ORİJİNAL ARAŞTIRMA ORIGINAL RESEARCH

DOI: 10.5336/biostatic.2024-105129

Çarpık Verilerin Dönüşüm Teknikleri ile İyileştirilmesi ve AUC Değerleri Üzerindeki Etkiler: Bir Simülasyon Çalışması

Improving Skewed Data Using Transformation Techniques and Impact on AUC Values: A Simulation Study

Gülcan GENCER[a]

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

ÖZET Amaç: Bu çalışmanın amacı, sağa ve sola çarpık sağlık verilerine uygulanan çeşitli veri dönüşümlerinin alıcı işletim karakteristiği [receiver operating characteristic (ROC)] eğrisi altında kalan alan [area under the curve (AUC)] üzerindeki etkilerini incelemektir. Özellikle, logaritmik, karekök, Box-Cox, Yeo-Johnson, Quantile, Rank, Robust Scale ve Inverse dönüşümlerin çarpık veri setlerinde AUC değerlerini nasıl etkilediği araştırılmıştır. **Gereç ve Yöntemler:** Çalışma kapsamında, farklı örneklem büyüklüklerinde sağa ve sola çarpık veri setleri oluşturulmuş ve bu veri setlerine farklı dönüşüm teknikleri uygulanmıştır. Her bir veri seti için AUC değerleri hesaplanmış ve çeşitli dönüşüm tekniklerinin bu değerlere etkisi simülasyon çalışmaları ile incelenmiştir. Python programlama dili kullanılarak 50≤n≤500 aralığında yer alan farklı n değerleri için veri türetilmiştir. **Bulgular:** Sağa çarpık verilere uygulanan dönüşüm teknikleri arasında, Quantile dönüşümü küçük örneklem boyutlarında yüksek AUC değerleri sağlamıştır. Logaritmik, karekök, Box-Cox ve Yeo-Johnson dönüşümleri ise çarpıklığı azaltarak benzer performans göstermiştir. Inverse dönüşümü küçük örneklemlerde düşük AUC değerleriyle etkisiz kalmıştır. Sola çarpık verilerde, Quantile dönüşümü küçük örneklemlerde etkili olurken, Inverse dönüşümü büyük örneklemlerde en iyi AUC değerini vermiştir. Box-Cox ve Yeo-Johnson dönüşümleri ise sola çarpık verilerde çarpıklığı azaltarak daha dengeli AUC değerleri elde edilmesini sağlamıştır. **Sonuç:** Bu çalışma, çarpık verilerin AUC değerleri üzerindeki etkilerini değerlendirerek, sağlık verisi analistlerine uygun dönüşüm tekniklerini seçme konusunda pratik rehberlik sunmaktadır. Quantile dönüşümü küçük örneklem boyutlarında sağa ve sola çarpık verilerde genellikle etkili bir yöntem olarak öne çıkarken, Inverse dönüşümü özellikle büyük örneklem boyutlarında sola çarpık verilerde etkili olabilir. Bu bulgular, çarpık veri setlerinde model performansını iyileştirmek için dönüşüm tekniklerinin dikkatli bir şekilde seçilmesi gerektiğini vurgulamaktadır.

Anahtar kelimeler: Veri dönüşümü; ROC eğrisi; eğri altındaki alan; çarpık veri; sağlık verisi analizi

