



YTU

YILDIZ TEKNİK ÜNİVERSİTESİ
Fen Edebiyat Fakültesi

YAPAY ZEKAYA GİRİŞ

3. Hafta

Veri nedir?

- ❑ Nitelik (attribute), bir nesnenin bir özelliğidir.
 - ❑ Meslek, Medeni Durum vb.
- ❑ Değer Kümesi, nesnelerin alabilecekleri değerleridir.
 - ❑ Meslek \in {Öğretmen, Hemşire, İşçi, Polis, Doktor, Avukat}
 - ❑ Gelir \in [5500,12500]
- ❑ Nitelikler ve bu niteliklere ait değerler, bir nesneyi oluşturur.
- ❑ Veri, nesneler ve nesnelerin niteliklerinden oluşan kümedir.

Nesneler

Nitelikler

SıraNo	Meslek	Medeni Durum	Gelir	Eğitim Durumu
1	Öğretmen	Bekar	8000	Lisansüstü
2	Hemşire	Bekar	6000	Lisans
3	İşçi	Evli	5000	Lise
4	İşçi	Bekar	7200	Lise
5	Polis	Evli	8500	Lisans
6	Öğretmen	Evli	8500	Lisans
7	Doktor	Evli	12000	Lisansüstü
8	İşçi	Bekar	5500	Lise
9	Polis	Evli	8250	Lisans
10	Avukat	Evli	12500	Lisans

Veri nedir?

- ❑ Nominal (Kategorik) Veri: Kategorilerden oluşan veri türüdür. 'Daha fazla' ifadesi ile kullanılmazlar.
 - ❑ Binary (İki Kategorili) Veri: Medeni Durum {Evli, Bekar}
 - ❑ İki'den Çok Kategorili Veri: Meslek {Öğretmen, Hemşire, İşçi, Polis, Doktor, Avukat}
- ❑ Ordinal (Sıralı) Veri: Kategorilerden oluşan ve kategorilerin sıra (önem, öncelik) bildirdiği veri türüdür. 'Daha fazla' ifadesi ile kullanılabilirler. Örneğin; Eğitim Durumu {Lise, Lisans, Lisansüstü}
- ❑ Aralıklı Veri: Eşit boyutta parçalara ayrılmış skala üzerinden ölçülen veri türüdür. Örneğin; Gelir [5000, 12500]
- ❑ Oransal Veri: Belli bir aralık içerisindeki sürekli değerlerden oluşan veri türüdür. Örneğin; Kilo {65.2, 68.1, 73.5, ...}

Nesneler

Nitelikler

SıraNo	Meslek	Medeni Durum	Gelir	Eğitim Durumu
1	Öğretmen	Bekar	8000	Lisansüstü
2	Hemşire	Bekar	6000	Lisans
3	İşçi	Evli	5000	Lise
4	İşçi	Bekar	7200	Lise
5	Polis	Evli	8500	Lisans
6	Öğretmen	Evli	8500	Lisans
7	Doktor	Evli	12000	Lisansüstü
8	İşçi	Bekar	5500	Lise
9	Polis	Evli	8250	Lisans
10	Avukat	Evli	12500	Lisans

Veriyi Tanımlayıcı Özellikler

- ❑ Amaç: Veriyi daha iyi anlamak.
 - ❑ Merkezi Eğilim Ölçüleri (Aritmetik Ortalama, Medyan, Mod)
 - ❑ Merkezi Yayılım Ölçüleri (Varyans, Standart Sapma, Çeyrekler (Quartiles))

Veriyi Tanımlayıcı Özellikler

☐ Merkezi Eğilim Ölçüleri:

☐ Aritmetik Ortalama:

- Anakütle: $\mu = \frac{\sum x}{N}$

- Örneklem: $\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$

5 7 4 6 8 16 11 7

$$\text{Aritmetik Ortalama} = \frac{5+7+4+6+8+16+11+7}{8} = 8$$

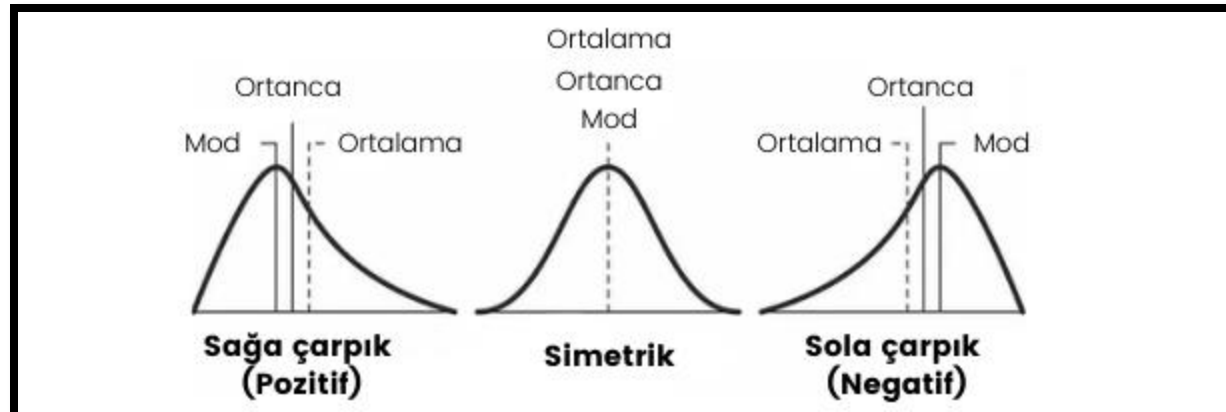
4 5 6 **7 7** 8 11 16

$$\text{Medyan} = \frac{7+7}{2} = 7$$

☐ Medyan (Ortanca): Veriler küçükten büyüğe ya da büyükten küçüğe sıralandığında ortada kalan değerdir.

