Türkiye Klinikleri Biyoistatistik Dergisi
Turkiye Klinikleri Journal of Biostatistics Turkiye Klinikleri J Biostat. 2024;16(2):

ORİJİNAL ARAŞTIRMA ORIGINAL RESEARCH DOI: 10.5336/biostatic.2024-103854

Örneklem Büyüklüğünün Tahmin-Doğrulama Metrikleri Üzerindeki Etkisinin İncelenmesi: Bir Simülasyon Çalışması

Examining the Effects of Sample Size on Forecast-Verification Metrics: A Simulation Study

İsmet DOĞAN[a], Nurhan DOĞAN[a]

[a]Afyonkarahisar Sağlık Bilimleri Üniversitesi Tıp Fakültesi, Biyoistatistik ve Tıbbi Bilişim AD, Afyonkarahisar, Türkiye

ÖZET Amaç: Bu çalışmanın amacı, yaygın kullanılan tahmin doğruluk metriklerinin tanıtılması ve farklı örneklem büyüklükleri için performanslarının karşılaştırılmasıdır. **Gereç ve Yöntemler:** Tahmin doğrulama metrikleri, bilimin çeşitli disiplinlerinde karar verme araçları olarak yaygın şekilde kullanılmaktadır. İki sonuçlu tahminler söz konusu olduğunda tahmin doğrulamaya yönelik ayırt etme kapasitesi metrikler kullanılarak değerlendirilebilir. Bu metriklere eşiğe bağlı metrikler denir. Farklı örneklem büyüklüklerinin tahmin-doğrulama metriklerinin performansı üzerindeki etkisini ortaya koymak amacıyla yaygın olarak kullanılan 9 farklı eşik bazlı metrik dikkate alınarak bir simülasyon çalışması yapılmıştır. Python-random kütüphanesi kullanılarak $10 \leq n \leq 1000$ aralığında 35 farklı n değeri için veri elde edilmiştir. Performans değerlendirmesinde Kappa katsayısı için literatürde önerilen değerler ve yorumlama düzeyleri dikkate alınmıştır. **Bulgular:** Farklı örneklem büyüklüklerinin dikkate alındığı bu çalışmadan elde edilen sonuçlardan, örneklem büyüklüğünün artırılması veya azaltılmasının tahmin doğrulama üzerindeki etkisinin neredeyse sabit olduğu tespit edilmiştir. Dikkate alınan tüm örneklem büyüklükleri için tahminlerin yaklaşık %50'sinin neredeyse tüm metrikler için "doğrulama yok veya önemsiz" yorumlama seviyesine sahip olduğu görülmüştür. Metrikler adil, orta, önemli ve mükemmel doğrulama düzeyleri bir arada dikkate alınarak liberal olma bakımından sıralandığında sıralama F, Odds Oranı Beceri Puanı, Kritik Başarı İndeksi, Peirce Beceri Puanı, Clayton Beceri Puanı, Tahmin Beceri İndeksi, Heidke Beceri Puanı, Kappa ve Gilbert Beceri Puanı şeklinde elde edilmiştir. **Sonuç:** Tahmin doğrulama metrikleri, örneklem büyüklüğünden ziyade gözlem değerlerinin 2x2 çapraz tablolardaki gözelere dağılımdan daha çok etkilenmektedir.

Anahtar kelimeler: Tahmin doğrulaması; iki sonuçlu olaylar; tahmin beceri puanı; doğrulama çalışmaları

ABSTRACT Objective: The aim of this study is to introduce commonly used forecast-verification metrics and compare their performance for different sample sizes. **Material and Methods:** Forecast verification metrics are widely used as decision support tools in various scientific disciplines. If the prediction results are binary, metrics can be used to evaluate the discriminative power for prediction verification. These metrics are called threshold-dependent metrics. In order to show the effect of different sample sizes on the performance of forecast-verification metrics, a simulation study was conducted considering nine different commonly used threshold-based metrics. Using the Python-random library, data were obtained for 35 different n values in the range of $10 \leq n \leq 1000$. For the performance evaluation, the values and interpretation levels recommended in the literature for the Kappa coefficient were taken into account. **Results:** From the results of this study, where different sample sizes were considered, it was found that the effect of increasing or decreasing the sample size on forecast verification was almost constant. It was observed that for all sample sizes considered, around 50 percent of the estimates had "none or none to low" levels of interpretation for almost all metrics. When the metrics were ranked in terms of liberality by considering fair, moderate, substantial and perfect levels of verification together, the order was obtained as F, Odds Ratio Skill Score, Critical Success Index, Peirce Skill Score, Clayton Skill Score, Prediction Skill Index, Heidke Skill Score, Kappa and Gilbert Skill Score. **Conclusion:** Forecast-verification metrics are more affected by the distribution of observations across cells in 2x2 crosstabs than by sample size.

Keywords: Forecast-verification; binary events; prediction skill score; validation studies

Correspondence: İsmet DOĞAN
Afyonkarahisar Sağlık Bilimleri Üniversitesi Tıp Fakültesi, Biyoistatistik ve Tıbbi Bilişim AD, Afyonkarahisar, Türkiye
E-mail: ismet.dogan@afsu.edu.tr

Peer review under responsibility of Turkiye Klinikleri Journal of Biostatistics.

Received: 10 Apr 2024 ***Received in revised form:*** 12 Jun 2024 ***Accepted:*** 13 Jun 2024 ***Available online:*** 28 Jun 2024

