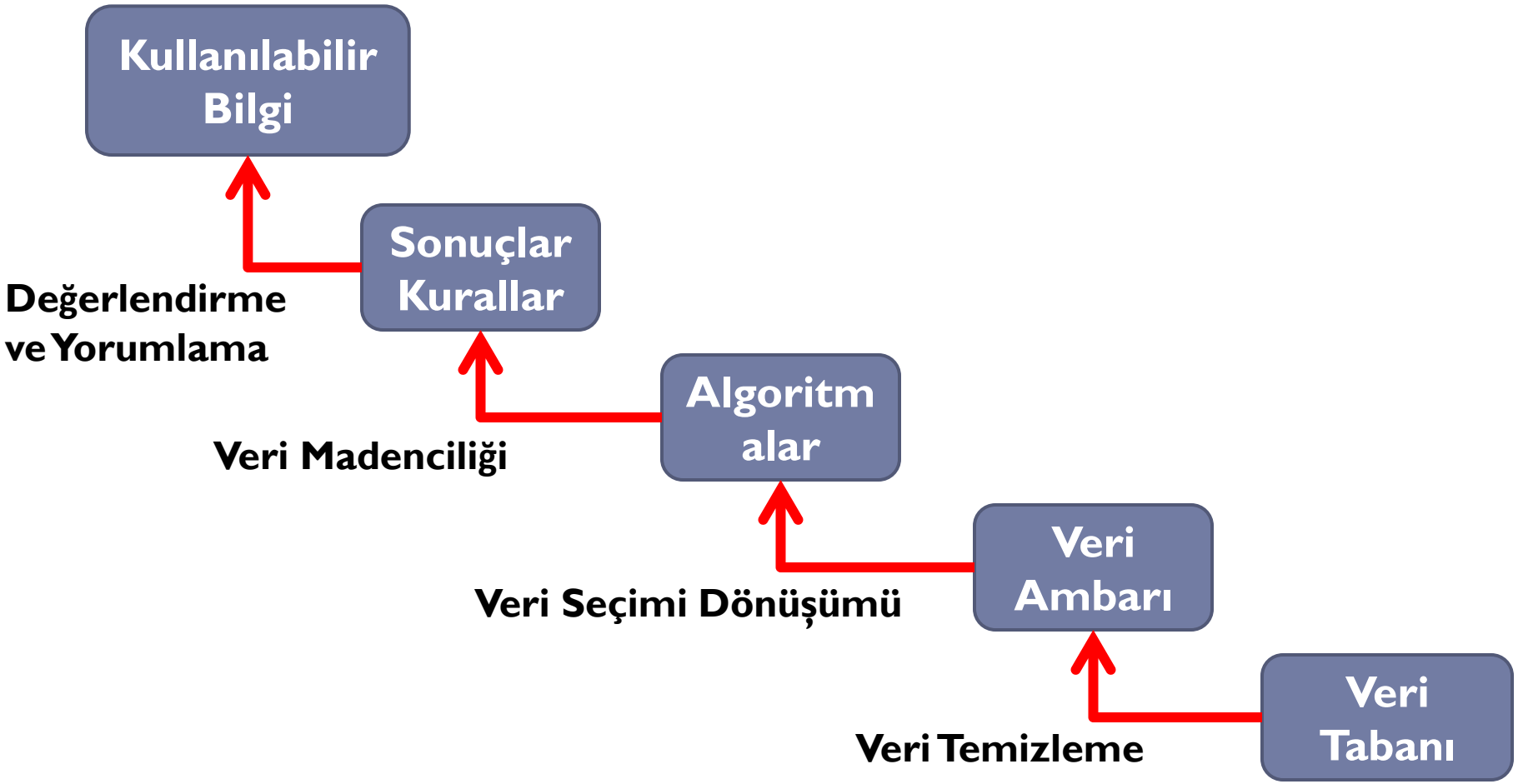
Burçlar	Marka	Alışveriş Günü	Alışveriş Miktar	…
Terazi	Torku	Pazartesi	5	
Aslan	Ülker	Salı	1	

OLAP (On-Line Analytical Processing)

Veri ambarları üzerinde, çeşitli taktik ve stratejik konular hakkında karar vermeye yardımcı olacak veri analizi ve sorgulama işlemlerine **OLAP** denir.

