VERİ MADENCİLİĞİ

Fırat İsmailoğlu, PhD

Sınıflandırıcıların Performanslarının Ölçülmesi

Kesinlik (Accuracy)

Bir sınıflandırıcını performansı genel olarak bir test setindeki elemanların yüzde kaçını doğru olarak sınıfladığı ölçülerek verilir. Bu ölçüye kesinlik (accuracy) denir.

ör.

	Gerçek Sınıf	**Tahmin Edilen Sınıf**	**Doğru Tahmin**
Test seti örnekleri	Kanser	Kanser	✓
	Kanser Değil	Kanser Değil	✓
	Kanser Değil	Kanser Değil	✓
	Kanser Değil	Kanser Değil	✓
	Kanser	Kanser Değil	✗
	Kanser	Kanser	✓
	Kanser Değil	Kanser	✗
	Kanser Değil	Kanser Değil	✓

$6 \rightarrow$ Doğru tahmin sayısı

$8 \rightarrow$ Toplam tahmin sayısı

$$\text{Kesinlik} = \frac{6}{8} \times 100 = 75$$

Aşırı Uyum (Overfitting)

Sınıflandırıcıların yaptığı iki tür hata vardır:

i. Eğitim Seti Hatası

ii. Test Set Hatası (Genelleştirme Hatası)

Eğitim seti hatası: sınıflandırıcın eğitim setinde yanliş sınıflandırdığı elemanların yüzdesidir.

Örneğin bir sınıfandıcı eğitildiği eğtim setindeki 50 örneğin 10'unu yanlış sınıflandırsın. Bu durumda bu sınıflandıcının eğitim seti hatası $\%20$ olur.

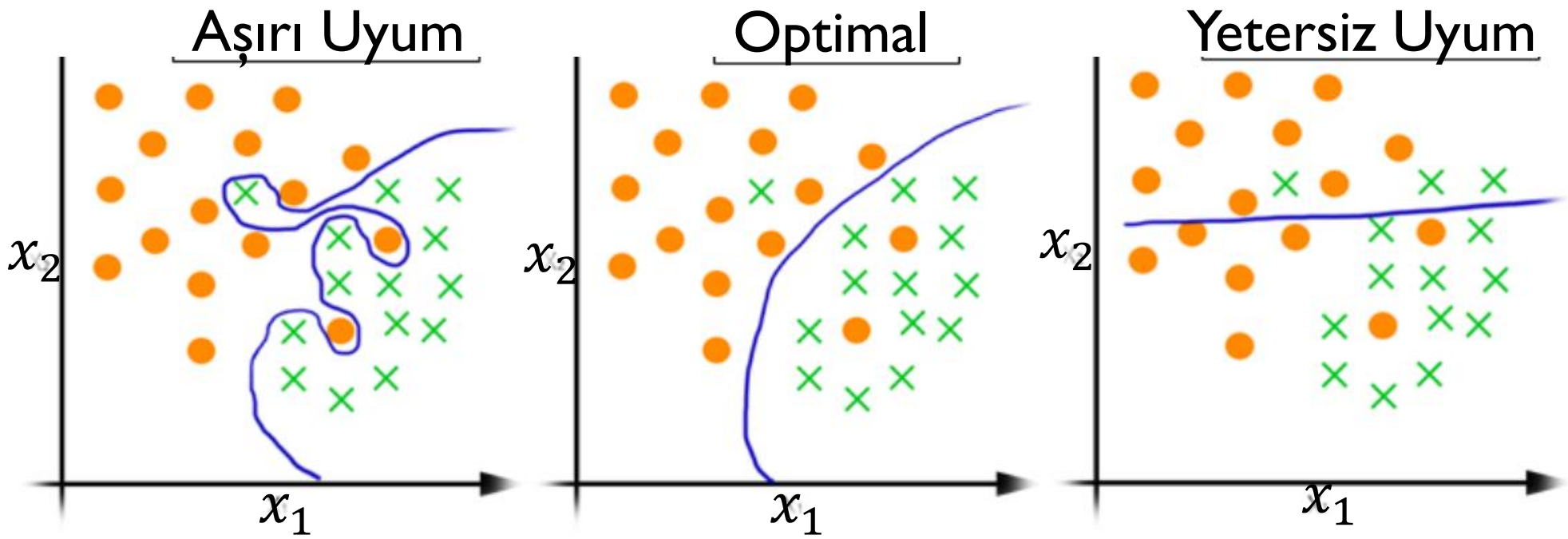
Test seti hatası: sınıflandırıcın test setinde yanlış sınıflandırdığı elemanların yüzdesidir.

Veri madenciliğinde esas olarak test seti hatası ile ilgilenilir. Çünkü bu, sınıflandırıcının ileride karşilacağı örnekleri ne kadar iyi sınıflandırıcağının bir göstergesidir. Yani sınıflandıcının ne kadar iyi genelleme yapabildiğinin bir göstergesidir.