ABSTRACT Objective: The purpose of this study is to investigate the effects of various data transformations applied to right and left skewed health data on the area under the receiver operating characteristic (ROC) curve (AUC). In particular, it was investigated how logarithmic, square root, Box-Cox, Yeo-Johnson, Quantile, Rank, Robust Scale and Inverse transformations affect AUC values in skewed data sets. **Material and Methods:** Within the scope of the study, right and left skewed data sets were created with different sample sizes and different transformation techniques were applied to these data sets. AUC values were calculated for each data set and the effects of various transformation techniques on these values were investigated with simulation studies. Data were generated for different n values in the range of 50≤n≤500 using the Python programming language. **Results:** Among the transformation techniques applied to right-skewed data, Quantile transformation provided high AUC values in small sample sizes. Logarithmic, Square Root, Box-Cox and Yeo-Johnson transformations showed similar performance by reducing skewness. Inverse transformation was ineffective with low AUC values in small samples. In left-skewed data, Quantile transformation was effective in small samples, while Inverse transformation gave the best AUC value in large samples. Box-Cox and Yeo-Johnson transformations provided more balanced AUC values by reducing skewness in left-skewed data. **Conclusion:** This study evaluates the effects of skewed data on AUC values and provides practical guidance to healthcare data analysts on selecting appropriate transformation techniques. Quantile transformation is generally effective for right and left skewed data in small sample sizes, while Inverse transformation can be especially effective for left skewed data in large sample sizes. These findings emphasize that transformation techniques should be carefully selected to improve model performance in skewed datasets.

Keywords: Data transformation; ROC curve; area under the curve; skewed data; health data analysis

KAYNAK GÖSTERMEK İÇİN:
Gencer G. Çarpık verilerin dönüşüm teknikleri ile iyileştirilmesi ve AUC değerleri üzerindeki etkiler: Bir simülasyon çalışması. Turkiye Klinikleri J Biostat. 2024;16(3):157-67.

Correspondence: Gülcan GENCER
Afyonkarahisar Sağlık Bilimleri Üniversitesi Tıp Fakültesi, Biyoistatistik ve Tıbbi Bilişim AD, Afyonkarahisar, Türkiye
E-mail: gulcan.gencer@afsu.edu.tr

Peer review under responsibility of Turkiye Klinikleri Journal of Biostatistics.

Received: 14 Aug 2024 ***Received in revised form:*** 12 Nov 2024 ***Accepted:*** 12 Nov 2024 ***Available online:*** 06 Dec 2024

2146-8877 / Copyright © 2024 by Türkiye Klinikleri. This in an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

Sağlık bilimleri alanında, veri analizi hasta sonuçlarını anlama ve iyileştirmede önemli bir rol oynar. Ancak, veri analistlerinin karşılaştığı önemli zorluklardan biri, özellikle tıbbi veri kümelerinde yaygın olan çarpık verilerin varlığıdır. Çarpık veriler, ister sağa ister sola çarpık olsun, uygun şekilde ele alınmadığında ön yargılı istatistiksel tahminlere ve yanlış çıkarımlara yol açabilir.[1,2] Bu tür verilerin analizi, özellikle doğruluk gerektiren tıbbi kararlar ve hasta tedavi süreçlerinde kritik bir öneme sahiptir.[3] Sağlık verilerinin karmaşıklığı ve çeşitliliği, veri dönüşüm tekniklerinin önemini artırır, çünkü bu teknikler verilerin dağılımını iyileştirerek model performansını artırabilir.[4] Ayrıca, doğru veri ön işleme adımlarının uygulanması, modelleme süreçlerinde ön yargıların ve hatalı çıkarımların önüne geçilmesinde hayati bir rol oynar.[5] Bu nedenle, sağlık bilimlerinde çarpık veri setlerinin doğru şekilde ele alınması, daha güvenilir ve geçerli sonuçların elde edilmesine katkı sağlar. Sağa çarpık veriler genellikle daha yüksek değerlerden oluşan uzun bir kuyruk olduğunda ortaya çıkar; bu, hastanede kalış süreleri, maliyet verileri veya küçük bir hasta alt kümesinde yükselmiş olabilecek biyobelirteç seviyeleri gibi değişkenlerde yaygındır. Tersine, sola çarpık veriler, belirli koşullarda tanı yaşı veya birkaç uç değerle ağırlıklı olarak kısa olan iyileşme süreleri gibi durumlarda görüldüğü gibi, daha düşük değerlerden oluşan uzun bir kuyruk olduğunda ortaya çıkar.[6,7] Bu çarpık dağılımlar, öngörücüler ile sonuçlar arasındaki ilişkiyi bozabilir ve bu da optimum olmayan model performansına ve yanlış risk tabakalaşmasına yol açabilir.[8] Bu çalışmanın amacı, sağlık veri kümelerindeki çarpıklığı düzeltmek için çeşitli veri dönüştürme tekniklerini keşfetmek ve bu tekniklerin alıcı işletim karakteristiği [receiver operating characteristic (ROC)] eğrisinin eğri altındaki alan [area under the curve (AUC)] üzerindeki etkilerini değerlendirmektir. Bu eğriler, tanı ve prognoz modellerinin performansını değerlendirmek için yaygın olarak kullanılan bir ölçüttür.[9,10] Özellikle, bu çalışma logaritmik, karekök, Box-Cox, Yeo-Johnson, Kuantil (Quantile) ve Ters (Inverse) dönüşümlerin sağa çarpık ve sola çarpık veriler üzerindeki etkilerine odaklanmıştır.[11,12]