Tahmin, hemen tüm disiplinler için ortak bir çabadır ve tahminlerin en iyi şekilde nasıl doğrulanabileceği sorusu, bilim insanlarının büyük bir kısmı için temel öneme sahiptir.[1] Genel olarak tahminlerin niteliğini ve değerini belirleme süreci ve uygulamasına tahmin değerlendirmesi denir. Bir tahminin kalitesinin belirlenmesinde ampirik değerlendirme (doğrulama) ve operasyonel değerlendirme olmak üzere iki tür değerlendirme söz konusudur. Bu değerlendirmelerin yanı sıra beceri de bir tahminin kalitesinin önemli bir ölçüsüdür. İyi kalitede bir tahmin aynı zamanda belirli bir referans temel çizgisinin üzerindeki doğruluk derecesi olan beceriyi de gösterebilir. Bir tahminin doğruluğu ve becerisi belirlenerek tahminin kalitesi geliştirilebilir ve gelecekte güvenle kullanılabilir.[2] Tahminler gelecekte ilgilenilen olay hakkında tamamen bilgisiz kalmaktansa, bir tahminin mevcut olmasının tercih edildiği inancıyla yapılmakta ve kullanılmaktadır. Tahminin ne kadar becerikli veya değerli olduğunu değerlendirerek bu inancı tahmin sonrası doğrulamak ise en az tahmin kadar önemlidir. Birçok disiplin tarafından kullanılan tahmin doğrulama, önceki tahminlerin bir örneğine veya örneklerine ve ilgili gözlemlere dayanarak bir tahmin sisteminin kalitesinin araştırılması ve değerlendirilmesi olarak tanımlanmaktadır.[3] Tahmin doğrulama bilimsel tahmin sisteminin önemli bir bileşeni olmasından dolayı birçok önemli amaca hizmet eder. Bu amaçlar, yapay zekâ başta olmak üzere tahmin teknolojisindeki son durumu ve tahmin kalitesindeki son eğilimleri değerlendirmeyi, tahmin prosedürlerini ve sonuçta tahminlerin kendisini iyileştirmeyi ve kullanıcılara tahminlerden etkin bir şekilde yararlanmak için gereken bilgileri sağlamayı içerir.[4] Tahmin doğruluğunun değerlendirilmesi tahmin modellerinin geliştirilmesinin temel konularından biridir.[5] Model performansını ölçmenin ayırt etme kapasitesi ve güvenirlik olmak üzere iki yönü vardır. Ancak ayırt etme kapasitesinin genel olarak güvenirlikten daha önemli olduğu ifade edilmektedir. Ayırt etme kapasitesi, bir modelin söz konusu fenomenin varlığı ve yokluğu arasında ayrım yapma yeteneğini ölçer. Güvenirlik, tahmin edilen değerler ile gözlenen değerler arasındaki uyumu ifade eder. Ayırt etme kapasitesini ve/veya güvenirliği değerlendirmek için çeşitli metrikler kullanılır. Bunların bir kısmı yalnızca iki düzeye sahip sonuçlara veya eşik adı verilen belirli bir kesme değeri (cut-off) kullanılarak iki sonuçlu çözüme dönüştürülmüş sürekli verilere uygulanabilir. Bu metriklere eşiğe bağlı metrikler denir. Eşiğe bağlı metrikler arasında genel doğruluk, duyarlılık, özgüllük, pozitif tahmin değeri, negatif tahmin değeri, olasılık oranı, gerçek beceri istatistiği, F metriği, Cohen'in kappası vb. yer alır. Eşiğe bağlı metrikler, tıbbi tanı, hava tahmini ve makine öğrenimi dâhil olmak üzere farklı alanlarda model doğruluğu değerlendirmesi için kullanılmaktadır. Metriklerin kullanımı büyük ölçüde belirli alanlarla sınırlıdır.[6] Çeşitli disiplinlerde tahminlere olan bağımlılığın artmasından dolayı, tahmin beceri metrikleri önerilmiştir. Tahmin becerisinin tahmin doğruluğu ile aynı şey olmadığını açıklığa kavuşturmak çok önemlidir. Tahmin becerisi, bir dizi tahminin bazı referans tahminlere göre göreceli doğruluğunu ifade etmektedir. Tahmin edilen olayın türüne, zorluk seviyesine ve tahminde kullanılan bilimsel bilginin durumuna bağlıdır.[7] 1980'lerin ortasından önce iki sonuçlu olayları tahmin etme becerisinin değerlendirilmesinde gözlenen değer ile tahmin edilen değer arasındaki farkı dikkate alan geleneksel yaklaşım kullanılmıştır. Bu yaklaşım, doğrulamaya yönelik *ölçüm odaklı* yaklaşım olarak bilinir ve şaşırtıcı derecede geniş bir dizi metriğin geliştirilmesine yol açmıştır. Bu metrikler genel olarak makul olsalar da belirli doğrulama sorunları için uygun metriklerin seçimine rehberlik edecek veya belirli metriklerin özelliklerini sistematik olarak tartışacak temel teorinin üzerinde anlaşmaya varılan arka plan konusunda çok az şey biliniyordu. Bu duruma çare bulmak amacıyla tahminlerin ve gözlemlerin ortak olasılık dağılımına dayalı doğrulama için genel bir çerçeve önerilmiş ve tahmin doğrulamayı bu ortak dağılımın istatistiksel özelliklerinin değerlendirilmesi süreci olarak tanımlanmıştır. Bu yaklaşım, *dağılım odaklı doğrulama* veya bazen *tanısal doğrulama* olarak bilinmektedir. Performans metriklerini değerlendirme sürecini ve bu metrikler için değerlendirme kriterlerinin geliştirilmesini ifade etmek üzere meta-doğrulama terimi türetilmiş ve metrikler meta-doğrulama kriterlerine göre karşılaştırılmıştır.[8] Performans metriğinin seçimi, herhangi bir sınıflandırma probleminde hangi yöntemin kullanılacağına karar vermede kritik öneme sahiptir. Farklı metrikler çok farklı seçimlere yol açabilir, dolayısıyla metrik hedeflerle eşleşmelidir.[9] Tahmin performansına ilişkin alternatif birçok performans metriği geliştirilmiştir. Ancak bu metriklerin bazıları yıllar içerisinde farklı isimler ile tekrarlı olarak literatüre kazandırılmıştır.[8] İki sonuçlu sınıflandırıcılar, tahmin değerini olay oluşumu ve gerçekleşmeme sınıflarına ayırmak için bir eşik kullanan sürekli bir tahmin ediciden elde edilir. İki sonuçlu sınıflandırma sonuçları için standart 2x2 çapraz tablosu Tablo 1'de gösterilmektedir.

TABLO 1: İki sonuçlu sınıflandırma için 2×2 çapraz tablo.

		Gerçek durum	
		Pozitif	**Negatif**
Tahmin sonucu	Pozitif	$GP = a$	$YP = b$
	Negatif	$YN = c$	$GN = d$

GP: Gerçek pozitif; YP: Yanlış pozitif; YN: Yanlış negatif; GN: Gerçek negatif.

TABLO 2: Eşiğe bağlı tahmin doğrulama metrikleri.

Metrik	Formül	Değer aralığı
PSI[7]	$\frac{\left\{\frac{\frac{a}{n}-\left[\frac{a+b}{n}*\frac{a+c}{n}\right]}{\sqrt{\frac{a+b}{n}*\frac{a+c}{n}}}+\frac{\frac{d}{n}-\left[\frac{b+d}{n}*\frac{c+d}{n}\right]}{\sqrt{\frac{b+d}{n}*\frac{c+d}{n}}}\right\}-\left\{\frac{\frac{b}{n}-\left[\frac{a+b}{n}*\frac{b+d}{n}\right]}{\sqrt{\frac{a+b}{n}*\frac{b+d}{n}}}+\frac{\frac{c}{n}-\left[\frac{a+c}{n}*\frac{c+d}{n}\right]}{\sqrt{\frac{a+c}{n}*\frac{c+d}{n}}}\right\}}{2}$	$-1 \le PSI \le 1$
CSI[8]	$\frac{a}{a+b+c}$	$0 \le CSI \le 1$
CSS[8]	$\frac{a}{a+b}-\frac{c}{c+d}$	$-1 \le CSS \le 1$
HSS[11]	$\frac{2\,(ad-bc)}{(a+b)(b+d)+(a+c)(c+d)}$	$-1 \le HSS \le 1$
GSS[8]	$\frac{a-\left[\frac{(a+b)(a+c)}{n}\right]}{a+b+c-\left[\frac{(a+b)(a+c)}{n}\right]}$	$-1/3 \le GSS \le 1$
PSS[8]	$\frac{ad-bc}{(a+c)(b+d)}$	$-1 \le PSS \le 1$
F[9]	$\frac{2a}{2a+b+c}$	$0 \le F \le 1$
ORSS[8]	$\frac{ad-bc}{ad+bc}$	$-1 \le ORSS \le 1$
Kappa[5]	$\frac{(\frac{a+d}{n})-\frac{(a+b)(a+c)+(c+d)(d+b)}{n^2}}{1-\frac{(a+b)(a+c)+(c+d)(d+b)}{n^2}}$	$-1 \le Kappa \le 1$

PSI: Tahmin Beceri İndeksi; CSI: Kritik Başarı İndeksi; CSS: Clayton Beceri Puanı; HSS: Heidke Beceri Puanı; GSS: Gilbert Beceri Puanı; PSS: Peirce Beceri Puanı; ORSS: Odds Oranı Beceri Puanı.