☐ Mod: En çok tekrar eden değerdir.

$$\text{Mod} = 7$$



Veriyi Tanımlayıcı Özellikler

☐ Merkezi Yayılım Ölçüleri:

☐ Varyans:

- Anakütle: $\sigma^2 = \frac{1}{N} \sum_{i=1}^n (x_i - \mu)^2$
- Örneklem: $s^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2$

☐ Standart Sapma: Varyansın kareköküdür.

☐ Çeyrekler (Quartiles):

- 1. Çeyrek (Q_1): %25
- 2. Çeyrek (Q_2 , Medyan): %50
- 3. Çeyrek (Q_3): %75

5 7 4 6 8 16 11 7

$$\text{Varyans} = \frac{(5-8)^2 + (7-8)^2 + (4-8)^2 + (6-8)^2 + (8-8)^2 + (16-8)^2 + (11-8)^2 + (7-8)^2}{8-1} \cong 14.857$$

$$\text{Standart Sapma} = \sqrt{14.857} = 3.854$$

4 5 6 **7 7** 8 11 16

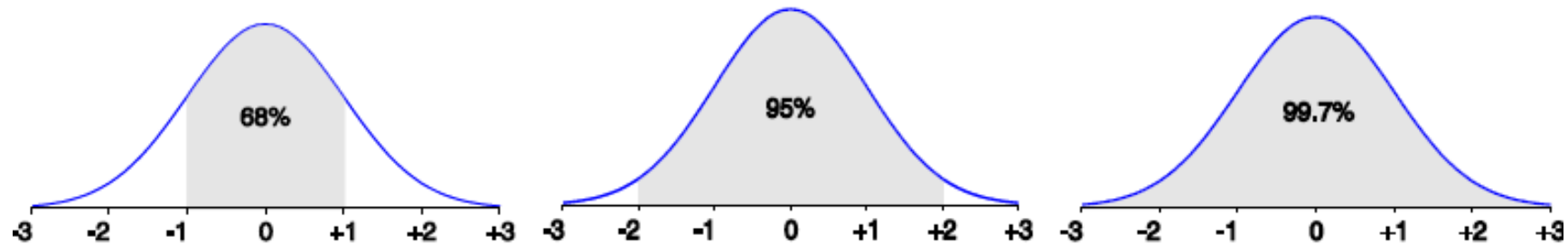
1. Çeyrek = $\frac{5+6}{2} = 5.5$ 3. Çeyrek = $\frac{8+11}{2} = 9.5$

2. Çeyrek = 7

Verinin Dağılımı

□ Normal Dağılım:

- Verilerin %68'i $\mu - \sigma$ ve $\mu + \sigma$ arasında
- Verilerin %95'i $\mu - 2\sigma$ ve $\mu + 2\sigma$ arasında
- Verilerin %99.7'u $\mu - 3\sigma$ ve $\mu + 3\sigma$ arasında



Veri Örnekleri

❑ Kirli Veri:

❑ Gerçek uygulamalarda toplanan veriler:

❑ Eksik olabilir: Bazı nitelik değerleri, bazı nesneler için girilmemiş olabilir.

❑ Meslek = ' '

❑ Gürültülü olabilir: Hatalı olabilir.

❑ Maaş = -10

❑ Tutarsız olabilir: Nitelik değerleri veya nitelik isimleri uyumsuz olabilir.

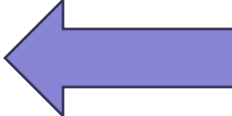
❑ Yaş = 35 iken Doğum Tarihi = 03/10/2004.

Veri Ön İşleme

- ❑ Veri Temizleme: Eksik nitelik değerlerini tamamlama, hatalı veriyi düzeltme, aykırılıkları saptama ve temizleme, tutarsızlıkları giderme
- ❑ Veri Birleştirme: Farklı veri kaynağındaki verileri birleştirme
- ❑ Veri Azaltma: Gerçek veri ile sonuçlar aynı kalacak şekilde veri dışında tutma
- ❑ Veri Dönüşümü: Normalizasyon

Veri Ön İşleme

- ❑ Veri Temizleme: Eksik, gürültülü (hatalı) veya tutarsız verilerin belirlenmesi, düzeltilmesi veya silinmesini ifade eder.
 - ❑ Eksik veri kayıtlarının nedenleri:
 - ❑ Veri toplandığı sırada bir nitelik değerinin elde edilememesi, bilinmemesi.
 - ❑ Veri toplandığı sırada bazı niteliklerin gerekliliğinin görülememesi.
 - ❑ İnsan, yazılım ya da donanma problemleri.
 - ❑ Gürültülü (hatalı) veri kayıtlarının nedenleri:
 - ❑ Hatalı veri toplama gereçleri.
 - ❑ Veri girişi problemleri.
 - ❑ Veri iletimi problemleri.
 - ❑ Teknolojik kısıtlar
 - ❑ Nitelik isimlerinde tutarsızlık.
 - ❑ Tutarsız veri kayıtlarının nedenleri:
 - ❑ Verinin farklı veri kaynaklarında tutulması.
 - ❑ İşlevsel bağımlılık kurallarına uyulmaması.