Bu dönüşümler uygulanarak ve AUC üzerindeki etkileri değerlendirilerek, biyoistatistikçilere ve sağlık verisi bilimcilerine analizlerinde çarpık verileri ele alma konusunda pratik rehberlik sağlanacaktır. Bu yöntem, özellikle tahmin modellerinin doğruluğunun klinik karar alma ve hasta sonuçlarını doğrudan etkileyebileceği tıbbi araştırma bağlamında önemlidir.[6,8,13]

Bu çalışmanın amacı, sağa ve sola çarpık sağlık verileri üzerinde uygulanan çeşitli veri dönüşümlerinin ROC eğrisi altında kalan alan değerlerine olan etkilerini incelemektir. Özellikle, sağa ve sola çarpık verilerde uygulanan logaritmik, karekök, Box-Cox, Yeo-Johnson, Quantile, Rank, Robust Scale ve Inverse dönüşümlerin AUC değerlerini nasıl etkilediğini araştırmak, bu dönüşümlerin biyoistatistiksel analizlerdeki önemini ortaya koymayı amaçlamaktadır. Çalışmada, her bir dönüşümün sağladığı AUC değerlerinin karşılaştırması yapılmış ve sağlık verilerinde hangi dönüşümün daha uygun olduğu değerlendirilmiştir.

GEREÇ VE YÖNTEMLER

Bu çalışmada, farklı veri setleri üzerinde uygulanan çeşitli dönüşüm tekniklerinin etkisini araştırmak amacıyla bir simülasyon çalışması gerçekleştirilmiştir. Çalışma kapsamında sağa çarpık, sola çarpık ve dengeli veri setleri kullanılmıştır. Sağa ve sola çarpık veri setleri, Python programlama dili kullanılarak rastgele türetilmiştir. Bu bağlamda, sağa ve sola çarpık veriler, scipy.stats kütüphanesindeki skewnorm.rvs fonksiyonu kullanılarak türetilmiştir. Bu fonksiyon, çarpıklık parametresi olarak tanımlanan "a" değerine göre $Z=a*X$ formülü üzerinden bir çarpıklık etkisi ekleyerek veriyi türetir. Burada; Z, çarpıklık etkisi uygulanmış, yani çarpık dağılıma sahip olacak şekilde dönüştürülmüş veri değeridir. Fonksiyon, sağa ve sola çarpıklık oluşturmak için $f(x|a)=2*\varphi(x)*\varphi(a*x)$ formülünü temel alır. Burada $\varphi(x)$, ortalaması 0, varyansı 1 olan normal dağılımın olasılık yoğunluk fonksiyonudur. $\varphi(a*x)$, çarpıklık parametresi "a" ile ayarlanan kümülatif dağılım fonksiyonudur. Pozitif a değeri sağa çarpıklık, negatif a değeri ise sola çarpıklık oluşturur; çarpıklık parametresi pozitif olduğunda sağa çarpık, negatif olduğunda ise sola çarpık bir dağılım oluşturmaktadır. Örneğin, a=10 değeri sağa çarpık bir dağılım üretirken, a=-10 değeri sola çarpık bir dağılım oluşturur. Veri setlerinin her biri için ROC eğrisi altında kalan alan değerleri hesaplanmış ve çeşitli dönüşüm tekniklerinin bu değerlere etkisi incelenmiştir. Simülasyon adımları kısaca şöyle özetlenebilir.