Her eşik ile bir çapraz tablo ilişkilendirilir ve bu tablodan birçok değer hesaplanabilir. Eşiğe bağlı tüm metrikler, 2x2 çapraz tablosunda yer alan bazı veya tüm ögelerine dayanmaktadır. Tanısal doğruluk çalışmalarında en sık kullanılan metrikler duyarlılık, özgüllük, pozitif ve negatif öngörü değerleridir. Tahmin modelleri tipik olarak en yüksekten en düşüğe pozitif öngörü değeri ve duyarlılığa göre sıralanır ve daha yüksek değerlere sahip olanlara öncelik verilir. Bununla birlikte, böyle bir yöntemin objektif olarak uygulanması zordur çünkü pozitif öngörü değeri ile duyarlılık arasında biri artarken diğeri azalan bir değiş-tokuş ilişkisi vardır. Bu durum, tanı modellerini doğruluk sırasına göre sıralamaya çalışırken hangi tahmine (pozitif öngörü değeri veya duyarlılık) öncelik verilmesi gerektiğini bilmeyi zorlaştırmaktadır.[10] Bu zorluğu giderebilmek için metrikler önerilmiştir. Önerilen metriklerin bazıları Tablo 2'de verilmiştir.

Tablo 2'den de görüldüğü üzere Kritik Başarı İndeksi [Critical Success Index (CSI)] ve F metrikleri 0-1 aralığında, diğer metrikler ise -1 ile +1 arasında değer almaktadır. Metrikler için daha yüksek değer, iyi tahmin anlamına gelir. +1 mükemmel doğrulamayı belirtir ve sıfır veya daha düşük değerler rastgeleden daha iyi olmayan bir performansı gösterir.[5,11] İki sonuçlu tahminler için tahmin kalitesi tipik olarak,

gerçek pozitif, yanlış pozitif, gerçek negatif ve yanlış negatiflerin sayısını dayalı beceri puanları kullanılarak ölçülür. Spesifik olarak beceri puanları, 2x2 çapraz tablodaki köşegen dışı değerlere dayalı dengesizliği farklı şekillerde hesaplayan basit aritmetik formüllere dayanır. Tahmin kalitesi için sıklıkla kullanılan tipik beceri puanları, *gerçek beceri istatistikleri* [true skill statistics (*TSS)]*, Heidke Beceri Puanı [Heidke Skill Score (HSS)] ve CSI'dır.[12] TSS, makine öğrenimi tabanlı modellemenin performansını değerlendirmek için kullanılan en pratik metriktir. 2x2 çapraz tablonun oluşturulmasında bir eşik değeri dikkate alındığından TSS'nin hesaplanması için en uygun eşiğin belirlenmesi gerekir. Eşik değer olarak genellikle duyarlılık ve özgüllük değerlerinin toplamının en büyük olduğu değer önerilir.[13] CSI ve F metrikleri, hem pozitif öngörü değerinin hem de duyarlılığın yüksek ve pozitif öngörü değeri ile duyarlılık değeri arasındaki farkın büyük olduğu durumlara öncelik vermektedir. Bu nedenle, CSI veya F metrikleri, tanısal doğruluk çalışmalarındaki pozitif öngörü değeri ve duyarlılığın yanı sıra tamamlayıcı metrikler olarak yararlı olabilir.[10] Çalışmada tipik olarak kullanılan tahmin doğruluk metriklerinin tanıtılması ve farklı örneklem büyüklükleri için performanslarının karşılaştırılması amaçlanmıştır. Araştırmada, Helsinki Deklarasyonu prensipleri dikkate alınmıştır.

GEREÇ VE YÖNTEMLER

Bir metriğin değerini hesaplamanın amacı, bir modelin başka bir model yerine kullanılmasıyla elde edilen kazanıma bir miktar bağlam kazandırmaktır. Mükemmel bir nokta tahmininin söz konusu olması durumunda metriğin değeri tahmin doğruluğundaki kazanç olarak tanımlanır.[1] Tahmin doğrulama değerlerinin yorumlanmasında ve performans değerlendirmesinde Kappa katsayısı için belirlenen değer aralıkları ve yorumlama seviyeleri diğer metriklere göre daha belirgin olduğundan Kappa katsayısı için literatürde önerilen ve Tablo 3'te verilen sınırlar ve yorumlama düzeyleri kullanılmıştır.[14]

TABLO 3: Tahmin doğrulama değerleri ve yorumlama düzeyleri.

Tahmin doğrulama beceri değeri	Yorumlama düzeyi
$Değer \leq 0$	Doğrulama yok
$0{,}00 < Değer \leq 0{,}20$	Önemsiz doğrulama
$0{,}20 < Değer \leq 0{,}40$	Adil düzeyde doğrulama
$0{,}40 < Değer \leq 0{,}60$	Orta düzeyde doğrulama
$0{,}60 < Değer \leq 0{,}80$	Önemli derecede doğrulama
$0{,}80 < Değer \leq 1{,}00$	Mükemmel doğrulama

Çalışmada Phyton-random kütüphanesi kullanılarak $10 \leq n \leq 1000$ aralığında yer alan 35 farklı n değeri (10, 15, 20, 25, 30, 35, 40, 45, 50, 55, 60, 65, 70, 75, 80, 85, 90, 95, 100, 125, 150, 175, 200, 225, 250, 300, 350, 400, 450, 500, 600, 700, 800, 900, 1000) için veri türetilmiştir. Veri türetimine ilişkin Python 3.9.10 (Python Software Foundation, ABD) programlama dili kullanılarak yazılan programda yer alan simülasyon kurgusunun detayları aşağıdaki gibidir:

Adım 1. Veri türetimi için gerekli değerler (toplam kaç adet veri türetilmek istendiği, kategorilere atanacak değerlerin toplamı, kategori sayısı) programa girilir.

Adım 2. Python random kütüphanesindeki choice fonksiyonu kullanılarak rastgele bir şekilde kategori (GP: Gerçek Pozitif, YP: Yanlış Pozitif, YN: Yanlış Negatif, GN: Gerçek Negatif) seçimi yapılır.

Adım 3. Choice fonksiyonu tekrar kullanılarak seçilen kategori için 0 ile istenen örneklem büyüklüğü arasında rastgele bir tam sayı seçimi yapılır. Örneğin istenen örneklem büyüklüğü 30 ise (0, 30) aralığında rastgele bir tam sayı seçilir.

Adım 4. Tüm kategoriler için ikinci ve üçüncü adımlar tekrarlanır k'inci kategori için bir tam sayı seçildikten sonra seçilen sayıların toplamına bakılır. Eğer bu toplam, istenilen toplamdan fazla ise bu geçersiz bir veri olacağından işleme baştan başlanır. Son kategoriye atanacak tam sayı ise istenilen toplamdan mevcut toplamın çıkarılmasıyla bulunur.

Adım 5. Türetilen verinin daha önce türetilip türetilmediğine bakılır. Aynı veri daha önce türetilmişse veri setine eklenmez.

Adım 6. İstenilen toplam veri sayısına ulaşana kadar birinci adımdan itibaren program bir döngü içinde tekrar çalıştırılır. (Bu çalışmada standardizasyonu sağlamak amacıyla her bir örneklem genişliği için 1.000 adet veri türetilmesi planlanmış ancak bazı örneklem büyüklükleri için tekrar etmeyen 1.000 tane veri bulmak imkânsız olduğundan farklı örneklem büyüklüklerinde veri türetilmiştir.)

Adım 7. Son olarak türetilen verilerden basit bir Excel uygulaması kullanılarak çalışmada dikkate alınan metriklere ait değerler hesaplanır.

BULGULAR

Farklı örneklem büyüklükleri ve doğrulama seviyeleri için çalışmada dikkate alınan metriklere ait tahmin doğrulama beceri yüzdeleri Tablo 4'te verilmiştir.