Aşırı Uyum (Overfitting)

Eğer bir sınıflandırıcı eğitim setindeki örnekleri iyi sınıflandırıyor (eğitim seti hatası düşük); test setindeki örnekleri kötü sınıflandırıyor (test seti hatası yüksek) ise, bu durumda sınıflandırıcıda aşırı uyum (overfitting) vardır denir.

Aşırı uyumun zıttı yetersiz uyumdur (underfitting) (yüksek eğitim seti hatası ve yuksek test seti hatasi)



Aşırı Uyum Görülme Nedenleri

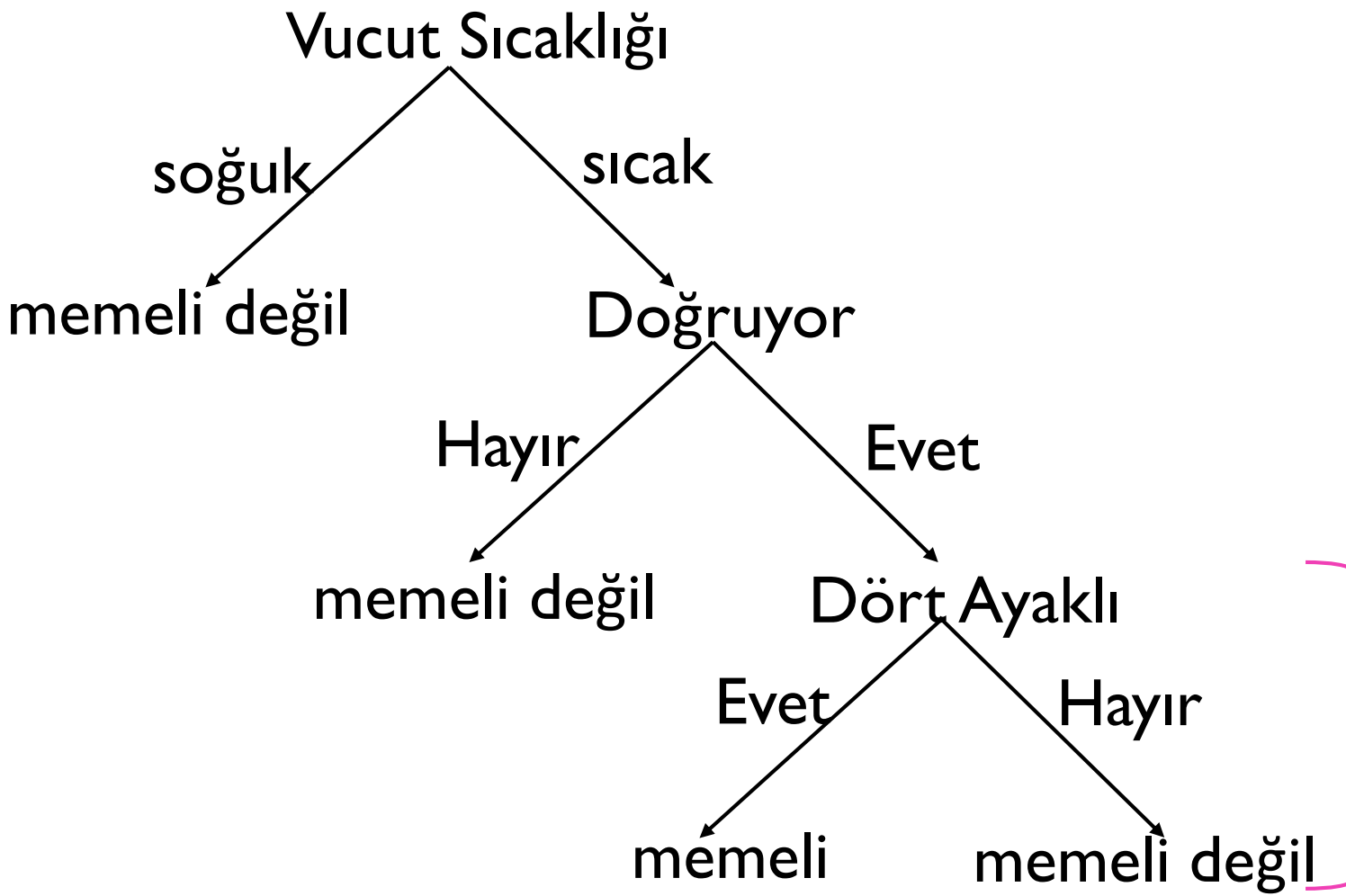
1. Yanlış Veri (Gürültü)

Veri setinde bazı sınıflar yanlış girilmiş olabilir. Bu durumda bu setten öğrenilen sınıflandırıcı test örneklerini doğru olarak sınıflandırmayabilir.