Veri Tabanlarındaki Sorguları	Veri Ambarlarındaki OLAP Sorguları
- Günlük ve Haftalık Satışlar - Satışların Alışlara Oranları - Depo Doluluk Oranı - En Az veya En Çok Satılan Ürünler	-Salı Günleri Süt ve Süt Ürünlerininin 10.000'i aşma Olasılığı - Kova Burcu Müşterilerden DVD Alanların Üç Ay İçinde Peynir Alma Olasılığı

Veri Madenciliği ve Bilgi Keşfi Süreci



Veri Madenciliği İçin Verilerin Hazırlanması

Veri madenciliğinin temelinde veriler bulunmaktadır. Dolayısıyla bu verilerle sağlıklı sonuç elde edilebilmesi için verilerin doğru olması önemlidir. Bu veriler elde edilirken, hatalı,eksik veya gereksiz veri girişi, gizlilik gibi durumlardan dolayı veriler yeniden düzenlenmesi gerekebilir. Düzenleme işlemleri genel olarak aşağıdaki iki başlık altında toplanabilir;

1- Verilerin Temizlenmesi

2- Verilerin Yeniden Yapılandırılması

Veri Önişleme

- **Veri Temizleme :** Eksik nitelik değerlerini tamamlama, hatalı veriyi düzeltme, aykırılıkları saptama ve temizleme, tutarsızlıkları giderme
- **Veri Birleştirme :** Farklı veri kaynağındaki verileri birleştirme
- **Veri Dönüşümü :** Normalizasyon ve biriktirme
- **Veri Azaltma :** Aynı veri madenciliği sonuçları elde edilecek şekilde veri miktarını azaltma

Veriyi Tanımlayıcı Özellikler

- Amaç : Veriyi daha iyi anlamak
  - Merkezi eğilim (central tendency), varyasyon, yayılma, dağılım
- Verinin Dağılım Özellikleri
  - Ortanca, en büyük, en küçük, sıklık derecesi, aykırılık, varyans
- Sayısal Nitelikler -> sıralanabilir değerler
  - Verinin dağılımı
  - Kutu grafik çizimi ve sıklık derecesi incelemesi

Merkezi Eğilimi Ölçme

- Ortalama
  - **Ağırlıklı ortalama :** Ağırlıklı ortalama, her bir verinin belirli bir ağırlıkla çarpılıp, bu çarpımların toplamının ağırlıkların toplamına bölünmesiyle bulunur.
  - **Kırpılmış ortalama :** Kırpılmış ortalama, veri setindeki uç değerlerin belirli bir yüzdesinin çıkarılmasıyla hesaplanan bir ortalamadır.

$$\{1, 2, 3, 4, 5, 6, 100\}$$

Bu veri setinde %20 kırpılmış ortalama hesaplamak için veri setinin %20'sini (yaklaşık 1.4 eleman, yani 1 eleman) uç değerlerden çıkarırız:

$$\{2, 3, 4, 5, 6\}$$

Kalan değerlerin ortalamasını alırız:

$$Kırpılmış\ Ortalama = \frac{2+3+4+5+6}{5}$$

$$Kırpılmış\ Ortalama = \frac{20}{5}$$

$$Kırpılmış\ Ortalama = 4$$

Ders	Not	Kredi
Matematik	80	3
Fizik	90	2
Kimya	70	1

$$Ağırlıklı\ Ortalama = \frac{(80 \cdot 3) + (90 \cdot 2) + (70 \cdot 1)}{3+2+1}$$

$$Ağırlıklı\ Ortalama = \frac{240 + 180 + 70}{6}$$

$$Ağırlıklı\ Ortalama = \frac{490}{6}$$

$$Ağırlıklı\ Ortalama \approx 81.67$$

Merkezi Eğilimi Ölçme

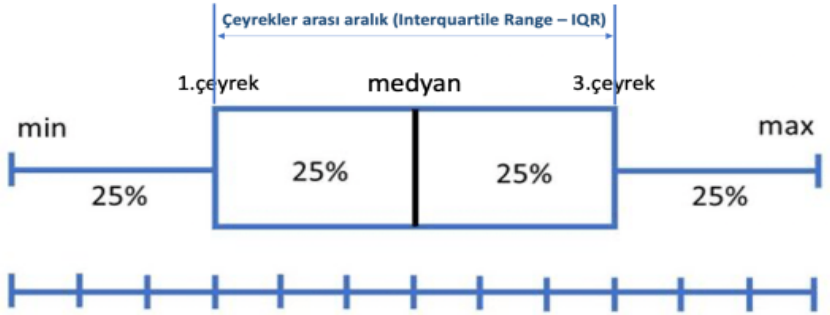
- Ortanca (median) : verinin tümü kullanılarak hesaplanır, veri sayısı tekse ortadaki değer, çift ise ortadaki iki sayının ortalaması medyanı verir.
- Mod
  - Veri içinde en sıklıkla görülen değer
  - Unimodal, bimodal, trimodal