Tablo 4'te verilen sonuçlar örneklem büyüklüğünün metriklerin tahmin doğrulama becerisi üzerinde etkisinin neredeyse sabit kaldığını göstermektedir. Örneklem büyüklüğünün artması ya da azalması hemen tüm metriklerde yorumlama seviyelerine ait beceri yüzdelerinde dikkate değer bir değişime neden olmamaktadır. HSS ve Kappa metrikleri birbirine en çok benzeyen metriklerdir. $n \geq 30$ için tüm metriklerin %50 oranında "doğrulama yok" yorumlama seviyesine sahip olduğu, genel olarak adil, orta, önemli ve mükemmel yorumlama seviyeleri olumlu olarak düşünüldüğünde tutuculuk bakımından metriklerin sıralaması Gilbert Beceri Puanı [Gilbert Skill Score (GSS)], HSS, Kappa, Tahmin Beceri İndeksi [Prediction Skill Index (PSI)], Clayton Beceri Puanı, Peirce Beceri Puanı [Peirce Skill Score (PSS)], Odds Oranı Beceri Puanı, CSI ve F şeklindedir.

TABLO 4: Metriklere ait tahmin doğrulama beceri yüzdeleri.

n	Doğrulama	CSI	CSS	HSS	GSS	PSS	F	ORSS	Kappa	PSI
10	Yok	-	45,1	41,9	41,9	45,1	-	48,4	41,9	48,4
	Önemsiz	41,8	20,1	30,6	36,3	20,1	29,5	5,7	30,6	12,6
	Adil	22,8	12,1	10,6	9,9	12,1	18,9	2,8	10,9	15,4
	Orta	16,8	9,5	7,4	6,0	9,5	17,9	4,1	7,4	12,6
	Önemli	11,6	6,8	6,3	2,8	6,8	19,6	4,1	6,3	6,5
	Mükemmel	7,0	6,4	3,2	3,2	6,4	14,0	35,0	2,8	4,5
15	Yok	-	47,2	45,5	45,5	47,2	-	48,9	45,5	48,9
	Önemsiz	40,2	18,6	25,6	34,4	18,6	26,9	5,8	25,6	15,9
	Adil	24,2	12,5	11,8	9,6	12,5	19,7	4,2	11,8	14,6
	Orta	17,8	9,7	8,4	6,1	9,7	20,5	5,8	8,5	11,1
	Önemli	11,8	6,9	4,9	2,7	6,9	20,1	4,2	4,9	6,1
	Mükemmel	6,0	5,1	3,9	1,7	5,1	12,8	31,0	3,8	3,4
20	Yok	-	47,8	46,7	46,7	47,7	-	48,7	46,7	48,7
	Önemsiz	38,0	18,4	24,8	33,0	19,4	25,4	6,6	24,8	17,2
	Adil	25,6	13,2	11,8	10,3	12,1	19,9	6,1	11,8	14,7
	Orta	19,5	9,6	8,4	6,0	9,4	20,6	5,3	8,5	11,0
	Önemli	10,8	7,4	5,9	2,3	7,3	22,3	5,7	5,9	6,3
	Mükemmel	6,1	3,6	2,3	1,6	4,2	11,8	27,6	2,2	2,1

25	Yok	-	47,7	47,0	47,0	48,1	-	48,8	47,0	48,8
	Önemsiz	39,5	18,7	23,8	32,7	17,2	25,1	5,2	23,8	16,7
	Adil	23,8	12,4	12,7	11,0	14,3	21,4	5,4	12,7	16,0
	Orta	18,1	10,2	8,0	5,7	9,7	18,4	5,7	8,0	9,7
	Önemli	13,5	6,8	5,8	2,2	7,2	22,6	7,5	5,9	6,4
	Mükemmel	5,1	4,3	2,7	1,4	3,6	12,5	27,4	2,6	2,4
30	Yok	-	50,9	50,4	50,4	51,0	-	51,4	50,4	51,4
	Önemsiz	40,0	18,7	23,2	32,2	18,6	24,4	6,3	23,2	18,3
	Adil	23,8	12,2	12,3	10,4	13,2	22,9	6,1	12,3	14,7
	Orta	21,7	10,3	8,2	4,2	9,1	19,6	6,9	8,2	9,4
	Önemli	10,1	5,1	3,7	1,8	5,3	22,6	6,7	3,8	4,6
	Mükemmel	4,4	2,8	2,2	1,0	2,8	10,5	22,4	2,1	1,6
35	Yok	-	49,8	49,5	49,5	49,8	-	50,2	49,5	50,2
	Önemsiz	37,5	17,4	21,6	31,5	17,6	23,3	6,2	21,6	17,8
	Adil	27,9	12,9	12,7	10,9	13,3	22,7	5,7	12,7	13,9
	Orta	19,0	9,8	9,2	4,9	10,8	22,3	6,4	9,2	11,0
	Önemli	11,6	7,0	4,7	1,9	5,7	22,3	7,7	4,8	5,1
	Mükemmel	4,0	3,0	2,3	1,3	2,7	9,4	23,8	2,2	2,0
40	Yok	-	49,2	48,9	48,9	49,2	-	49,6	48,9	49,6
	Önemsiz	35,3	19,3	23,8	32,0	19,4	20,8	6,9	23,8	19,8
	Adil	28,7	12,0	11,8	11,3	12,4	22,8	7,1	11,9	13,2
	Orta	19,0	9,9	8,5	4,9	9,7	24,6	6,0	8,5	10,4
	Önemli	12,2	7,0	5,3	2,1	6,5	20,0	7,3	5,3	5,4
	Mükemmel	4,8	2,5	1,7	0,8	2,7	11,8	23,1	1,6	1,6
45	Yok	-	48,4	48,2	48,2	48,4	-	48,6	48,2	48,6
	Önemsiz	38,3	18,1	22,5	33,6	19,0	24,2	6,1	22,5	18,9
	Adil	25,8	15,7	14,3	9,9	14,4	20,6	7,4	14,3	15,9
	Orta	20,3	8,7	7,3	5,3	8,4	22,4	7,4	7,3	9,1
	Önemli	12,2	6,5	5,5	2,3	7,1	21,9	9,1	5,6	6,0
	Mükemmel	3,4	2,6	2,2	0,7	2,6	10,9	21,5	2,1	1,5
50	Yok	-	50,1	50,0	50,0	50,2	-	50,3	50,0	50,3
	Önemsiz	37,0	17,1	21,0	29,8	16,1	21,7	4,9	21,0	18,0
	Adil	27,4	12,8	12,4	12,1	13,2	22,8	6,5	12,4	14,1
	Orta	19,2	11,6	9,6	4,4	10,8	22,4	7,1	9,6	10,7
	Önemli	11,7	5,0	4,8	3,0	6,2	22,0	7,5	4,9	5,3
	Mükemmel	4,7	3,3	2,2	0,7	3,4	11,1	23,6	2,1	1,7
55	Yok	-	49,4	49,4	49,4	49,6	-	49,7	49,4	49,7
	Önemsiz	36,8	18,4	22,7	30,9	17,2	22,7	5,9	22,7	18,7
	Adil	27,4	13,0	11,9	9,7	14,4	22,0	6,9	11,9	14,3
	Orta	19,5	7,8	6,8	6,5	8,4	22,6	6,8	6,8	8,6
	Önemli	11,7	8,0	7,1	2,6	7,0	20,5	9,0	7,1	7,1
	Mükemmel	4,6	3,3	2,1	0,9	3,3	12,2	21,6	2,1	1,6
60	Yok	-	49,4	49,4	49,4	49,7	-	49,7	49,4	49,7
	Önemsiz	38,6	19,0	22,4	31,6	17,5	22,5	6,9	22,4	18,7
	Adil	26,3	13,1	12,4	10,5	13,9	24,2	6,0	12,4	14,3
	Orta	20,4	8,4	8,1	5,4	9,3	21,3	7,6	8,1	9,4
	Önemli	10,8	7,4	5,6	2,0	6,7	21,5	8,7	5,6	6,1
	Mükemmel	3,9	2,7	2,1	1,1	2,8	10,5	21,0	2,1	1,7
65	Yok	-	49,8	49,8	49,8	49,9	-	49,9	49,8	49,9
	Önemsiz	37,3	17,6	21,3	32,1	18,1	23,3	6,1	21,3	18,3
	Adil	26,3	15,8	15,4	11,4	14,6	22,0	7,0	15,4	17,7
	Orta	21,6	9,3	8,1	3,6	10,5	22,2	8,5	8,1	8,6
	Önemli	10,8	4,6	3,6	2,6	4,7	22,7	9,4	3,6	4,0
	Mükemmel	4,0	2,8	1,8	0,5	2,1	9,8	19,0	1,8	1,5