İsim	Vucut Sıcaklığı	Doğuruyor	Dört Ayaklı	Kış Uykusu	Sınıf (Memeli)
Kirpi	Sıcak	Evet	Evet	Evet	Evet
Kedi	Sıcak	Evet	Evet	Hayır	Evet
Yarasa	Sıcak	Evet	Hayır	Evet	Hayır
Balina	Sıcak	Evet	Hayır	Hayır	Hayır
Semender	Soğuk	Hayır	Evet	Evet	Hayır
K. Ejder	Soğuk	Hayır	Evet	Hayır	Hayır
Piton	Soğuk	Hayır	Hayır	Evet	Hayır
Somon	Soğuk	Hayır	Hayır	Hayır	Hayır
Kartal	Sıcak	Hayır	Hayır	Hayır	Hayır
Lepistes	Soğuk	Evet	Hayır	Hayır	Hayır

Yanlış Sınıflandırılmış

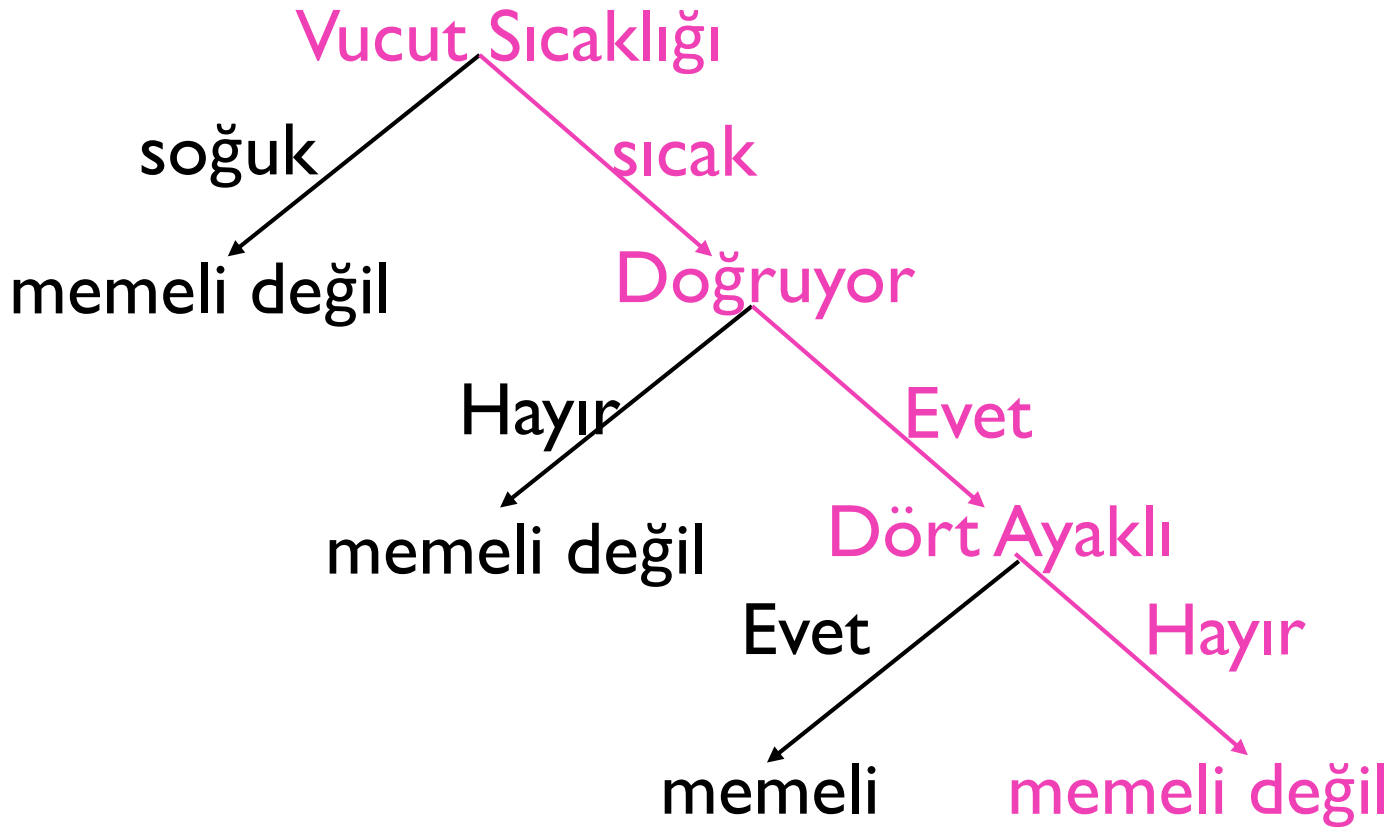
Bu veri setinde yarasa ve yunus gerçekte memeli olmasına rağmen memeli değil olarak girilmiş. Bu '<u>hatalı</u>' veri setinden elde edilen karar ağacı aşağıdaki gibi olur:



Karar ağacının bu parçası yanlış girilen sınıflar yüzünden oluşmuştur!