En zor belirlenebilen ve düzeltilebilen veridir. Çok dikkat edilmesi gerekir. Veri girişi sırasında kısıtlamalarla minimum düzeye indirgenmeye çalışılır.

Veri Ön İşleme

☐ Eksik veriler nasıl tamamlanır?

- ☐ Eksik nitelik değerleri olan veri kayıtlarını kullanmamak.
 - ☐ Eksik nitelik değerlerini elle doldurmak.
 - ☐ Eksik nitelik değerleri için global bir değişken kullanmak (Null, bilinmiyor, ...).
 - ☐ Eksik nitelik değerlerini o niteliğin ortalama değeri ile doldurmak.
 - ☐ Aynı sınıfa ait kayıtların nitelik değerlerinin ortalaması ile doldurmak.
 - ☐ Olasılığı en fazla olan nitelik değerleriyle doldurmak.
-

Veri Ön İşleme

❑ Gürültülü veriler nasıl düzeltilir?

- ❑ Bölümleme (Binning): Veriler sıralanır ve eşit aralıklarla bölümlere ayrılır. Her bölme ortalamayla, medyanla ve sınır verileri ile ifade edilir.
 - ❑ Eğri Uydurma (Regression): Veriler regresyon fonksiyonlarına uydurulur.
 - ❑ Kümeleme (Clustering): Veriler benzerlik durumlarına göre gruplandırılır. Aykırı ve aşırı değerler belirlenir ve silinir.
 - ❑ İnsanlar tarafından hatalı verilerin algılanması: Şüpheli değerlerin insanlar tarafından bulunması ve kontrol edilmesi.
-

Veri Ön İşleme

- ❑ Veri Birleştirme: Farklı kaynaklardan alınan verilerin tutarlı bir şekilde birleştirilmesini ifade eder.
 - ❑ Birleştirme işlemi yapılırken;
 - ❑ Tutarsız veri oluşturmamaya dikkat etmek gerekmektedir !
 - ❑ Kanseri teşhisini ortaya koymak için birleştirilen veride hastaların bıyıklı olduğu bilgisinin eklenmesi.

Veri Ön İşleme

❑ Veri Azaltma: Verinin çok fazla olduğu durumlarda, bu verilerin algoritmalar tarafından analiz edilmesi çok uzun zaman alabilir ve hatalar oluşabilir.

❑ Veri Azaltma Türleri:

❑ Boyut Azaltma: Önemsiz niteliklerin silinmesi veya kaldırılması.

❑ Wavelet dönüşüm (Wavelet transforms)

❑ Temel Bileşenler Analizi (Principle Component Analysis, PCA)

❑ Niteliklerin seçimi veya yaratılması

❑ Gözlem Azaltma:

❑ Eğri Uydurma (Regression and Log-Linear Models)

❑ Histogram, kümeleme (clustering), örnekleme (sampling)

❑ Veri küpleri

Veri miktarı çok fazla olduğu zaman veri madenciliği algoritmalarının çalışması ve sonuç üretmesi çok uzun sürebilir. Bu durumda veriyi azaltmak, başarıyı arttırır ancak sonucun (nerdeyse) hiç değişmemesi gerekir.

Veri Ön İşleme

❑ Veri Dönüşümü: Verilen niteliklerden yeni niteliklerin oluşturulmasıdır.

❑ Genelleme: Özetlemek, Veri küpü oluşturmak.

❑ Normalizasyon (İstatistiksel Normalleştirme):

1. Veriler arasında çok fazla farklılık olduğu durumda, verileri tek bir düzen içerisine almaya yarar. Veriyi daha küçük aralıklara indirgemeyi sağlar.
2. Farklı ölçekleme sistemindeki verileri aynı benzer düzen içerisine toplayarak karşılaştırabilmeye olanak tanır. Buradaki amaç, matematiksel fonksiyonlar kullanarak farklı sistemlerde bulunan verileri, ortak bir sisteme taşımak ve karşılaştırılabilir hale getirmektir.

Veri Ön İşleme

❑ Örnek:

❑ Veri türleri nelerdir?

- ❑ Meslek => Nominal, Medeni Durum => Nominal, Çocuk Sayısı => Oran, Gelir => Oran, Eğitim Durumu => Aralıklı

❑ Eksik veri var mıdır? (Varsa) eksik veriler için hangi işlemler yapılabilir?