70	Yok	-	51,1	51,0	51,0	51,1	-	51,2	51,0	51,2
	Önemsiz	38,9	18,3	21,4	30,8	17,4	21,5	6,4	21,4	18,7
	Adil	28,3	13,5	12,5	10,7	13,6	24,6	7,1	12,5	14,0
	Orta	16,9	8,5	8,9	5,2	9,7	24,0	6,7	8,9	9,7
	Önemli	12,4	6,7	4,6	1,6	5,6	18,5	8,9	4,6	5,1
	Mükemmel	3,5	1,8	1,6	0,7	2,5	11,4	19,6	1,6	1,3
75	Yok	-	48,8	48,7	48,7	48,8	-	48,9	48,7	48,9
	Önemsiz	37,0	18,3	22,4	30,9	18,2	23,0	6,5	22,4	19,6
	Adil	27,5	13,7	13,2	10,7	13,2	20,8	7,3	13,2	14,2
	Orta	19,0	8,8	7,3	5,5	9,7	24,4	7,6	7,3	9,6
	Önemli	10,3	6,8	5,9	3,1	6,2	19,7	8,3	5,9	5,3
	Mükemmel	6,2	3,5	2,5	1,1	3,8	12,1	21,3	2,5	2,4
80	Yok	-	51,2	51,1	51,1	51,2	-	51,2	51,1	51,2
	Önemsiz	36,8	16,8	21,5	29,7	18,2	21,9	6,1	21,5	18,6
	Adil	28,2	12,9	11,9	11,4	12,2	23,4	6,9	11,9	13,7
	Orta	18,5	11,0	8,8	5,1	10,0	22,5	7,9	8,8	9,7
	Önemli	12,4	5,8	5,0	1,7	6,3	20,4	8,5	5,0	5,4
	Mükemmel	4,1	2,3	1,7	1,0	2,1	11,8	19,3	1,7	1,3
85	Yok	-	48,6	48,6	48,6	48,6	-	48,6	48,6	48,6
	Önemsiz	38,4	18,5	23,2	32,2	18,8	23,6	6,0	23,2	19,0
	Adil	25,4	14,5	13,2	11,9	15,1	23,7	7,4	13,2	16,5
	Orta	20,1	10,2	9,0	5,2	9,7	19,8	6,9	9,1	9,3
	Önemli	12,4	6,1	4,5	1,6	5,7	22,0	9,9	4,5	5,5
	Mükemmel	3,7	2,0	1,5	0,5	2,1	10,9	21,1	1,4	1,0
90	Yok	-	48,6	48,6	48,6	48,6	-	48,6	48,6	48,6
	Önemsiz	37,3	17,6	21,8	31,9	18,3	22,4	7,4	21,8	19,7
	Adil	28,3	14,9	13,5	11,5	14,5	21,4	6,3	13,5	14,8
	Orta	17,9	8,7	8,7	4,4	9,6	25,3	7,4	8,7	9,4
	Önemli	13,0	6,7	5,0	2,5	6,6	18,5	9,0	5,0	5,4
	Mükemmel	3,5	3,4	2,4	1,1	2,4	12,4	21,2	2,4	2,0
95	Yok	-	48,6	48,5	48,5	48,6	-	48,7	48,5	48,7
	Önemsiz	37,3	17,6	21,5	31,7	17,1	22,1	6,0	21,5	18,5
	Adil	24,3	14,6	13,9	11,9	13,6	21,7	6,9	13,9	15,7
	Orta	20,4	11,4	9,6	5,4	11,6	19,8	7,5	9,6	10,8
	Önemli	13,5	5,6	5,0	2,0	6,9	24,6	9,5	5,0	5,2
	Mükemmel	4,5	2,1	1,5	0,5	2,1	11,8	21,3	1,5	1,1
100	Yok	-	53,1	53,0	53,0	53,1	-	53,2	53,0	53,2
	Önemsiz	38,9	15,8	20,7	29,1	17,2	23,0	5,3	20,7	17,5
	Adil	27,5	13,0	11,2	9,4	11,8	24,6	6,4	11,2	12,4
	Orta	16,6	9,0	7,1	5,4	8,9	21,5	6,7	7,2	9,2
	Önemli	12,3	7,0	6,3	2,2	6,2	19,2	8,6	6,3	6,0
	Mükemmel	4,7	2,0	1,7	0,9	2,7	11,7	19,7	1,6	1,6
125	Yok	-	50,7	50,7	50,7	50,8	-	50,8	50,7	50,8
	Önemsiz	36,2	17,7	21,9	30,1	18,3	22,7	6,0	21,9	19,6
	Adil	28,3	13,9	12,2	9,9	12,7	20,1	6,8	12,2	13,6
	Orta	18,9	8,6	7,1	5,6	8,4	24,7	8,1	7,2	8,0
	Önemli	12,5	6,5	5,8	2,4	6,6	20,9	9,3	5,8	6,2
	Mükemmel	4,1	2,6	2,3	1,3	3,2	11,6	19,0	2,2	1,8
150	Yok	-	50,0	50,0	50,0	50,0	-	50,0	50,0	50,0
	Önemsiz	36,7	17,3	20,8	30,8	18,0	20,5	6,1	20,8	18,3
	Adil	28,8	14,0	13,8	12,4	13,8	23,7	7,1	13,8	15,3
	Orta	18,0	10,3	10,2	4,6	11,2	24,4	6,4	10,3	11,2
	Önemli	12,9	5,9	3,3	1,6	5,0	20,2	11,4	3,3	4,0
	Mükemmel	3,6	2,5	1,9	0,6	2,0	11,2	19,0	1,8	1,2