Adım 1: Veri setleri belirli parametrelere göre türetilmiştir. Sağa çarpık veri seti, skewnorm.rvs (a=10, size=n_samples) fonksiyonu ile üretilmiştir. Sola çarpık veri seti ise skewnorm.rvs (a=-10, size=n_samples) fonksiyonu ile türetilmiştir. Dengeli veri seti, norm.rvs (size=n_samples) fonksiyonu ile standart normal dağılımdan üretilmiştir. Çarpık veri setleri için her bir veri noktası, ikili sınıflandırma amacıyla rastgele 0 veya 1 olarak etiketlenmiştir. Bu sınıflandırma, parametreli Bernoulli dağılımı temel alınarak yapılmıştır. Bernoulli dağılımı, her bir gözlem için iki olasılıktan birini alma (0 veya 1) üzerine kurulu olup, başarı olasılığı p=0,5 olarak belirlenmiştir. Bu sayede, her bir gözlem için başarı veya başarısızlık olarak iki olasılık sağlanmıştır. Böylece sağa ve sola çarpık veri setlerinin ikili grup değişkenleri eşit olasılıkla rastgele türetilmiştir.

Adım 2: Çeşitli dönüşüm teknikleri kullanılarak veri setleri üzerinde dönüşümler gerçekleştirildi.

Adım 3: Parametrik olmayan bir yöntem olan Mann-Whitney U istatistiği, her bir dönüşüm sonrası veri setleri için AUC değerlerini hesaplamak için ROC eğrisi altında kalan alanı tahmin etmek amacıyla kullanıldı.

Adım 4: Her bir veri seti ve dönüşüm için elde edilen AUC değerleri kaydedildi ve karşılaştırıldı.

Adım 5: Grafikler kullanılarak, AUC değerleri görselleştirildi ve dönüşümlerin etkileri görsel olarak analiz edildi.

Adım 6: Burada bahsedilen tüm simülasyonlar 10.000 tekrar üzerinden ve çeşitli örnek hacimleri dikkate alınarak gerçekleştirilmiştir. Her veri seti ve dönüşüm tekniği için bu 10.000 tekrardan elde edilen AUC değerlerinin ortalaması sunulmuştur. Simülasyon çalışmaları Python programlama dili ve Scipy kütüphanesi kullanılarak gerçekleştirilmiştir.

Bu çalışmada kullanılan dönüşüm teknikleri aşağıdaki gibidir:

1. Log Dönüşümü: $X' = log(X+1)$ (1)

2. Karekök Dönüşümü: $X' = \sqrt{X}$ (2)

3. Box-Cox Dönüşümü: Box-Cox dönüşümü George E.P. Box ve David Cox tarafından önerilen λ parametreli üstel bir dönüşümdür. Dönüşüm eşitlik (3)'te verildiği gibi tanımlanmıştır. Lambda değeri, en uygun çarpıklığı düzeltmek amacıyla en uygun olacak şekilde optimize edilerek belirlenmiştir.[6]

$$x_i^{(\lambda)} = \begin{pmatrix} \frac{x_i^{\lambda}-1}{\lambda} \; eğer \; \lambda \neq 0 \\ ln(x_i) \; eğer \; \lambda = 0 \end{pmatrix} \quad (3)$$