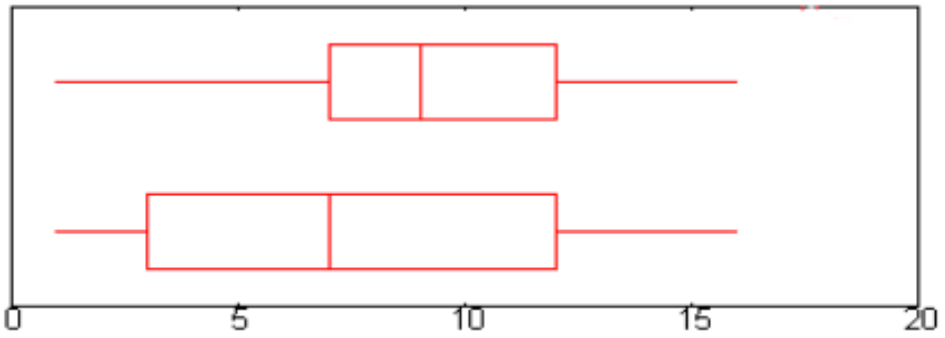
Örneğin 2, 1, 5, 4, 5, 1, 2, 3, 5 serisi sıralanırsa 1, 1, 2, 2, 3, 4, 5, 5, 5 serisi elde edilir. Bu seri 9 elemanlı olduğundan ortadaki, yani 5. eleman (medyan) olacaktır. 5. eleman 3 sayısıdır.

Örneğin 1, 5, 4, 5, 1, 3, 5 serisinde en çok tekrar eden sayı 5 sayısıdır ve frekansı 3'tür (3 kez tekrar etmiş).
Bazi serilerin 2 modu olabilir. Örneğin 1, 5, 4, 5, 3, 4, 5, 2, 3, 4 serisinde 4 ve 5 sayıları 3'er kez tekrar etmişlerdir. Bu durumda bu serinin 4 ve 5 olmak üzere iki modu vardır.

Veri Dağılımını Ölçme

- Çeyrek (quartile) : nitelik değerler küçükten büyüğe sıralandıktan sonra, %25'lik olarak bölünür, Q1, Q2, Q3, Q4
- Dörtlü Aralık : IQR = Q3-Q1
- Beş Sayılı Özet : min, Q1, median, Q3, max
- Kutu Grafiği Çizimi
- Q1 ve Q3 arasında bir kutu,
- Kutu içinde ortanca noktayı gösteren bir çizgi,
- Kutudan min ve max noktalarına çizilen bir çizgiden oluşur.
- Aykırılıklar ise 1,5x IQR değerinden küçük/büyük olan değerlerdir.





Örneklem 1 (üstte) : X_{min}=1 , Q1= 7, X_{med}= 9 , Q3 = 12 , X_{maks} = 16.

Örneklem 2 (altta) : X_{min}=1 , Q1= 3 , X_{med}= 7 , Q3 = 12 , X_{maks} = 16.

Kutu Grafiği

Adım 1: Veriyi Küçükten Büyüğe Sırala.
Örneğin, elimizde şu veri olsun:

3,7,8,5,12,14,21,15,18,14

Sıralı hali:

3,5,7,8,12,14,14,15,18,21

Adım 2: Çeyrekleri ve Medyanı Belirle.
•**Medyan (Q_2)**: Ortadaki değer. 10 gözlem olduğu için ortadaki iki değerin ortalaması alınır: (12+14)/2=13

•**Birinci Çeyrek (Q_1)**: İlk yarının medyanı: (5+7)/2=6

Üçüncü Çeyrek (Q_3): İkinci yarının medyanı: (15+18)/2=16.5

Adım 3: IQR ve Aykırı Değer Sınırlarını Hesapla.
IQR: Q3−Q1=16.5−6=10.5

Aykırı Değer Sınırları:

Alt Sınır: Q1−1.5×IQR=6−1.5×10.5 =6−15.75=−9.75
Üst Sınır: Q3+1.5×IQR=16.5+15.75=32.25

Bu aralığın dışında kalan değerler aykırı kabul edilir. (Bu örnekte yoktur.)