İsim	Vucut Sıcaklığı	Doğuruyor	Dört Ayaklı	Kış Uykusu	Sınıf (Memeli)
İnsan	Sıcak	Evet	Hayır	Hayır	Evet
Yunus	Sıcak	Evet	Hayır	Evet	Evet

Test seti örnekleri



Memeli olan İnsan ve Yunus karar ağacının hatalı parçasından dolayı yanlış (memeli değil) olarak sınıflandırılır!

Aşırı Uyum Görülme Nedenleri

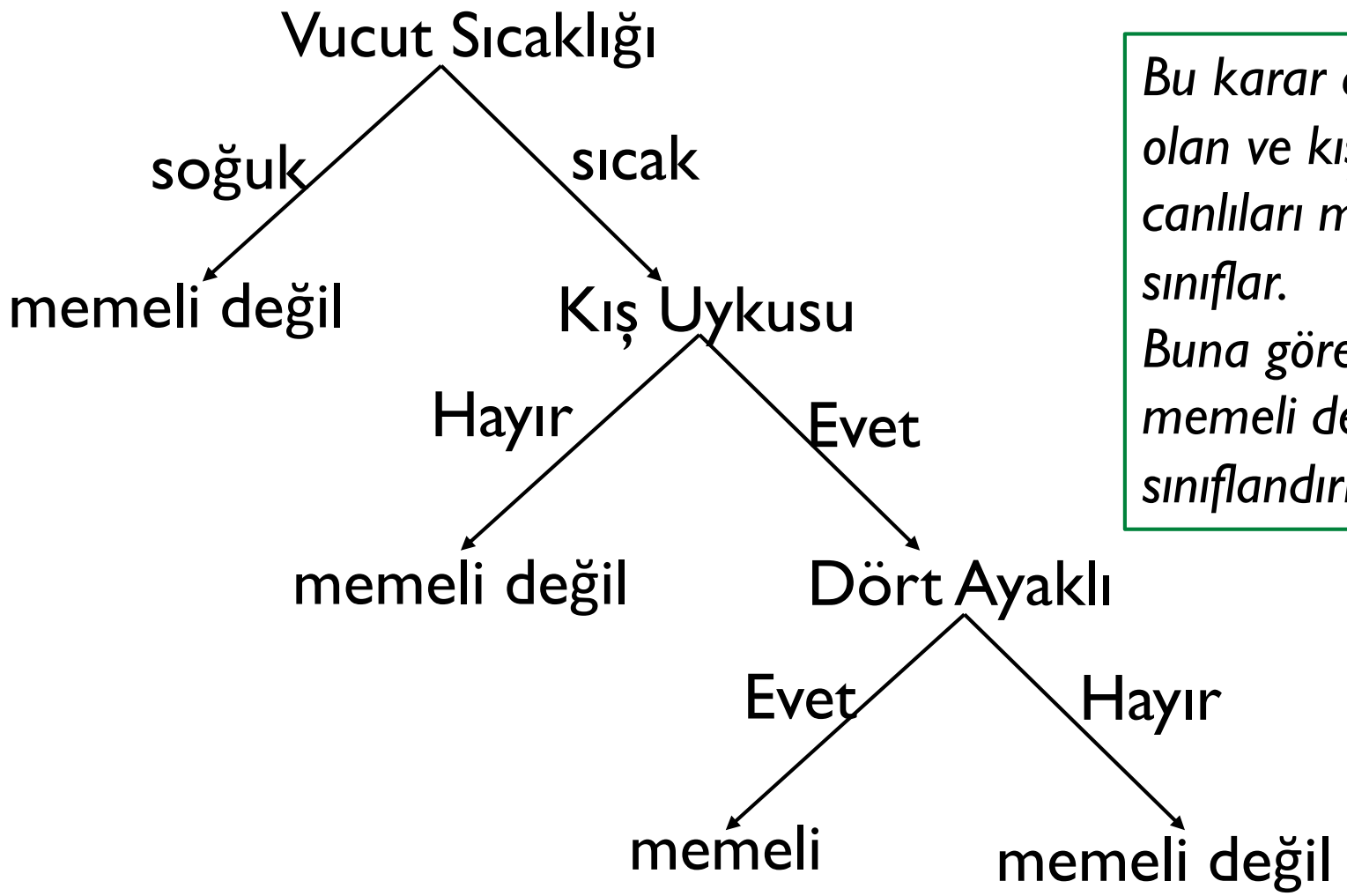
2. Yeterince Veri Olmaması

Eğer verilen eğitim seti küçükse, bu küçük setten öğrenilen sınıflandırıcıda aşırı uyum görülmesi muhtemeldir.

Örnek olarak varsayalımki eğitim setimiz yalnızca aşağıdaki 5 örnekten oluşşsun.