175	Yok	-	52,2	52,2	52,2	52,2	-	52,2	52,2	52,2
	Önemsiz	37,5	17,6	21,1	30,1	17,9	22,0	6,1	21,1	18,3
	Adil	28,1	14,4	12,3	10,1	12,1	24,0	7,6	12,3	14,0
	Orta	18,5	8,2	7,3	4,7	9,7	22,6	6,5	7,4	9,3
	Önemli	11,1	5,4	5,2	2,0	5,4	19,7	8,8	5,2	4,4
	Mükemmel	4,8	2,2	1,9	0,9	2,7	11,7	18,8	1,8	1,8
200	Yok	-	48,5	48,5	48,5	48,5	-	48,5	48,5	48,5
	Önemsiz	35,1	19,8	22,0	30,7	18,3	20,0	8,5	22,0	20,6
	Adil	27,4	12,0	13,0	11,4	13,4	22,4	6,2	13,0	13,8
	Orta	21,1	10,7	8,5	5,9	10,4	23,7	7,2	8,5	9,4
	Önemli	12,4	6,4	5,9	3,1	6,6	22,6	9,4	6,0	6,1
	Mükemmel	4,0	2,6	2,1	0,4	2,8	11,3	20,2	2,0	1,6
225	Yok	-	51,3	51,3	51,3	51,4	-	51,4	51,3	51,4
	Önemsiz	37,4	16,5	20,0	28,1	16,9	23,0	6,7	20,0	18,0
	Adil	29,5	14,0	13,2	14,2	13,2	22,3	6,4	13,2	14,3
	Orta	20,0	10,3	9,5	4,6	11,3	25,2	6,6	9,5	10,4
	Önemli	9,9	6,1	4,4	1,5	4,9	20,4	9,7	4,5	4,6
	Mükemmel	3,2	1,8	1,6	0,3	2,3	9,1	19,2	1,5	1,3
250	Yok	-	52,0	52,0	52,0	52,0	-	52,0	52,0	52,0
	Önemsiz	38,4	15,8	19,6	28,4	16,5	22,7	6,6	19,6	17,4
	Adil	27,4	14,5	13,4	12,2	14,1	23,4	5,4	13,4	15,3
	Orta	17,1	9,7	8,3	5,0	9,5	23,0	7,2	8,4	8,8
	Önemli	13,1	5,8	4,7	1,4	5,4	19,0	11,2	4,7	4,7
	Mükemmel	4,0	2,2	2,0	1,0	2,5	11,9	17,6	1,9	1,8
300	Yok	-	51,3	51,3	51,3	51,3	-	51,3	51,3	51,3
	Önemsiz	34,3	16,8	20,9	29,9	16,5	20,5	5,3	20,9	17,9
	Adil	29,8	13,5	11,9	10,8	14,0	23,1	6,9	11,9	14,5
	Orta	19,1	10,4	8,9	5,3	10,1	23,9	8,5	8,9	9,5
	Önemli	13,0	5,6	5,3	2,0	6,1	20,0	9,0	5,4	5,6
	Mükemmel	3,8	2,4	1,7	0,7	2,0	12,5	19,0	1,6	1,2
350	Yok	-	51,4	51,4	51,4	51,4	-	51,4	51,4	51,4
	Önemsiz	38,2	15,3	19,2	30,0	16,1	23,2	6,0	19,2	17,3
	Adil	25,1	15,1	14,3	10,7	15,3	21,8	5,4	14,3	15,7
	Orta	20,8	10,2	8,5	5,4	8,5	20,7	7,9	8,6	9,6
	Önemli	12,1	6,2	5,3	2,4	7,3	22,8	11,4	5,3	5,0
	Mükemmel	3,8	1,8	1,3	0,1	1,4	11,5	17,9	1,2	1,0
400	Yok	-	49,7	49,7	49,7	49,7	-	49,7	49,7	49,7
	Önemsiz	36,1	17,8	20,6	31,6	17,1	22,3	6,2	20,6	18,0
	Adil	27,7	15,3	15,6	12,4	16,3	22,0	6,7	15,6	16,9
	Orta	21,1	9,5	8,6	4,4	10,2	23,1	7,8	8,6	10,1
	Önemli	11,5	6,1	4,6	1,5	5,1	21,9	10,9	4,7	4,5
	Mükemmel	3,6	1,6	0,9	0,4	1,6	10,7	18,7	0,8	0,8
450	Yok	-	47,7	47,7	47,7	47,7	-	47,7	47,7	47,7
	Önemsiz	36,2	20,9	23,6	32,2	20,0	20,2	8,4	23,6	21,2
	Adil	26,6	12,2	12,2	10,8	13,3	23,5	7,7	12,2	14,1
	Orta	20,6	9,9	8,4	5,5	9,6	23,0	7,0	8,4	9,2
	Önemli	11,7	6,0	5,5	2,8	6,5	21,9	9,6	5,6	5,4
	Mükemmel	4,9	3,3	2,6	1,0	2,9	11,4	19,6	2,5	2,4
500	Yok	-	49,9	49,9	49,9	49,9	-	49,9	49,9	49,9
	Önemsiz	38,7	20,8	23,5	33,5	21,5	23,5	8,9	23,5	21,7
	Adil	28,1	13,0	13,7	9,6	12,3	24,1	7,4	13,8	14,2
	Orta	17,6	8,5	6,4	4,6	8,8	22,1	6,8	6,4	8,2
	Önemli	11,3	5,4	5,0	1,9	5,4	18,7	8,9	4,9	4,7
	Mükemmel	4,3	2,4	1,5	0,5	2,1	11,6	18,1	1,5	1,3

600	Yok	-	47,3	47,3	47,3	47,3	-	47,3	47,4	47,3
	Önemsiz	34,7	18,6	21,8	33,3	18,1	19,6	8,0	21,8	20,1
	Adil	25,9	15,3	14,4	10,5	15,2	23,0	7,1	14,4	15,8
	Orta	22,3	10,1	8,7	5,6	10,3	21,6	8,8	8,7	9,5
	Önemli	12,3	6,5	5,6	2,8	5,8	23,8	9,6	5,6	6,1
	Mükemmel	4,8	2,2	2,2	0,5	3,3	12,0	19,2	2,1	1,2
700	Yok	-	49,8	49,8	49,8	49,8	-	49,8	49,8	49,8
	Önemsiz	35,7	18,2	21,4	30,5	18,9	21,4	7,9	21,4	19,4
	Adil	27,3	13,8	13,8	11,3	12,6	21,3	6,3	13,8	14,6
	Orta	19,2	10,0	7,8	5,8	9,8	23,7	6,8	7,8	8,9
	Önemli	14,4	6,0	5,6	2,4	6,8	21,3	9,1	5,7	6,1
	Mükemmel	3,4	2,2	1,6	0,2	2,1	12,3	20,1	1,5	1,2
800	Yok	-	47,8	47,8	47,7	47,8	-	47,8	47,8	47,8
	Önemsiz	37,1	19,3	23,3	33,0	18,7	20,5	6,7	23,3	21,2
	Adil	25,7	13,9	12,6	9,7	14,3	23,7	7,0	12,7	14,0
	Orta	20,1	8,5	8,4	7,3	10,8	21,7	8,0	8,4	9,6
	Önemli	12,5	8,5	6,2	1,9	6,2	22,4	10,6	6,1	6,4
	Mükemmel	4,6	2,0	1,7	0,4	2,2	11,7	19,9	1,7	1,0
900	Yok	-	50,5	50,5	50,5	50,5	-	50,5	50,5	50,5
	Önemsiz	35,0	18,3	21,4	30,6	17,9	21,6	7,5	21,4	19,6
	Adil	28,5	14,3	13,6	11,7	14,1	20,2	7,5	13,6	14,9
	Orta	21,7	9,5	8,2	4,9	9,9	24,8	6,8	8,3	8,9
	Önemli	10,4	5,5	4,7	1,9	6,0	23,3	10,5	4,7	4,8
	Mükemmel	4,4	1,9	1,6	0,4	1,6	10,1	17,2	1,5	1,3
1000	Yok	-	50,0	50,0	50,0	50,0	-	50,0	50,1	50,0
	Önemsiz	38,1	17,8	23,0	32,5	19,7	24,0	6,8	23,0	20,4
	Adil	26,9	15,5	12,9	9,8	12,8	23,2	6,7	12,9	14,8
	Orta	20,2	9,4	8,1	5,8	10,3	20,9	8,9	8,1	9,2
	Önemli	11,4	5,8	4,8	1,9	5,9	22,3	8,7	4,8	4,5
	Mükemmel	3,4	1,5	1,2	-	1,3	9,6	18,9	1,1	1,1
Genel	Yok	-	49,7	49,5	49,5	49,7	-	49,9	49,5	49,9
	Önemsiz	37,3	18,0	22,1	31,4	18,1	22,5	6,5	22,1	18,8
	Adil	27,0	13,7	13,0	10,9	13,6	22,4	6,6	13,0	14,7
	Orta	19,5	9,6	8,3	5,2	9,8	22,4	7,2	8,3	9,6
	Önemli	11,9	6,3	5,2	2,2	6,2	21,2	8,9	5,2	5,4
	Mükemmel	4,3	2,6	2,0	0,8	2,6	11,4	20,8	1,9	1,6

CSI: Kritik Başarı İndeksi; CSS: Clayton Beceri Puanı; HSS: Heidke Beceri Puanı; GSS: Gilbert Beceri Puanı; PSS: Peirce Beceri Puanı; ORSS: Odds Oranı Beceri Puanı; PSI: Tahmin Beceri İndeksi.