Adım 4: Minimum ve Maksimum Değerleri Seç.
•Minimum: 3
•Maksimum: 21
3. Sonuç Olarak:
•**Kutu (Box)**: Q1 = 6, Medyan = 13, Q3 = 16.5
•**Bıyıklar (Whiskers)**: Minimum = 3, Maksimum = 21
•**Aykırı Değerler**: Yok

Veri Dağılımını Ölçme

- Standart Sapma : Verilerin (notların) aritmetik ortalamadan sapmalarının karelerinin aritmetik ortalamasının kare köküdür.

σ : standart sapma

X_i: i inci öğrencinin notu

μ: ilgili dersin aritmetik ortalaması

n: öğrenci sayısı

$$\sigma = \sqrt{\frac{\sum (X_i - \mu_x)^2}{n}}$$

- Varyans : Varyans, verilerin aritmetik ortalamadan sapmalarının karelerinin toplamıdır. Standart sapmanın karesi varyans'a eşittir.

$$s^2 = \frac{\sum (X_i - \mu_x)^2}{n}$$

Veri Dağılımını Ölçme

Diyelim ki bir şehirde son 5 gün boyunca gündüz sıcaklıkları şu şekilde ölçüldü:

Gün	Sıcaklık (°C)
1	20
2	22
3	18
4	25
5	21

1. Ortalamayı Bulalım:

$$\mu = \frac{20 + 22 + 18 + 25 + 21}{5} = 21.2^\circ\text{C}$$

2. Varyansı Hesaplayalım:

Önce her değerin ortalamaya göre farkının karesini alalım:

$$(20 - 21.2)^2 = 1.44$$

$$(22 - 21.2)^2 = 0.64$$

$$(18 - 21.2)^2 = 10.24$$

$$(25 - 21.2)^2 = 14.44$$

$$(21 - 21.2)^2 = 0.04$$

Bunların ortalamasını alırsak:

$$\sigma^2 = \frac{1.44 + 0.64 + 10.24 + 14.44 + 0.04}{5} = \frac{26.8}{5} = 5.36$$

Bu, **varyanstır** ve sıcaklık biriminin karesindedir (**°C²**), bu yüzden yorumlaması zordur.

3. Standart Sapmayı Hesaplayalım:

$$\sigma = \sqrt{5.3} \approx 2.32^\circ\text{C}$$

Veri Madenciliği İçin Verilerin Hazırlanması

Verilerin Temizlenmesi: Bu aşamada kirli (kayıp veya gürültülü) veriler temizlenir.

Kayıp veriler için yapılabilecek işlemler;

1- Kayıp verinin bulundu kayıtları veriler kümesinden çıkarmak

2- Kayıp verileri elle tek tek tamamlamak

3- Tüm kayıp verilere aynı veriyi girmek

4- Kayıp verilere tüm verilerin ortalama değerinin verilmesi

5- Regresyon yöntemi ile kayıp verilerin tahmin değerinin verilmesi

Veri Madenciliği İçin Verilerin Hazırlanması

Gürültülü (aykırı değerler içeren veriler) veriler için düzgünleştirme işlemleri yapılır.

1-Aritmetik ortalamaya göre düzgünleştirme,

2- Sınırlar yardımıyla düzgünleştirme,

3- Uç noktadaki verilerin ortalaması yardımıyla düzgünleştirme

Veri Madenciliği İçin Verilerin Hazırlanması

1-Aritmetik ortalamaya göre düzgünleştirme,

D : {31, 12, 14, 11, 9, 5, 4, 1, 8, 6, 3, 2} elimizde bu veriler olsun;

Öncelikle veriler küçükten büyüğe sıralanır,

D:{1, 2, 3, 4, 5, 6, 8, 9, 11, 12, 14, 31} daha sonra eşit derinlikte alt kümelere bölünür,

D1:{1,2,3,4}, D2:{5,6,8,9}, D3:{11,12,14,31} sonra kümelere ait aritmetik ortalamalar kümelere verilir.

D1:{2.5,2.5,2.5,2.5}, D2:{7,7,7,7}, D3:{17,17,17,17} sonuç olarak küme;

D:{2.5,2.5,2.5,2.5,7,7,7,7,17,17,17,17} olarak düzgünleştirilir.