İsim	Vucut Sıcaklığı	Doğuruyor	Dört Ayaklı	Kış Uykusu	Sınıf (Memeli)
Semender	Soğuk	Hayır	Evet	Evet	Hayır
Lepistes	Soğuk	Evet	Hayır	Hayır	Hayır
Kartal	Sıcak	Hayır	Hayır	Hayır	Hayır
Gece Kuşu	Sıcak	Hayır	Hayır	Evet	Hayır
Ornitorenk	Sıcak	Hayır	Evet	Evet	Evet

Bu 5 örnek kullanilarak elde edilen karar ağacı şöyledir:



Bu karar ağacı vucut sıcaklığı sıcak olan ve kış uykusuna yatmayan canlıları memeli değil olarak sınıflar.
Buna göre insanlar ve yunuslar memeli değil olarak sınıflandırılırlar; bu yanlıştır.

Karışıklık Matrisi (Confusion Matrix)

Sınıflandırıcımızın performansını ölçerken yalnızca doğru olarak sınıfladığı örneklerin yüzdesini hesaplamak çoğu kez yeterli değildir.

Örnek olarak varsayalımki test setimizde 95 kişi kanser değil ve 5 kişi de kanserli olsun. Eğer bu durumda sınıflandırıcımız tahmin ettiği 100 kişinin tamamına kanser değil derse kesinliği $\%95$ olur. Bu, ilk bakışta sınıflandıcımızın başarılı olduğu anlamına gelebilir.

Fakat bu önemsiz (trivial) bir başarıdır. Gerçekten bulmak istediğimiz 5 kanserli hastanın tamamı kanserli değil olarak sınıflandırılmıştır ve bu hastalar tespit edilememiştir.

Test setinde sınıfların birbirlerine gore sayilari orantisiz ise (bir yada bir kac sinif baskınsa) bu durumda genel olarak sınıflandırıcının kesinligine degil karışıklık matrisindeki performansına bakarız.

Karışıklık Matrisi (Confusion Matrix)

Karışıklık matirisi, satır toplamları sınıfların gerçek sayıları; sütün toplamları sınıfların tahmin edilen sayıları olan matristir.

		Tahmin Edilen Sınıflar: Pozitif	Tahmin Edilen Sınıflar: Negatif
Gerçek Sınıflar	Pozitif	Gerçek Pozitif (GP)	Yanlış Negatif (YN)
	Negatif	Yanlış Pozitif (YP)	Gerçek Negatif (GN)

Gerçek Pozitif (True Positive): Gerçekte pozitif olan ve sınıflandırıcı tarafından da pozitif olarak tahmin edilen test örneklerinin sayısı.

Gerçek Negatif (True Negative):Gerçekte negatif olan ve sınıflandırıcı tarafından da negatif olarak tahmin edilen test örneklerinin sayısı.

Karışıklık Matrisi (Confusion Matrix)

Yanlış Negatif (False Negative) Gerçekte pozitif olan fakat sınıflandırıcı tarafından negatif olarak tahmin edilen test örneklerinin sayısı (yani yanlışlıkla negatif denmis aslında pozitif olan örnekler) .

Yanlış Pozitif (False Positive) Gerçekte negatif olan fakat sınıflandırıcı tarafından postive olarak tahmin edilen test örneklerinin sayısı (yani yanlışlıkla pozitif denmis aslında negatif olan örnekler)

		Tahmin Edilen Sınıflar		
		Pozitif	Negatif	Toplam
Gerçek Sınıflar	Pozitif	GP	YN	Toplam Pozitif
	Negatif	YP	GN	Toplam Negatif

Karışıklık Matrisinden Kesinlik Elde Edilmesi

$$Kesinlik = \frac{GP + GN}{Toplam\ Pozitif + Toplam\ Negatif}$$

$GP + GN$ → Doğru olarak tahmin edilmiş toplam test örneği sayısı

$Toplam\ Pozitif + Toplam\ Negatif$ → Test seti eleman sayısı

Gerçek Pozitif Oranı (TP Rate – Sensivity)

$$GP\ Oran = \frac{GP}{GP + YN} = \frac{GP}{Toplam\ Pozitif}$$

GP → Doğru olarak tahmin edilmiş pozitif örnekler sayısı

$Toplam\ Pozitif$ → Test setindeki pozitif örnekler sayısı

Gerçek pozitif oranı, pozitif örnekler icinde doğru olarak sınıflanan örneklerin oranıdır.

Bir baska ifadeyle sınıflandırıcının test setindeki pozitif örnekleri yakalama gücüdür.

Gerçek pozitif oranın bir başka adı hassaslık (sensivity) dır.

Gerçek Negatif Oranı (TN Rate – Specificity)

$$GN\ Oran = \frac{GN}{GN + YP} = \frac{GN}{Toplam\ Negatif}$$

GN → Doğru olarak tahmin edilmiş negatif örnekler sayısı

$Toplam\ Negatif$ → Test setindeki negatif örnekler sayısı

Gerçek negatif oranı, negatif örnekler icinde doğru olarak sınıflanan örneklerin oranıdır.