TARTIŞMA

Literatürde mevcut tahmin doğrulama metrikleri bazı kayda değer benzerlikler ve farklılıklar içermektedir. Tahmin doğrulama alanındaki temel, kavramsal ve metodolojik gelişmelerin ve tartışmaların başlangıcı 1884 yılından 1893 yılına kadar olan dönemi ifade etmektedir. Bir asırdan uzun bir süre geçmesine rağmen tahmin doğrulama, sürekli olarak daha fazla yöntem ve tekniğin icat edilmesi ve/veya yeniden keşfedilmesiyle önemli gelişmeler yaşamaktadır.[7,15] Literatürde fazlasıyla tartışıldığı gibi tahmin doğrulama karmaşık ve çok boyutlu bir sorundur. Tahminler ve gözlemler arasındaki ilişkileri koruyan dağılım odaklı doğrulama yaklaşımları, tahmin kalitesinin birçok yönünü temsil etmede ölçüm odaklı yöntemlerden daha kapsamlıdır. Metrikler eksikliklerine rağmen birçok durumda kullanışlı olabilir ve sıklıkla kullanılmaya devam edebilir. Tahmin kalitesine ilişkin metriklerin, performansı anlamayı ve iyileştirmeyi amaçlayan araştırmaların bitiş noktası olmaktan ziyade başlangıç noktası olduğu açıktır.[16] Tanı koymada kullanılan yöntemleri doğruluklarına göre sıralamayı daha kolay ve daha objektif hâle getirmek için pozitif öngörü değerini ve duyarlılığı tek bir metrikte birleştirmenin yeni yollarını düşünmeye açıkça ihtiyaç vardır. Bu amaçla, CSI veya F metriğinin

kullanılması önerilmektedir.[17,18] CSI ve F metrikleri, pozitif öngörü ve duyarlılık değerlerini, ayrı ayrı hesaplamaya çalışmaktan ziyade tanısal doğruluk açısından yorumlanması ve derecelendirilmesi daha kolay olan kullanışlı tek bir ölçümde birleştirir. CSI, hava durumu tahminleri dışında nadiren kullanılmaktadır. Sağlık alanında CSI'yı kullanan herhangi bir çalışma henüz yayımlanmamıştır. Her ne kadar F metriğinin kullanımı yapay zekâ alanında yaygın olsa da diğer alanlarda bir tanısal doğruluk aracı olarak uygulanabilirliği ve rahatlığı henüz tam olarak kanıtlanmamıştır.[10] İşlem karakteristik eğrisi altında kalan alan [area under the curve (AUC)], doğruluk, duyarlılık ve özgüllük model performansı için sıklıkla kullanılan metriklerdir. AUC, tahmin doğruluğunun eşikten bağımsız bir ölçüsü olması nedeniyle en yaygın kullanılan metrik olmasına rağmen son zamanlarda model performansını abarttığı ve alternatif performans metriklerini gerektirdiği için eleştirilmektedir. Beceri istatistikleri (skill score) model performansının en pratik ölçüsü olarak kabul edilmektedirler.[13] İki sonuçlu olayların deterministik tahminleri yapıldığında, gerçek negatiflerin sayısının tanımlanmasının zor olduğu veya performans metriklerinin çoğunun hesaplamalarında baskın olacak kadar büyük olduğu durumlarda tercih edilen metrikler F ve CSI'dır. Hem F hem de CSI metrikleri gerçek negatiflerin sayısından bağımsızdır.[19] Ancak sınıflandırma bağlamında gerçek negatiflerin sayısı nadiren önemsizdir. Örneğin bazı hastaların kanser olduğunun bilindiği kişisel sağlık kayıtlarından oluşan bir veri tabanının kullanıldığı bir çalışmada, gerçekten kanser olan hastaların sınıflandırılması açıkça odak noktası olmakla birlikte kanser olmayan kaç hastanın, hastalığa sahip olmayan olarak doğru şekilde sınıflandırıldığı da büyük önem taşımaktadır. Bu nedenle CSI ve F metrikleri böyle bir sınıflandırma problemi için uygun bir değerlendirme metriği olmayacaktır.[9] Thapliyal ve Singh tarafından yapılan çalışmada, CSI'nın gerçek negatiflerin sayısının dikkate almamasından dolayı ön yargılı bir metrik olduğu belirtilmiştir. TSS, yanlış pozitiflerin oranına göre gerçek pozitiflerin oranını ifade eder ve gerçek pozitiflerin sayısı yanlış negatiflerin sayısından fazla olduğu sürece değeri pozitif olacaktır. $TSS = 0$ olması hiçbir becerinin olmadığını gösterir. Negatif değerler ters tahminleri temsil eder ve tüm evet tahminlerinin hayır ile değiştirilmesiyle pozitif beceriye dönüştürülebilir veya bunun tersi de geçerlidir. Tahmin doğrulama metriklerinden TSS ve HSS yaklaşımları karşılaştırılmış ve nadir olayları tahmin etmede HSS metriğinin TSS metriğinden daha üstün olduğu sonucuna ulaşılmıştır.[11] Hogan ve Mason tarafından yapılan çalışmada, iki sonuçlu olaylara yönelik deterministik tahmin becerisinin doğrulanması alanında kullanılan 18 farklı metrik, meta-doğrulama kriterleri dikkate alınarak karşılaştırılmış ve PSS metriğinin diğer metriklere tercih edilmesi gerektiği sonucuna ulaşılmıştır.[8] Sitthiyot ve Holasut tarafından yapılan çalışmada, PSI metriğinin nasıl çalıştığını göstermek için 24 farklı veri seti kullanılmıştır. Sonuçlar, PSI'nın her zaman aynı değeri tahmin etmek için aynı sonucu vermekle kalmayıp, aynı zamanda nadir veya aşırı olayların doğru tahmini için de önemsiz olmayan sonuçlar verdiğini göstermiştir. Üstelik PSI, farklı beceri sonuçları vererek nadir veya aşırı olayların mükemmel tahmini ile rastgele olayların tahmini arasındaki farkı ayırt edebilmektedir.[7] CSI, ilgilenilen olayın tahmin edilmesi, gözlemlenmesi ya da her ikisinin birden olması koşuluyla, gerçek pozitiflerin koşullu olasılık tahmini olarak kabul edilebilir. Gerçek negatiflerin sayısına herhangi bir bağımlılık içermemesi nadir olaylar için bir performans metriği olarak yaygın şekilde kullanılmasına yol açmıştır. Gerçek negatiflerin sayısı önemli hâle gelebileceği için çok yaygın olaylar için kesinlikle uygun değildir.[8] Tüm bu açıklamalardan da anlaşılacağı üzere literatürde yer alan tartışmalar daha çok 2x2 çapraz tablolarda yer alan değerleri (gerçek pozitif, yanlış pozitif, yanlış negatif, gerçek negatif) dikkate alan tartışmalardır. Literatürde örneklem büyüklüğü ile tahmin-doğrulama metrikleri arasındaki ilişkinin doğrudan incelendiği bir çalışma bulunmamaktadır. Dolayısıyla çalışmadan elde edilen sonuçların tartışılması mümkün olmamıştır.