- ❑ Evet, örnek veride eksik gözlem vardır.
- ❑ Eksik gözlem içeren satırlar incelendiğinde, Sıra numarası 1 olan satırda dolu olması beklenen 5 bilginin 2'si mevcuttur. Yani bu satırın %60'ı boştur. Dolayısıyla silinebilir.
- ❑ Silinen satırdan sonra kalan eksik gözlemlere bakıldığında, Medeni Durum bilgisi 9 kişi için dolu olması gerekirken 4 kişi için doludur. Yani bu sütunun yarısından fazlası boştur. Dolayısıyla bu sütun silinebilir.
- ❑ Tablonun son haline bakıldığında tek bir eksik bilgi olduğu görülmektedir. Sıra numarası 8 olan kişiye ait Gelir. Buradaki eksikliği gidermek için aşağıdaki uygulamalar yapılabilir.
 - ❑ 'Bilinmiyor' şeklindeki bir ifade ile doldurulabilir.
 - ❑ Gelir sütununun ortalaması alınarak (ya da medyanı bulunarak), o değerle doldurulabilir.
 - ❑ Eğitim Durumu Lise olanların ortalaması alınarak ya da Meslek bilgisi İşçi olanların ortalaması alınarak ya da Eğitim Durumu bilgisi Lise olup Meslek bilgisi İşçi olanların Gelir ortalaması alınarak doldurulabilir.
 - ❑ Eksik veriyi içeren satırı çıkararak, bir regresyon modeli oluşturup, o regresyon modeli ile Gelir bilgisini tahmin ederek doldurulabilir.

SıraNo	Meslek		Çocuk Sayısı	Gelir	Eğitim Durumu
1					
2	Hemşire		0	6000	Lisans
3	İşçi		3	5000	Lise
4	İşçi		0	7200	Lise
5	Polis		2	8500	Lisans
6	Öğretmen		1	8500	Lisans
7	Doktor		1	12000	Lisansüstü
8	İşçi		1		Lise
9	Polis		0	8250	Lisans
10	Avukat		3	12500	Lisans

Veri Ön İşleme

☐ Normalizasyon Türleri:

- ☐ Z-Skor Normalizasyonu
 - ☐ Min-Max Normalizasyonu
 - ☐ Medyan Z-Skor Normalizasyonu
 - ☐ Sigmoid Normalizasyonu
 - ☐ D_Min_Max Normalizasyonu
-

Veri Ön İşleme

❑ Normalizasyon Türleri:

❑ Z-Skor Normalizasyonu: $x'_i = \frac{x_i - \mu_i}{\sigma_i}$

- x' : Normalize edilmiş veriyi,
- x_i : Girdi değerini,
- μ_i : Girdi setinin ortalamasını,
- σ_i : Girdi setinin standart sapmasını ifade etmektedir.

Veri Ön İşleme

□ Normalizasyon Türleri:

□ Min-Max Normalizasyonu: $x'_i = \frac{x_i - x_{min}}{x_{max} - x_{min}}$

- x' : Normalize edilmiş veriyi,
- x_i : Girdi değerini,
- x_{min} : Girdi seti içerisinde yer alan en küçük sayıyı,
- x_{max} : Girdi seti içerisinde yer alan en büyük sayıyı ifade etmektedir.

Min-Max Normalizasyon yöntemi, verileri doğrusal olarak normalize eder. Minimum, bir verinin alabileceği en küçük değer iken maksimum verinin alabileceği en yüksek değeri ifade eder. Bir veri, Min-Max Normalizasyonu yöntemi ile 0-1 aralığına indirgenir.

Veri Ön İşleme

❑ Normalizasyon Türleri:

❑ Medyan Normalizasyonu: $x'_i = \frac{x_i - \text{Medyan}(x_i)}{\text{Medyan}(x_i)}$

- x' : Normalize edilmiş veriyi,
- x_i : Girdi değerini,
- $\text{Medyan}(x_i)$: Ortanca değeri ifade eder.

Medyan Normalizasyonu yöntemi, aşırı (uç) değerlerin olduğu veriler için oldukça kullanışlıdır. Çünkü medyan aşırı sapmalardan etkilenmez.

Veri Ön İşleme

❑ Normalizasyon Türleri:

❑ Sigmoid Normalizasyonu: $x'_i = \frac{1}{1+e^{x_i}}$

- x' : Normalize edilmiş veriyi,
- x_i : Girdi değerini,

Veri Ön İşleme

□ Normalizasyon Türleri:

□ D_Min_Max Normalizasyonu: $x'_i = (b - a) * \frac{x_i - x_{min}}{x_{max} - x_{min}} + a$

- x' : Normalize edilmiş veriyi,
 - x_i : Girdi değerini,
 - x_{min} : Girdi seti içerisinde yer alan en küçük sayısı,
 - x_{max} : Girdi seti içerisinde yer alan en büyük sayısı,
 - b : Fonksiyonun tanımlı olduğu üst sınırı,
 - a : Fonksiyonun tanımlı olduğu alt sınırı ifade eder.
-

Veri Ön İşleme

Örnek:

Deneyim(Yıl)	Birikim	Z_skor_Birikim	Min_Max_Birikim	Medyan_Z_Skor_Birikim	D_Min_Max_Birikim
6	100,000.00	-1.06	0.00	-0.69	1.00
7	250,000.00	-0.50	0.23	-0.23	3.08
15	750,000.00	1.37	1.00	1.31	10.00
15	150,000.00	-0.87	0.08	-0.54	1.69
18	400,000.00	0.06	0.46	0.23	5.15
34	650,000.00	1.00	0.85	1.00	8.62

□ Z-Skor Normalizasyonu $x'_i = \frac{x_i - \mu_i}{\sigma_i}$

□ Min-Max Normalizasyonu $x'_i = \frac{x_i - x_{min}}{x_{max} - x_{min}}$

□ Medyan Z-Skor Normalizasyonu $x'_i = \frac{x_i - Medyan(x_i)}{Medyan(x_i)}$

□ D_Min_Max Normalizasyonu $x'_i = (b - a) * \frac{x_i - x_{min}}{x_{max} - x_{min}} + a$

Normalizasyon işlemi yapıldıktan sonra iki sütun arasında (varsa) nasıl bir ilişki olduğu gözlemlenebilir.



