Basitçe Accuracy, Precision, Recall ve F1 Score

Preprint · September 2021

1 author:

Kahraman Kostas
Heriot-Watt University
6 PUBLICATIONS **15** CITATIONS

SEE PROFILE

Basitçe Accuracy, Precision, Recall ve F1 Score

Kahraman Kostas

Özet

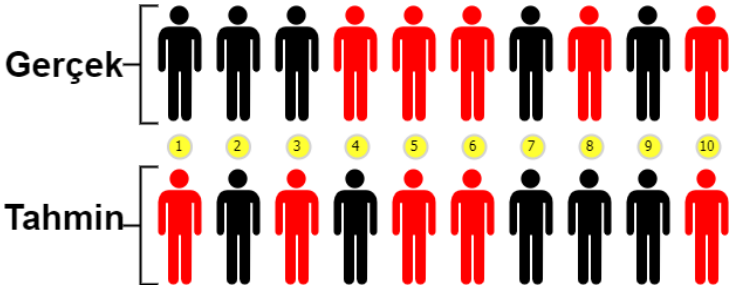
Bir sınıflandırma probleminin/işleminin ne kadar başarılı olduğunu anlamak için çeşitli değerlendirme kriterleriyle sonuçlarımızı analiz etmemiz gerekir. Bu yazıda en basit şekilde bu kriterler nelerdir ve nasıl kullanılırlar sorularına ışık tutacağız. Bu bağlamda Accuracy, Precision, Recall ve F1 Score kavramlarını oldukça basitçe açıklayacağız.

Anahtar Kelimeler: Accuracy · Doğruluk · Precision · Kesinlik · Recall · Duyarlılık · F1 Score

Örnek Olay

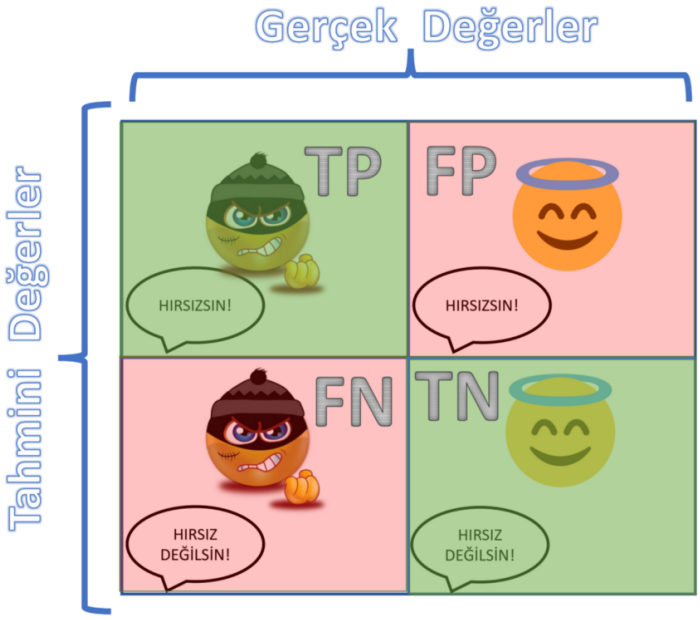
Her şeyden önce problemimizi belirleyelim/yaratalım: Hayali bir markete güvenlik şefi olarak alınmak üzere başvurduğunuzu düşünün. Marketin sahibi size, mahallede hırsızlık oranlarının çok yüksek olduğunu bunun için sezgileri kuvvetli birini aradıklarını söyleyerek bir uygulama yapmanızı istiyor. Bu işlem sonucunda sizin ne kadar başarılı olduğunuz gözlemlenecek. Uygulama şöyle: Marketten çıkan 10 müşterinin hırsız olup olmadığını tahmin edeceksiniz. Bu "Tahmin verileri"ni oluşturacak. Market çıkışında bu 10 müşterinin hepsinin üstü aranacak ve hırsız olup olmadıklarına bakılarak "Gerçek veriler" oluşturulacak. Daha sonra bu iki veri kıyaslanarak sizin ne kadar başarılı olduğunuz analiz edilecek.