SONUÇ

Farklı örneklem büyüklüklerinin dikkate alındığı bu çalışmadan elde edilen sonuçlardan, örneklem büyüklüğünün artması ya da azalmasının tahmin doğrulama metriklerinin yorumlama seviyeleri için etkisinin neredeyse sabit olduğu belirlenmiştir. Tahmin doğrulama, örneklem büyüklüğünden ziyade gözlem değerlerinin 2x2 çapraz tablolardaki gözelere dağılımından daha çok etkilenmektedir. Dolayısıyla yapılacak çalışmalarda, 2x2 çapraz tabloların gözelerinde yer alan tüm değerlerin eşit olduğu, gerçek pozitif ve gerçek negatiflerin

toplam sayısının, yanlış pozitif ve yanlış negatiflerin toplam sayısından fazla olduğu ya da bunun tam tersi durumlar vb. dikkate alınmalıdır. Dikkate alınan tüm örneklem büyüklükleri için metriklerin hemen hemen tamamında tahminlerin yaklaşık %50'sinin "doğrulama yok ya da önemsiz" yorumlama seviyelerine sahip oldukları görülmüştür. Farklı örneklem büyüklüklerinden elde edilen metrik değerleri adil, orta, önemli ve mükemmel yorumlama seviyeleri birlikte dikkate alındığında GSS metriğinin diğer metriklere göre en tutucu, F metriğinin ise en liberal metrik olduğu sonucuna ulaşılmıştır.

Finansal Kaynak

Bu çalışma sırasında, yapılan araştırma konusu ile ilgili doğrudan bağlantısı bulunan herhangi bir ilaç firmasından, tıbbi alet, gereç ve malzeme sağlayan ve/veya üreten bir firma veya herhangi bir ticari firmadan, çalışmanın değerlendirme sürecinde, çalışma ile ilgili verilecek kararı olumsuz etkileyebilecek maddi ve/veya manevi herhangi bir destek alınmamıştır.

Çıkar Çatışması

Bu çalışma ile ilgili olarak yazarların ve/veya aile bireylerinin çıkar çatışması potansiyeli olabilecek bilimsel ve tıbbi komite üyeliği veya üyeleri ile ilişkisi, danışmanlık, bilirkişilik, herhangi bir firmada çalışma durumu, hissedarlık ve benzer durumları yoktur.

Yazar Katkıları

Bu çalışma hazırlanırken tüm yazarlar eşit katkı sağlamıştır.

KAYNAKLAR

1. Wheatcroft E. Interpreting the skill score form of forecast performance metrics. Int J Forecast. 2019;35(2):573-9. chrome-extension://efaidnbmnnnibpcajpcglclefindmkaj/https://www.lse.ac.uk/CATS/Assets/PDFs/Publications/Papers/2019/Interpreting-the-skill-score-form-of-forecast-performance-metrics.pdf
2. Roeger C, Stull R, McClung D, Hacker J, Deng X, Modzelewski H. Verification of mesoscale numerical weather forecasts in mountainous terrain for application to avalanche prediction. WAF. 2003;18(6):1140-60. https://www.researchgate.net/publication/242140599_Verification_of_Mesoscale_Numerical_Weather_Forecasts_in_Mountainous_Terrain_for_Application_to_Avalanche_Prediction
3. Jolliffe IT, Stephenson DB. Epilogue: New directions in forecast verification. In: Jolliffe IT, Stephenson DB, eds. Forecast Verification: A Practitioner's Guide in Atmospheric Science. 2nd ed. West Sussex: John Wiley and Sons; 2012. p.221-30.
4. Murphy AH, Winkler RL. A general framework for forecast verification. MWR. 1987;115(7):1330-8. chrome-extension://efaidnbmnnnibpcajpcglclefindmkaj/https://www.spaceweather.gov/sites/default/files/images/u30/Murphy%2C%20A.H.%2C%20and%20R.L.%20Winkler%2C%201987.pdf
5. Allouche O, Tsoar A, Kadmon R. Assessing the accuracy of species distribution models: prevalence, kappa and the true skill statistic (TSS). J Appl Ecol. 2006;43(6):1223-32. https://besjournals.onlinelibrary.wiley.com/doi/10.1111/j.1365-2664.2006.01214.x
6. Liu C, White M, Newell G. Measuring the accuracy of species distribution models: a review. 18th World IMACS / MODSIM Congress, Cairns, Australia 13-17 July 2009. https://www.researchgate.net/publication/288962465_Measuring_the_accuracy_of_species_distribution_models_a_review
7. Sitthiyot T, Holasut K. On the evaluation of skill in binary forecast. TWE. 2022;40(3):33-54. chrome-extension://efaidnbmnnnibpcajpcglclefindmkaj/https://arxiv.org/pdf/2209.04686
8. Hogan R, Mason I. Deterministic forecasts of binary events. In: Jolliffe IT, Stephenson DB, eds. Forecast Verification: a Practitioner's Guide in Atmospheric Science. 2nd ed. West Sussex: John Wiley and Sons; 2012. p.31-59.
9. Christen P, Hand DJ, Kirielle N. A review of the F-measure: its history, properties, criticism, and alternatives. ACM Comput Surv. 2023;56(3):1-24. chrome-extension://efaidnbmnnnibpcajpcglclefindmkaj/https://dl.acm.org/doi/pdf/10.1145/3606367
10. Mbizvo GK, Simpson CR, Duncan SE, Chin RFM, Larner AJ. Critical success index or F measure to validate the accuracy of administrative healthcare data identifying epilepsy in deceased adults in Scotland. Epilepsy Res. 2024;199:107275. PMID: 38128202.
11. Thapliyal R, Singh B. Heavy rainfall forecasting for Dehradun capital city during monsoon season 2020. MAUSAM. 2023;74(1):141-50. https://mausamjournal.imd.gov.in/index.php/MAUSAM/article/view/4951/5582
12. Guastavino S, Piana M, Benvenuto F. Bad and good errors: value-weighted skill scores in deep ensemble learning. IEEE Trans Neural Netw Learn Syst. 2024;35(2):1993-2002. PMID: 35776819.
13. Yoon S, Lee WH. Application of true skill statistics as a practical method for quantitatively assessing CLIMEX performance. Ecol Indic. 2023;146(6):109830. https://www.sciencedirect.com/science/article/pii/S1470160X22013036
14. McHugh ML. Interrater reliability: the kappa statistic. Biochem Med (Zagreb). 2012;22(3):276-82. PMID: 23092060; PMCID: PMC3900052.
15. Murphy AH. The Finley affair: a signal event in the history of forecast verification. WAF. 1996;11(1):3-20. chrome-extension://efaidnbmnnnibpcajpcglclefindmkaj/https://www.spaceweather.gov/sites/default/files/images/u30/Murphy%2C%20A.H.%2C%201996%20-%20The%20Finely%20Affair.pdf

16. Roebber PJ. Visualizing multiple measures of forecast quality. WAF. 2009;24(2):601-8. https://journals.ametsoc.org/view/journals/wefo/24/2/2008waf2222159_1.xml
17. Schaefer JT. The critical success index as an indicator of warning skill. WAF. 1990;5(4):570-5. https://journals.ametsoc.org/view/journals/wefo/5/4/1520-0434_1990_005_0570_tcsiaa_2_0_co_2.xml
18. Powers DM. What the F-measure doesn't measure: features, flaws, fallacies and fixes. ArXiv. 2015;abs/1503.06410. https://arxiv.org/abs/1503.06410
19. Jolliffe IT. The Dice coefficient: a neglected verification performance measure for deterministic forecasts of binary events. Meteorol Appl. 2016;23(1):89-90. https://rmets.onlinelibrary.wiley.com/doi/10.1002/met.1532