Yani sınıflandırıcının test setindeki negatif örnekleri yakalama gücüdür.

Gerçek negatif oranın bir başka adı belirginlik (specificity) dır.

ör.

		Tahmin Edilen Sınıflar		
		Pozitif	Negatif	Toplam
Gerçek Sınıflar	Pozitif	87	14	101
	Negatif	4	53	57

GP=87 örnek gerçekte pozitifmiş ve sınıflandırıcı tarafından da pozitif olarak sınıflandırılmış.

YN=14 örnek gerçekte pozitifmiş fakat sınıflandırıcı tarafından negatif olarak sınıflandırılmış.

YP=4 örnek gerçekte negatifmiş fakat sınıflandırıcı tarafından pozitif olarak sınıflandırılmış.

GN=53 örnek gerçekte negatifmiş ve sınıflandırıcı tarafından da negatif olarak sınıflandırılmış.

$$Kesinlik = \frac{87+53}{101+57} = 0.88 \qquad GP\ Oran = \frac{87}{101} = 0.86 \qquad GN\ Oran = \frac{53}{57} = 0.93$$

Ne Zaman Gerçek Pozitif Oran - Ne Zaman Gerçek Negatif Oran Önemlidir?

Diyelimki elimizde bir kanser sınıflandırma veri seti var. Bu veri seti için gerçekte pozitif (kanser) olan kisileri negatif (kanser değil) olarak sınıflandırmayı hiç istemeyiz; kanser olan kişilerin tamamını yakalamak isteriz.

YN: Gerçekte kanser olup bizim kanser degil dedigimiz kişi sayısı (düşük olmalı)

GP: Gerçekte kanser olup bizim de kanser dedigimiz kişi sayısı (yüksek olmalı)

Bu durumda bizim için Gerçek Pozitif Oran $\frac{GP}{GP+YN}$ yüksek olmalıdır.

Soru: Bu veri seti için YP ve GN nin ne anlama geldiklerini yazınız.

YP:

GN:

Ne Zaman Gerçek Pozitif Oran - Ne Zaman Gerçek Negatif Oran Önemlidir?

Diyelimki elimizde bir spam mail sınıflandırma veri seti var. Burada pozitif sınıf mailin spam olduğunu (istenmeyen mail olduğunu) negatif sınıfsa mailin ham olduğunu (istenilen mail olduğunu) gösteriyor.

Bu veri seti için gerçekte negatif (ham) mailleri pozitif (spam) olarak sınıflandırmayı hiç istemeyiz.

YP (yanlışlıkla spam olarak işaretlenmiş mail sayısı) değerinin olabildiğince düşük olmasını isteriz.

Öte yandan GN (ham olarak sınıflandırılmış ham mailler sayısı) olabildiğince yüksek olmasını isteriz.

Bu durumda bizim için Gerçek Negtif Oran $\frac{GN}{GN+YP}$ yüksek olmalıdır.

K – Katlı Çapraz Doğrulama (K –Fold Cross Validation)