Şekil 1'de gerçek ve tahmin verilerini görüyorsunuz. Kırmız renk hırsızı, siyah ise normal müşteriyi gösteriyor. Şekil 1 incelenirse bazı tahminlerinizde başarılı bazılarında ise başarısız olduğunuzu görürsünüz. Bu tahminleri 4 grup altında toplayabiliriz [1]. Bunlar :

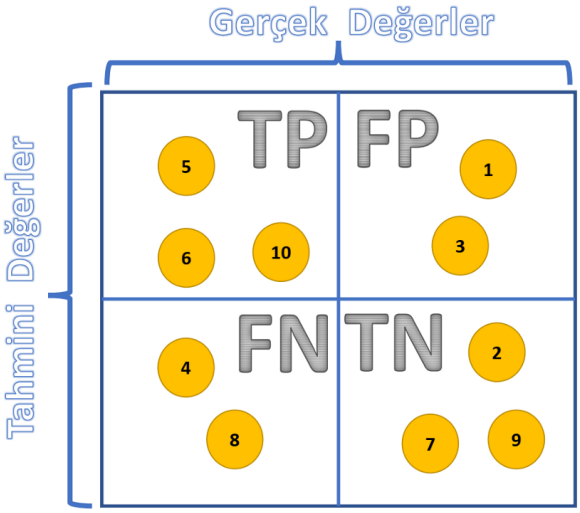


Şekil 1: Tahmin ve Gerçek verilerin görselleştirilmesi. Kırmızı: hırsız, siyah: normal müşteri.

- **TP** (True Positive): Bir kişiye hırsız dediniz ve o kişi hırsız.
- **TN** (True Negative): Bir kişiye hırsız değil dediniz ve o kişi hırsız değil.
- **FP** (False Positive): Bir kişiye hırsız dediniz ve o kişi hırsız değil.
- **FN** (False Negative): Bir kişiye hırsız değil dediniz ve o kişi hırsız.



Şekil 2: TP, TN, FP ve FN kavramlarının görselleştirilmesi.



Şekil 3: Tahminlerimizin confusion matrix üzerindeki dağılımı.

Bu kavramlar Şekil 2'de görselleştirilmiştir. Bu dört grubu, bir arada gösteren tabloya confusion matrix (Hata Matrisi) [1] diyoruz. Tahminlerimizin confusion matrix üzerindeki dağılımını Şekil 3'de görebilirsiniz. Şekil ürerindeki sarı daireler tahmin edilen müşterilerin numaralarını gösterir. Değerlendirme yöntemleri arasında en yaygın olarak kullanılan yöntem, **Accuracy (Doğruluk)**' tur.

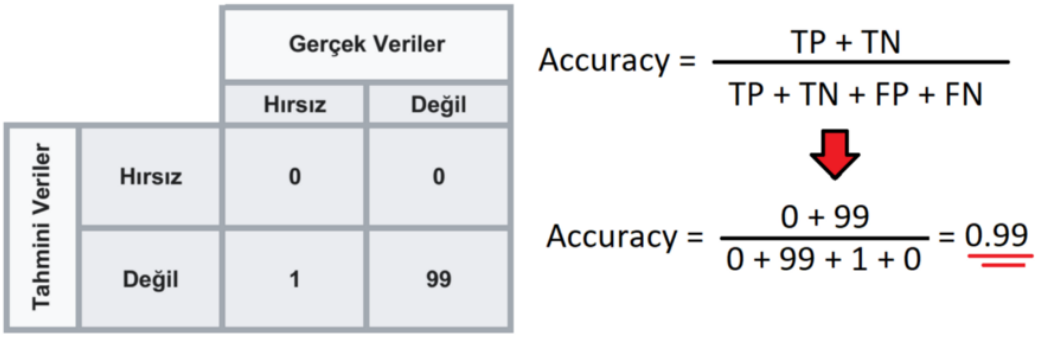
Accuracy (Doğruluk)

Accuracy (Doğruluk) [2] tüm doğru cevaplarınızın (TP ve TN), tüm cevaplarınıza (TP, TN, FP, FN) oranı olarak açıklanabilir (bakınız Denklem 1).