4. Yeo-Johnson Dönüşümü: Transformasyonlarında çok sık kullanılan yöntemlerden biri de Yeo-Johnson dönüşümüdür. Dönüşüm eşitlik (4)'te verildiği gibi tanımlanmıştır. Lambda değeri, en uygun çarpıklığı düzeltmek amacıyla en uygun olacak şekilde optimize edilerek belirlenmiştir.[14]

$$x_i^{(\lambda)} = \begin{pmatrix} \frac{[(x_i+1)^{\lambda}-1]}{\lambda} \; eğer \; \lambda \neq 0, x_i \geq 0 \\ ln(x_i+1) \; eğer \; \lambda = 0, \; x_i \geq 0 \\ -\frac{[(-x_i+1)^{2-\lambda}-1]}{2-\lambda} \; eğer \; \lambda \neq 2, x_i < 0 \\ -ln(-x_i+1) \; eğer \; \lambda = 2, \; x_i < 0 \end{pmatrix} \quad (4)$$

5. Ters (Inverse) Dönüşüm: $X' = \frac{1}{X+1}$ (5)

6. Kuantil (Quantile) Dönüşüm: Quantile dönüşümü, orijinal veri dağılımını verilerin sıralarına göre tekdüze veya normal bir dağılıma eşler. Bu çalışmada normal dağılım kullanılmıştır. Öncelikle, verilerin ampirik kümülatif dağılım fonksiyonu hesaplanır. Orijinal verilerin sıraları (veya kuantillerini) tekdüze veya normal bir dağılımla eşlenir. Dönüşüm eşitlik (6)'da verildiği gibi tanımlanmıştır.

$$X' = \varphi^{-1}\big(F(x)\big) \quad (6)$$

Burada, X' orijinal veri seti, $F(x)$ ampirik kümülatif dağılım fonksiyonu, φ^{-1} standart normal dağılımın ters kümülatif dağılım fonksiyonudur.[15]

7. Rank Dönüşümü: Rank dönüşümü, bir veri kümesindeki her veri noktasına bir sıra atar. Bu dönüşüm, esasen verileri sıralarıyla değiştirir, sırayı korur ancak orijinal değerleri korumaz. Sıralamalar genellikle eşitlik (7)'de verildiği gibi hesaplanır:

$$X_i' = Rank(X_i) \quad (7)$$

X_i orijinal veri setindeki i'nci veri noktasıdır, $Rank(X_i)$, tüm veri noktaları arasında X_i'nin sıralamasıdır.[16]

8. Robust Scaling: Robust Scaling, klasik standartlaştırma yerine sağlam konum ve ölçek tahmincileri kullanılarak gözlem değerlerinin standartlaştırılması için kullanılan yöntemdir. Dönüşüm eşitlik (8)'de verildiği gibi tanımlanmıştır.

$$X' = \frac{X - Median(X)}{IQR(X)} \quad (8)$$

X orijinal veriyi, $Median(X)$ verinin medyanını, $IQR(X)$ çeyrekler arası aralığı ifade eder.[17]

AUC HESAPLAMA YÖNTEMİ

Parametrik olmayan yöntem olan Mann-Whitney U testi AUC hesaplama yöntemi olarak kullanılmıştır. İlk defa Bamber tarafından önerilen Mann-Whitney U istatistiği ile ROC eğrisi altında kalan alan tahmini eşitlik (9) ile elde edilmiştir.[18]

$$A = \frac{U}{m*n} \quad (9)$$

Burada U, Mann Whitney istatistiği, m ve n ise hasta ve kontrol gruplarındaki birey sayısını göstermektedir. Her dönüşüm için ayrı ayrı AUC değerleri hesaplanmış ve sonuçlar karşılaştırılmıştır. Elde edilen AUC değerleri, görselleştirme için farklı grafik türleriyle sunulmuştur. Bu yöntemlerle elde edilen sonuçlar, sağlık bilimlerinde karşılaşılan sağa ve sola çarpık veriler üzerinde uygun dönüşümler uygulanarak model performansının nasıl iyileştirilebileceğini göstermektedir. Mann-Whitney U testi, parametrik olmayan bir yöntem olarak iki bağımsız grup arasındaki sıralı farklılıkları ölçen bir testtir ve özellikle ROC eğrisinin altında kalan alanın parametrik olmayan bir tahmini olarak yaygın şekilde kullanılmaktadır.[18]