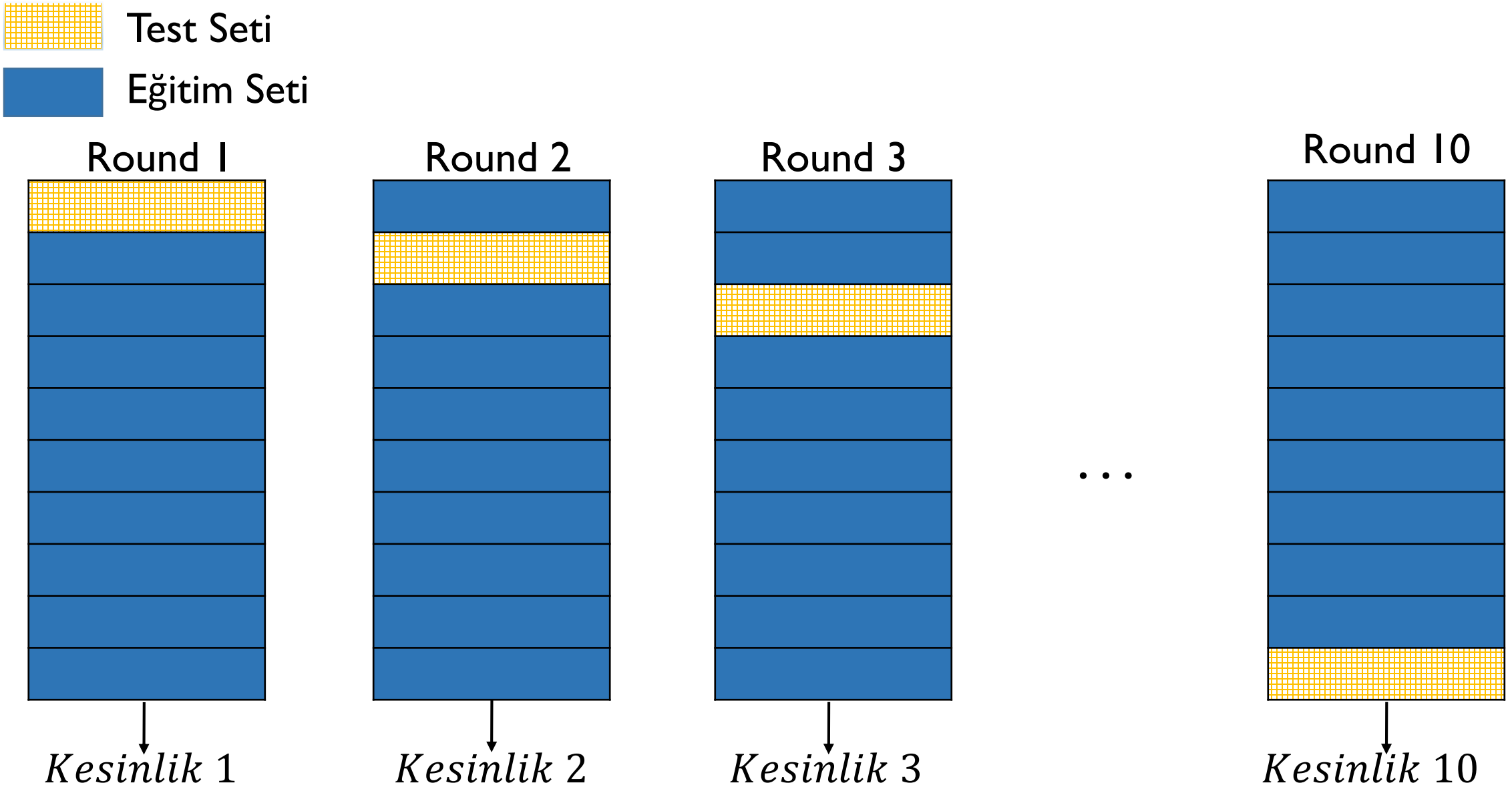
Bir sınıflandırıcı inşa ederken verilen veri setini $\%70 - \%30$ oranında böler; $\%70'$ni sınıflandırma algoritmasını eğitmek için kullanır, $\%30'$nu sınıflandırıcıyı test etmek için kullanırız demiştik.

Fakat bazen veri seti yeterince büyük olmaz. Bu verinin $\%70$ 'ini kullanarak iyi bir sınıflandırıcı elde edemeyiz. Bu durumda '*k*-katlı çapraz doğrulama' yöntemini kullanırız.

K – Katlı Çapraz Doğrulama Yöntemi

1. Genel olarak *k* 10 alınır.
2. Veri seti rastgele 10 parçaya bölünür (enine).
3. Oluşan her bir parça için:

   Geriye kalan 9 parça eğtim seti olur. Bu eğitim setinden bir sınıflandırıcı elde edilerek bu parçadaki elemanlar sınıflandırılır. Böylece buradan bir kesinlik elde edilir.
4. Elde edilen 10 kesinlik değerinin ortalaması alinarak sonuç kesinlik değeri elde edilir.