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{1}$$

Sık kullanılmasına rağmen accuracy'nin bir dezavantajı vardır: Dengesiz dağılıma sahip gruplarda sağlıklı sonuç vermez. Şöyle bir örnekle açıklayalım: yine markette hırsızlığı tespit etmeye çalıştığınızı farzedelim. İnsanların pek azı hırsızdır. Diyelim ki çalınan ürünleri tespit eden bir alarm sisteminiz var ve ne yazık ki bozuk. 100 kişilik müşterileriniz içerisinde yalnızca bir hırsız olsun. Siteminiz bozuk olduğu için bu kişiyi tespit edemediniz ancak günün sonunda başarınızı accuracy ile ölçerseniz %99 oranında başarılı olduğunuz

görülür (Şekil 4'i inceleyin). Aslında hiç fena değil! Sisteminiz tamamen arızalı iken bile %99 başarı oranı sağlıyor ki bu tamamen yanıltıcı ve hiç istediğimiz bir şey değil. İşte tam olarak bu yüzden accuracy dışında da çeşitli değerlendirme yöntemleri kullanılmaktadır.



Şekil 4: Örnek için yaratılan confusion matrix ve Accuracy'nin hesaplaması.

Recall ve Precision

Recall (Duyarlılık/Hassasiyet) [3]: Doğru tespit ettiğimiz Pozitif sınıfların (TP, doğru tahmin ettiğimiz hırsızlar), Tüm pozitiflere oranı (bizim doğru tahmin etmemizden bağımsız olarak gerçekten hırsız olanlar, yani TP+FN). Şekil 4'te ilk hücrenin ilk sütuna bölümü Recall değerini verir (Denklem 3).

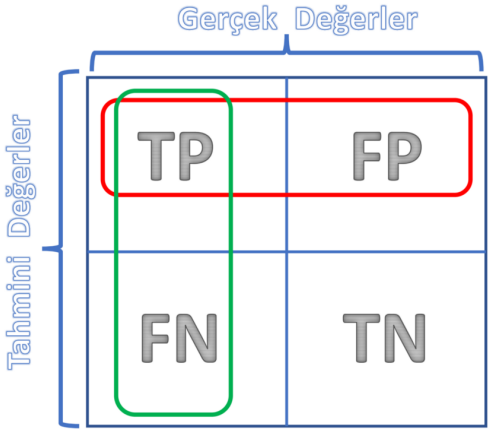
$$Recall = \frac{TP}{TP + FN} \tag{2}$$

Precision (Kesinlik) [3] : doğru tespit ettiğimiz Pozitif sınıfların (TP, doğru tahmin ettiğimiz hırsızlar) tüm hırsız diye etiketlediğimiz/adlandırdığımız verilere oranıdır (TP+FP). Başka bir değişle bildiğimiz hırsızların sayısının, bildiğimiz hırsızlar ve yanlış alarmların toplamına oranıdır. Şekil 4'te ilk hücrenin ilk satıra bölümü Precision değerini verir (bakınız Denklem 3).

$$Precision = \frac{TP}{TP + FP} \tag{3}$$

Burada precision/ recall arasındaki dengeden de bahsetmek gerekir. Yaptığınız işe göre precision/ recall arasında tercih yapmanız gerekebilir. Ek örnekle açıklayalım: Market örneğinde, bazen hırsızların çaldıkları şeyleri çok

iyi sarıp sarmaladıklarında, cihazınızı kandırabildiklerini düşünelim. Marketteki alarmınızın bir hassasiyet ayarı olsun, eğer bu ayarı yükseltirseniz, sarıp sarmalanan ürünleri da yakalıyor (gerçek hırsızlar yakalanıyor / TP artıyor). Ancak, bunun bir dezavantajı var, bu alarm bazen üzerinde çeşitli metal eşya taşıyan normal müşteriler için de ötüyor (yanlış alarmlar — FP yükseliyor). Yaptığımız ayar neticesinde, TP oranını yükselttik, FN sayısında bir değişiklik olmadı bu yüzden Recall değerimiz yükselmeye başladı. Diğer taraftan yaptığımız bu hassas ayar yanlış alarm sayısını yükselttiği için Precision değerimiz düşmeye başladı.



Şekil 5: Örnek confusion matrix.