Hafta_3_normalizasyon

Performans Değerlendirme Teknikleri

- ❑ Kurulacak modelin hedef değişkenine göre temelde iki farklı şekilde model kurulabilir: Sınıflandırma Modelleri, Tahmin (Regresyon) Modelleri.
 - ❑ Sınıflandırma Modelleri: Hedef değişkenin kategorik veri tipinde olduğu modellerdir.
 - ❑ Bir e-ticaret sitesinde müşterilerin bir ürünü satın alıp almayacağını bulmaya çalışırken sınıflandırma modeli oluşturulur.
 - ❑ Tahmin (Regresyon) Modelleri: Hedef değişkenin sürekli (sayısal, aralıklı veya oran) veri tipinde olduğu modellerdir.
 - ❑ Bir evin fiyatını; evin konumu, büyüklüğü, oda sayısı gibi verilerle bulmaya çalışırken tahmin (regresyon) modeli oluşturulur.
 - ❑ Bir model oluşturulduktan sonra, bu model ile yapılan tahminlerin ne kadar doğru olduğuna dair değerlendirme yapılması gerekmektedir. Sınıflandırma ve tahmin modelleri için kurulan modeli değerlendirme teknikleri farklılık göstermektedir.
-

Performans Değerlendirme Teknikleri

❑ Sınıflandırma Modelleri İçin Performans Değerlendirme Teknikleri:

- ❑ Karmaşıklık matrisi (Confusion Matrix), sınıflandırma modellerinin performansını değerlendirmek için kullanılan temel bir araçtır. Bu matris, modelin yaptığı doğru ve yanlış sınıflandırmaları dört kategoriye ayırarak sonuçları görselleştirir. Bu dört kategori şunlardır:

		Actual (Gerçek)	
		0 (True)	1 (False)
Prediction (Tahmin)	0 (Positive)	TP	FP
	1 (Negative)	FN	TN

- ❑ TP (True Positive): Gerçekte 0 olan bir olayın, model sonucunda 0 olarak tahmin edilmesi. Doğru pozitif sayısı.
- ❑ FP (False Positive): Gerçekte 1 olan bir olayın, model sonucunda 0 olarak tahmin edilmesi. Yanlış pozitif sayısı.
- ❑ FN (False Negative): Gerçekte 0 olan bir olayın, model sonucunda 1 olarak tahmin edilmesi. Yanlış negatif sayısı.
- ❑ TN (True Negative): Gerçekte 1 olan bir olayın, model sonucunda 1 olarak tahmin edilmesi. Doğru negatif sayısı.

True ve false değeri, bu modele dair gerçek sonuçları, positive ve negative ise modele dair tahminleri göstermektedir.

Performans Değerlendirme Teknikleri

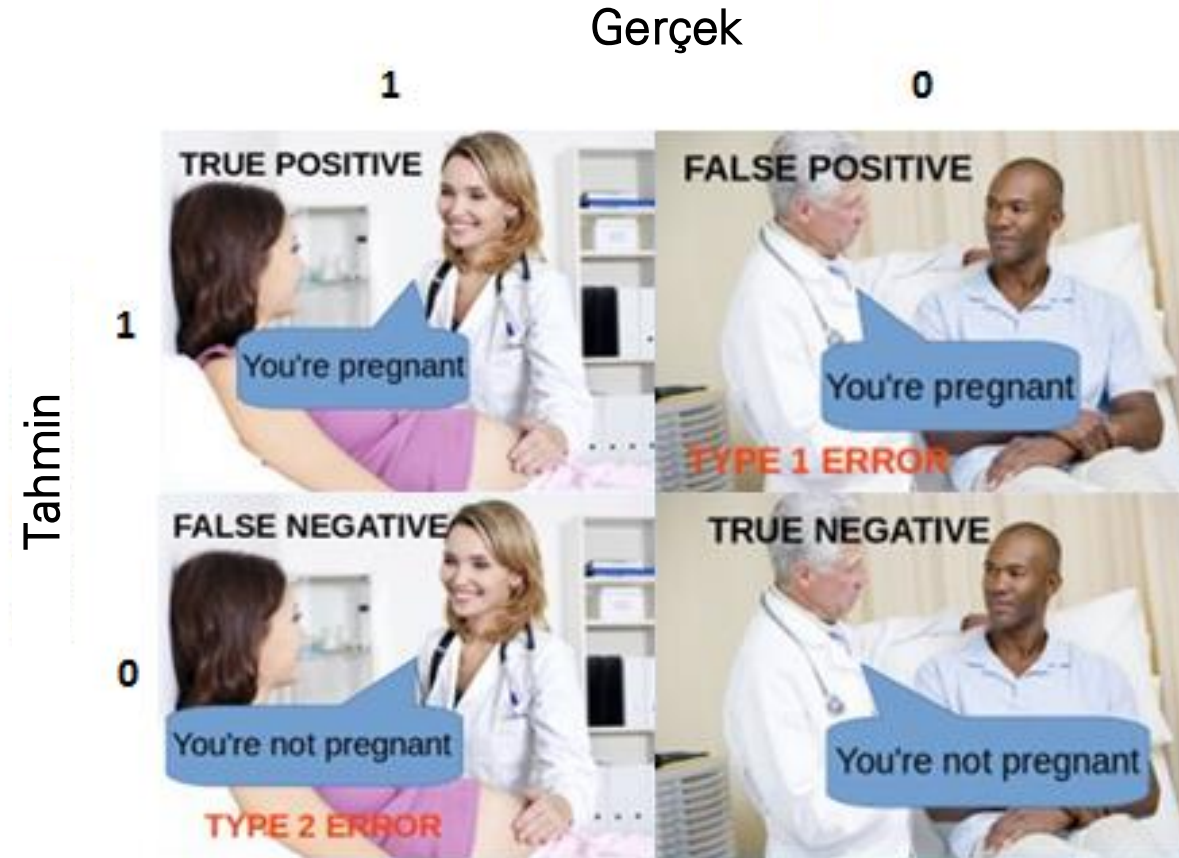
❑ Sınıflandırma Modelleri İçin Performans Değerlendirme Teknikleri:

		Actual (Gerçek)	
		0 (True)	1 (False)
Prediction (Tahmin)	0 (Positive)	TP	FP
	1 (Negative)	FN	TN

- ❑ TP (True Positive): Churn edeceğini tahmin ettiğimiz müşterilerimiz (positive), gerçekten churn etmiş (true).
 - ❑ FP (False Positive): Churn edeceğini tahmin ettiğimiz müşterilerimiz (positive), gerçekte churn etmemiş (false). — > 1. Tip Hata (Type 1 Error)
 - ❑ FN (False Negative): Churn etmeyecek dediğimiz müşteriler (negative), gerçekte churn etmiş (false). — > 2. Tip Hata (Type 2 Error)
 - ❑ TN (True Negative): Churn etmeyecek dediğimiz müşteriler (negative), gerçekte churn etmemiş (true).
-

Performans Değerlendirme Teknikleri

❑ Sınıflandırma Modelleri İçin Performans Değerlendirme Teknikleri:



Performans Değerlendirme Teknikleri

❑ Sınıflandırma Modelleri İçin Performans Değerlendirme Teknikleri:

		Actual (Gerçek)	
		0 (True)	1 (False)
Prediction (Tahmin)	0 (Positive)	TP	FP
	1 (Negative)	FN	TN

❑ Doğruluk (Accuracy): Doğru tahminlerin toplam veri kümesine oranıdır.

$$Accuracy = \frac{TP + TN}{TP + FP + FN + TN}$$

Performans Değerlendirme Teknikleri

❑ Sınıflandırma Modelleri İçin Performans Değerlendirme Teknikleri:

		Actual (Gerçek)	
		0 (True)	1 (False)
Prediction (Tahmin)	0 (Positive)	TP	FP
	1 (Negative)	FN	TN

❑ Kesinlik (Precision): Pozitif olarak tahmin edilen verilerin kaçının gerçekten pozitif olduğunu gösterir.

$$Precision = \frac{TP}{TP + FP}$$

Performans Değerlendirme Teknikleri

❑ Sınıflandırma Modelleri İçin Performans Değerlendirme Teknikleri:

		Actual (Gerçek)	
		0 (True)	1 (False)
Prediction (Tahmin)	0 (Positive)	TP	FP
	1 (Negative)	FN	TN

❑ Duyarlılık (Recall): Geliştirilen modelin pozitif olanların kaçını yakaladığını gösterir.

$$Recall = \frac{TP}{TP + FN}$$

Performans Değerlendirme Teknikleri

❑ Sınıflandırma Modelleri İçin Performans Değerlendirme Teknikleri:

		Actual (Gerçek)	
		0 (True)	1 (False)
Prediction (Tahmin)	0 (Positive)	TP	FP
	1 (Negative)	FN	TN

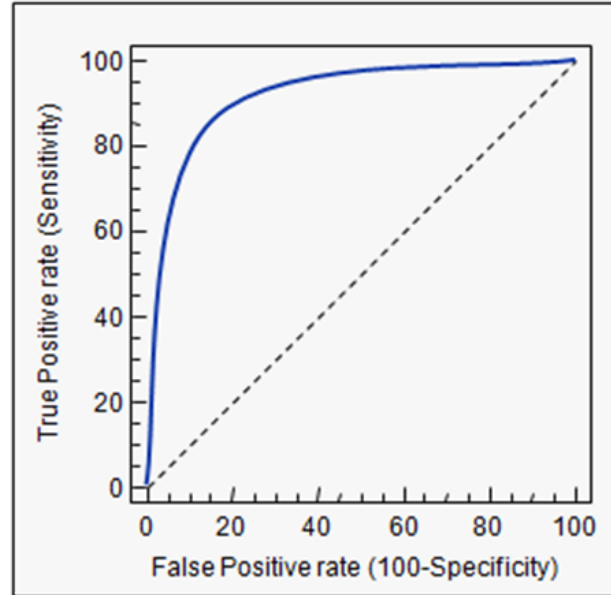
❑ F1 Skor: F1 skoru, kesinlik ve duyarlılık değerlerinin harmonik ortalamasıdır. Sınıf dağılımı benzer olduğunda doğruluk kullanılırken, dengesiz veri setlerinde F1 skor daha iyi bir metriktir.