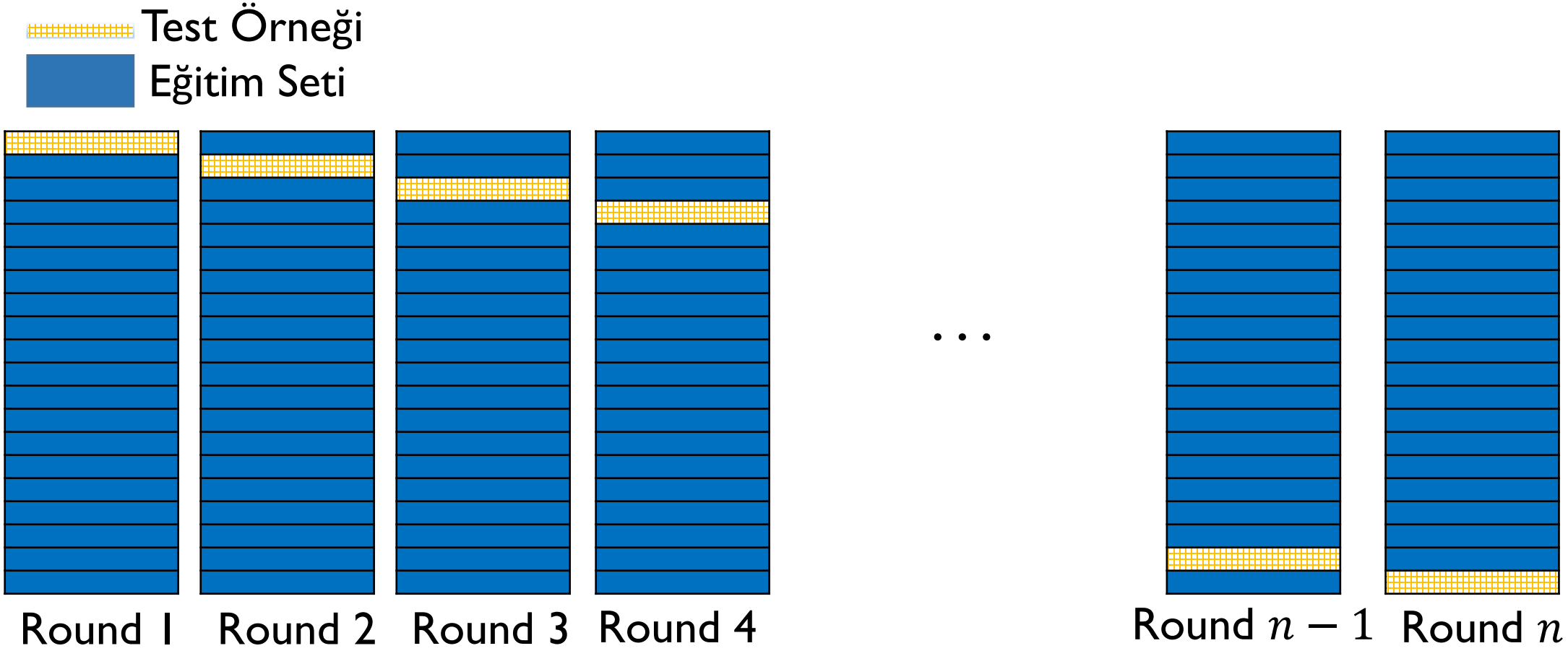


$$Final\ Kesinlik = \frac{1}{10}(Kesinlik1 + Kesinlik\ 2 + Kesinlik\ 3 + \cdots + Kesinlik\ 10)$$

Birini Dışarıda Bırakma (Leave One Out)

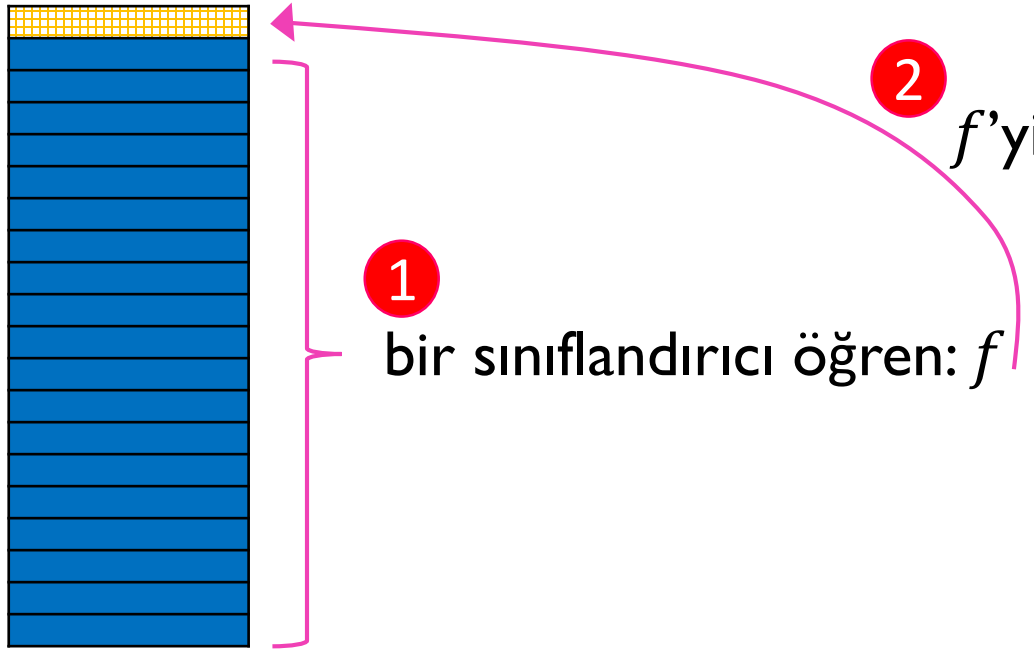
K-katlı çapraz doğrulama yönteminde *K*'yı maksimum, veri setindeki örnek sayısı kadar alabiliriz. Bu şekilde her round'da test seti yalnızca bir örnekten oluşur, diğer tüm örnekler eğitim setini oluşturur.

Bu yönteme 'birini dışarıda bırakma' yöntemi denir.



Birini Dışarıda Bırakma (Leave One Out)

Her bir roundda



3 Eğer test örneğini doğru sınıflandırsan, doğru sınıflandırdığın örnek sayısını 1 artır.