Veri Madenciliği İçin Verilerin Hazırlanması

2- Sınırlar yardımıyla düzgünleştirme,

D : {31, 12, 14, 11, 9, 5, 4, 1, 8, 6, 3, 2} elimizde bu veriler olsun;

Öncelikle veriler küçükten büyüğe sıralanır,

D:{1, 2, 3, 4, 5, 6, 8, 9, 11, 12, 14, 31} daha sonra eşit derinlikte alt kümelere bölünür,

D1:{**1**,2,3,**4**}, D2:{**5**,6,8,**9**}, D3:{**11**,12,14,**31**} sonra alt kümelere ait uç noktalar belirlenir. Küme elemanları hangi, uç noktasına daha yakınsa o sınır değeri alınır.

D1:{1,1,4,4}, D2:{5,5,9,9}, D3:{11,11,11,31} sonuç olarak küme;

D:{1,1,4,4,5,5,9,9,11,11,11,31} olarak düzgünleştirilir.

Veri Madenciliği İçin Verilerin Hazırlanması

3- Uç noktadaki verilerin ortalaması yardımıyla düzgünleştirme

D : {31, 12, 14, 11, 9, 5, 4, 1, 8, 6, 3, 2} elimizde bu veriler olsun;

Öncelikle veriler küçükten büyüğe sıralanır,

D:{1, 2, 3, 4, 5, 6, 8, 9, 11, 12, 14, 31} daha sonra eşit derinlikte alt kümelere bölünür,

D1:{**1**,2,3,**4**}, D2:{**5**,6,8,**9**}, D3:{**11**,12,14,**31**} sonra alt kümelere ait uç noktalar belirlenir.

D1=(4-1)/4 = 0.75, D2=(9-5)/4=1 ve D3=(31-11)/4=5 olur.

D1: {0.75, 0.75, 0.75, 0.75}, D2:{1, 1, 1, 1}, D3:{5, 5, 5, 5} sonuç olarak küme;

D: {0.75, 0.75, 0.75, 0.75,1, 1, 1, 1,5, 5, 5, 5} olarak düzgünleştirilir.

Veri Madenciliği İçin Verilerin Hazırlanması

Verilerin Yeniden Yapılandırılması : Bu aşamada veriler amaca ve kullanılacak yönteme göre uygunlaştırılmalıdır. Örn. Karar ağaçları kullanılarak bir işlem yapılacaksa sürekli değerler yerine aralıklı değerler kullanılmalıdır. 550-15000 TL arasındaki değişen ücret yerine, 550-1000, 1001-2000, 2001-4000,4001-12000, 12001-15000 gibi aralıklara bölebiliriz. Genellikle normalizasyon işlemleri kullanılır. Normalizasyon çeşitleri;

1- Min-Max yöntemi

2- Sıfır Ortalama yöntemi

Veri Madenciliği İçin Verilerin Hazırlanması

Min-Max yöntemi : Bu yöntem ile sınır değerleri belli olan bir değerin 0-1 aralığında veya farklı bir değer aralığında normalize etmek için kullanırız. Formülü,

$$s' = \frac{s - min}{maks - min} . (AraMaks - AraMin) + AraMin$$

s' : Normalize edilen değer

s : Aralıktaki normalize edilecek değer

min : Aralığın minimum değeri

$maks$: Aralığın maksimum değeri

$AraMin$: Normalize edilmek istenen aralığın minimum değeri

$AraMaks$: Normalize edilmek istenen aralığın maksimum değeri

Veri Madenciliği İçin Verilerin Hazırlanması

Min-Max yöntemi

Örn. 18-89 arasında değerler alınabilsin 37 değerini 0-1 aralığında normalize edelim.

S_{yeni} = (((37-18)/(89-18)) * (1-0))+0 = 0,268 olarak bulunacaktır.

Örn. 37 değeri 1-5 arasında normalize edilmek istenirse,

S_{yeni} = (((37-18)/(89-18)) * (5-1))+1 = 2 olarak bulunacaktır.

Veri Madenciliği İçin Verilerin Hazırlanması

Sıfır Ortalama yöntemi : Bu normalizasyon yönteminde ortalama ve standart sapma değerleri kullanılarak normalizasyon işlemi yapılmaktadır.

S_{yeni} = (s-ort)/std

Önceki verilen örnekte yaş değişkenlerinin ortalması 48,96 ve standart sapması 19,11 olursa 37 değerinin normalizasyonu = (37-48.96) / 19.11 = -0.62 olarak bulunur.