$$F1\ skor = 2 * \frac{Kesinlik * Duyarlilik}{Kesinlik + Duyarlilik}$$

Performans Değerlendirme Teknikleri

❑ Sınıflandırma Modelleri İçin Performans Değerlendirme Teknikleri:

- ❑ ROC Eğrisi: Yanlış pozitif oranı ve gerçek pozitif oranı göz önünde bulundurularak, x ekseninde ve y ekseninde 0'dan 100'e kadar olan değerlerin üzerinde bir eğri oluşturulur. Bu eğrinin altında kalan alana Area Under Curve (AUC) adı verilir. Bu alanın büyük olması modelin başarılı olduğunu gösterir. Grafikte yer alan mavi çizgi; ne kadar geniş bir alan kaplıyorsa modelin tahmin başarısı o kadar yüksek, ortadaki kesikli çizgiye ne kadar yakınsa modelin başarı oranı o kadar düşüktür.



Performans Değerlendirme Teknikleri

❑ Örnek:

		Actual (Gerçek)	
		0 (True)	1 (False)
Prediction (Tahmin)	0 (Positive)	1 (TP)	0 (FP)
	1 (Negative)	7 (FN)	92 (TN)

❑ Doğruluk?
$$Accuracy = \frac{TP + TN}{TP + FP + FN + TN} = \frac{1 + 92}{1 + 0 + 92 + 7} = 0.93$$

❑ Kesinlik?
$$Precision = \frac{TP}{TP + FP} = \frac{1}{1 + 0} = 1.00$$

❑ Duyarlılık?
$$Recall = \frac{TP}{TP + FN} = \frac{1}{1 + 7} = 0.125$$

❑ F1 Skor?
$$F1\ skor = 2 * \frac{Kesinlik * Duyarlilik}{Kesinlik + Duyarlilik} = 2 * \frac{1 * 0.125}{1 + 0.125} = 2 * \frac{0.125}{1.125} = 0.22$$

Performans Değerlendirme Teknikleri

❑ Örnek:

		Actual (Gerçek)	
		0 (True)	1 (False)
Prediction (Tahmin)	0 (Positive)	13	5
	1 (Negative)	7	28

❑ Doğruluk? $Accuracy = \frac{TP + TN}{TP + FP + FN + TN} = \frac{13 + 28}{10 + 5 + 7 + 28} = 0.82$

❑ Kesinlik? $Precision = \frac{TP}{TP + FP} = \frac{13}{11 + 5} = 0.81$

❑ Duyarlılık? $Recall = \frac{TP}{TP + FN} = \frac{13}{11 + 7} = 0.72$

❑ F1 Skor? $F1\ skor = 2 * \frac{Kesinlik * Duyarlilik}{Kesinlik + Duyarlilik} = 2 * \frac{0.81 * 0.72}{0.81 + 0.72} = 2 * \frac{0.58}{1.53} = 0.75$

Performans Değerlendirme Teknikleri

❑ Tahmin (Regresyon) Modelleri İçin Performans Değerlendirme Teknikleri:

- ❑ Ortalama Hata (Mean Error, ME): Oluşturulan modelin öngördüğü, tahmin değerleri ile gerçek değerlerin arasındaki ortalama hatadır. Bu bağlamdaki hata, bir ölçümdeki belirsizlik veya tahmin değeri ile gerçek değer arasındaki farktır.

$$ME = \frac{1}{n} \sum_{j=1}^n e_j$$

$$e_j = A_j - \hat{A}_j$$

- A_j : Gerçek değer.
 - \hat{A}_j : Modelin öngördüğü tahmin değeri.
 - e_j : Modelin öngördüğü tahmin değerleri ile gerçek değerlerin arasındaki fark.
-

Performans Değerlendirme Teknikleri

❑ Tahmin (Regresyon) Modelleri İçin Performans Değerlendirme Teknikleri:

- ❑ Ortalama Yüzde Hata (Mean Percentage Error, MPE): Bir modelin tahmin ettiği değerler ile gerçek değerler arasındaki farkın ortalama yüzdesidir. MPE, daha çok birden fazla tahmin modelinin karşılaştırılmasında kullanılır. MPE değeri hesaplanırken tahmin hatalarının mutlak değerlerinden ziyade gerçek değerler kullanıldığı için pozitif ve negatif tahmin hataları birbirini dengeleyebilir. Bu ölçütün bir dezavantajı, tek bir gerçek değerın sıfır olması durumunda tanımlanamaz.

$$MPE = \frac{100}{n} \sum_{j=1}^n \frac{e_j}{A_j}$$

Performans Değerlendirme Teknikleri

❑ Tahmin (Regresyon) Modelleri İçin Performans Değerlendirme Teknikleri:

- ❑ Ortalama Mutlak Hata (Mean Absolute Error, MAE): İki sürekli değişken arasındaki farkın ölçüsüdür. MAE değeri, kolay yorumlanabilir olduğu için regresyon modellerinde sıkça kullanılmaktadır. MAE, yönlerini dikkate almadan bir dizi tahmindeki hataların ortalama büyüklüğünü ölçen, tüm tekil hataların ortalamada eşit olarak ağırlıklandırıldığı doğrusal bir skordur.

$$MAE = \frac{1}{n} \sum_{j=1}^n |e_j|$$

Performans Değerlendirme Teknikleri

❑ Tahmin (Regresyon) Modelleri İçin Performans Değerlendirme Teknikleri:

- ❑ Kök Ortalama Kare Hata (Root Mean Squared Error, RMSE): Bir modelin, tahminleyicinin tahmin ettiği değerler ile gerçek değerleri arasındaki uzaklığın bulunmasında sıklıkla kullanılan, hatanın büyüklüğünü ölçen kuadratik bir metriktir. RMSE değeri, 0'dan ∞ 'a kadar değerlik alabilir. RMSE değerinin sıfır olması, modelin hiç hata yapmadığı anlamına gelir. RMSE, büyük hataları daha fazla cezalandırmanın avantajına sahiptir. Bu yüzden bazı durumlara daha uygun olabilir.

$$RMSE = \sqrt{\frac{\sum_{j=1}^n e_j^2}{n}}$$

$$RMSE = \sqrt{MSE}$$

Performans Değerlendirme Teknikleri

❑ Tahmin (Regresyon) Modelleri İçin Performans Değerlendirme Teknikleri:

- ❑ Ortalama Mutlak Yüzde Hata (Mean Absolute Percentage Error, MAPE): Regresyon modellerinde tahminlerin doğruluğunu ölçmek için sıklıkla kullanılır. Gerçek değerler arasında sıfır içerenler varsa, sıfır ile bölünme olacağı için MAPE hesaplanamaz.

$$MAPE = \frac{100}{n} \sum_{j=1}^n \frac{|e_j|}{|A_j|}$$

Performans Değerlendirme Teknikleri

❑ Örnek:

A_j (Actual)	\hat{A}_j (Prediction)	$e_j = A_j - \hat{A}_j$	e_j / A_j	$ e_j $	$[e_j]^2$	$(e_j) / (A_j)$
4	6	-2	-0.50	2	4	0.50
8	7	1	0.13	1	1	0.13
15	15	0	0.00	0	0	0.00
16	15	1	0.06	1	1	0.06
23	18	5	0.22	5	25	0.22
46	34	12	0.26	12	144	0.26

❑ Ortalama Hata?

$$ME = \frac{1}{n} \sum_{j=1}^n e_j = 2.83$$

❑ Ortalama Yüzde Hata?

$$MPE = \frac{100}{n} \sum_{j=1}^n \frac{e_j}{A_j} = 2.76$$

❑ Ortalama Mutlak Hata?

$$MAE = \frac{1}{n} \sum_{j=1}^n |e_j| = 3.5$$

❑ Kök Ortalama Kare Hata ?

$$RMSE = \sqrt{\frac{\sum_{j=1}^n e_j^2}{n}} = 5.4$$

❑ Ortalama Mutlak Yüzde Hata

$$MAPE = \frac{100}{n} \sum_{j=1}^n \frac{|e_j|}{|A_j|} = 19.43$$



Hafta_3_regresyo
n_metrik_1

Performans Değerlendirme Teknikleri

❑ Örnek:

$A_j(\text{Actual})$	$\hat{A}_j(\text{Prediction})$	$e_j = A_j - \hat{A}_j$	e_j / A_j	$ e_j $	$[e_j]^2$	$(e_j)/(A_j)$
5	10	-5	-1.00	5	25	1.00
15	15	0	0.00	0	0	0.00
7	11	-4	-0.57	4	16	0.57
11	14	-3	-0.27	3	9	0.27
23	25	-2	-0.09	2	4	0.09
28	26	2	0.07	2	4	0.07

❑ Ortalama Hata?

$$ME = \frac{1}{n} \sum_{j=1}^n e_j = -2.0$$

❑ Ortalama Yüzde Hata?

$$MPE = \frac{100}{n} \sum_{j=1}^n \frac{e_j}{A_j} = -30.99$$

❑ Ortalama Mutlak Hata?

$$MAE = \frac{1}{n} \sum_{j=1}^n |e_j| = 2.67$$

❑ Kök Ortalama Kare Hata ?

$$RMSE = \sqrt{\frac{\sum_{j=1}^n e_j^2}{n}} = 3.11$$

❑ Ortalama Mutlak Yüzde Hata

$$MAPE = \frac{100}{n} \sum_{j=1}^n \frac{|e_j|}{|A_j|} = 33.38$$



Hafta_3_regresyon_metrik_2

Kaynakça

- ❑ Bonaccorso, G. (2018). Machine Learning Algorithms: Popular algorithms for data science and machine learning. Packt Publishing Ltd.
- ❑ Yılmaz, A. (2021). Yapay Zekâ. Kodlab Yayın Dağıtım Yazılım Ltd. Şti.
- ❑ Çelik, Ş., Köleoğlu, N., & Çemrek, F. (2022). Veri Madenciliği ve Makine Öğrenmesi İle Farklı Alanlarda Uygulamalar. Holistence Publications.