Başka bir deyişle sisteminizin hassasiyetini (Recall) arttırmanız kesinlik (Precision) değerinizin düşmesine neden olabilir. Burada önemli olan şey sisteminizin öncelikleridir. Önceliklerinizi analiz ederek bu dengeyi iyi ayarlamanız gerekir. Örneğin hava alanlarında abratılı bir güvenlik protokolü uygulanır. Bu yüzden her yolcu yüksek hassasiyetle, suçluymuşcasına aranır. Bu işlem Recall değerini en iyilerken, Precision değerini oldukça düşürür. Bunun nedeni uçuş güvenliğinin öne çıkmasıdır. Bu işlem sırasında yüksek yanlış alarm/FP çıkması göze alınır ve yüksek risk engellenir. Alışveriş merkezlerinde ise daha düşük hassasiyetli bir güvenlik politası öne çıkar böylece güvenlik ve müşteri memmuniyeti dengesi gözetilir.

Accuracy'ye alternatif olabilecek bir başka değerlendirme yöntemi ise F1 Score'dur. Dengesiz dağılıma sahip veriler için F1 Score'un kullanımı daha doğru olacaktır.

F1 Score

F1 Score [4], Precision ve Recall'un harmonik ortalamasıdır (Denklem 4). Harmonik ortalamanın normal ortalamadan farkı, güçsüz tarafın yanında olmasıdır. F1 Score'u, yüksek değeri cezalandırır, böylece bu iki değerden çok yüksek olanın düşük olanı manipüle etmesinin önüne geçer. Örneğin Şekil 4 verilen örnekte accuracy %99 çıkarken F1 Scoru 0 olacak ve bu çıktı sisteminizde yolunda gitmeyen bir şeyler olduğu konusunda size ipucu verecektir. Harmonik ortalama ve normal ortalamanın farkını sezmek için Şekil 6'de verilen değerleri inceleyebilirsiniz.

$$F1\ Score = \frac{Recall^{-1} + Precision^{-1}}{2} = 2 * \frac{Recall * Precision}{Recall + Precision} \qquad (4)$$

Sayı 1	Sayı 2	Ortalama	Harmonik Ortalama
1	100	50.50	1.98
2	99	50.50	3.92
3	98	50.50	5.82
4	97	50.50	7.68
5	96	50.50	9.50
6	95	50.50	11.29
7	94	50.50	13.03
8	93	50.50	14.73
9	92	50.50	16.40
10	91	50.50	18.02

Sayı 1	Sayı 2	Ortalama	Harmonik Ortalama
45	56	50.50	49.90
46	55	50.50	50.10
47	54	50.50	50.26
48	53	50.50	50.38
49	52	50.50	50.46
50	51	50.50	50.50
51	50	50.50	50.50
52	49	50.50	50.46
53	48	50.50	50.38
54	47	50.50	50.26

Sayı 1	Sayı 2	Ortalama	Harmonik Ortalama
91	10	50.50	18.02
92	9	50.50	16.40
93	8	50.50	14.73
94	7	50.50	13.03
95	6	50.50	11.29
96	5	50.50	9.50
97	4	50.50	7.68
98	3	50.50	5.82
99	2	50.50	3.92
100	1	50.50	1.98

Şekil 6: harmonik ortalama vs normal ortalama.

References

[1] Wikipedia, "Confusion matrix — Wikipedia, the free encyclopedia," http://en.wikipedia.org/w/index.php?title=Confusion%20matrix&oldid=1088688630, 2022, [Online; accessed 27-July-2022].

[2] ——, "Accuracy and precision — Wikipedia, the free encyclopedia," http://en.wikipedia.org/w/index.php?title=Accuracy%20and%20precision&oldid=1089401220, 2022, [Online; accessed 27-July-2022].

[3] ——, "Precision and recall — Wikipedia, the free encyclopedia," http://en.wikipedia.org/w/index.php?title=Precision%20and%20recall&oldid=1100737023, 2022, [Online; accessed 27-July-2022].

[4] ——, "F-score — Wikipedia, the free encyclopedia," http://en.wikipedia.org/w/index.php?title=F-score&oldid=1099088638, 2022, [Online; accessed 27-July-2022].