Bu çalışma, Helsinki Deklarasyonu prensiplerine uygun olarak gerçekleştirilmiştir.

BULGULAR

Tablo 1, farklı örneklem büyüklüklerinde sağa çarpık verilere uygulanan çeşitli dönüşümler sonrası elde edilen çarpıklık katsayıları, orijinal veriler ile AUC değerlerini göstermektedir. Çarpıklık katsayıları, verinin asimetrikliğini yansıtır ve dönüşüm teknikleriyle bu çarpıklık önemli ölçüde azaltılabilir. Örneğin, orijinal verinin çarpıklık katsayısı genellikle yüksektir ve dönüşüm teknikleri (özellikle Box-Cox ve Yeo-Johnson) bu değeri belirgin şekilde düşürmektedir. Bu durum, verinin dağılımını daha simetrik hâle getirerek model performansını (AUC) iyileştirme potansiyeline sahip olduğunu göstermektedir. Özellikle, Log, Box-Cox, ve Yeo-Johnson dönüşümleri, çarpıklığı en etkili şekilde azaltan dönüşümler olarak öne çıkmaktadır. Örneklem büyüklüğünün n=50 olması durumunda, en yüksek AUC değeri, Quantile dönüşümüyle elde edilmiştir (0,6694). Bu durum, Quantile dönüşümünün küçük örneklem boyutlarında sağa çarpık verileri normalize etmede etkili olabileceğini gösterir. Inverse dönüşümü, daha düşük bir AUC değeri (0,4900) ile diğer dönüşümlerin gerisinde kalmaya devam etmektedir; logaritmik, karekök, Box-Cox, ve Yeo-Johnson dönüşümleri benzer AUC değerleri sunmaktadır ve bu dönüşümler, sağa çarpık verilerin dağılımını iyileştirmede birbirine yakın performans sergilemektedir. Inverse dönüşümü, oldukça düşük bir AUC değeri (0,3346) vermiştir, bu

da bu dönüşümün sağa çarpık verilerde etkin olmadığını işaret eder. Örneklem büyüklüğünün n=75 olması durumunda, yine en yüksek AUC değeri Quantile dönüşümünden elde edilmiştir (0,6893). Bu, Quantile dönüşümünün daha büyük örneklem boyutlarında da etkili olduğunu gösterir.

Diğer dönüşümler arasında, logaritmik, karekök, Box-Cox ve Yeo-Johnson benzer AUC değerleri göstermektedir, bu da bu yöntemlerin sağa çarpık verilerin dönüştürülmesinde etkili olduğunu teyit eder. Inverse dönüşümü yine düşük bir AUC değeri (0,3520) ile en düşük performansı sergilemiştir. Örneklem büyüklüğünün n=100 olması durumunda, Log ve Quantile dönüşümleri nispeten yüksek AUC değerleri sunmaktadır. Özellikle Quantile dönüşümü (0,5615), verilerin dağılımını iyileştirmede yine öne çıkmaktadır.

Inverse dönüşümü, önceki örneklem boyutlarına benzer şekilde, düşük bir AUC değeri (0,4318) ile düşük performans göstermektedir. Örneklem büyüklüğünün n=250 olması durumunda, Inverse dönüşümü diğer dönüşümlerden daha yüksek bir AUC değeri (0,5208) elde etmiştir. Bu durum, Inverse dönüşümünün daha büyük örneklem boyutlarında sağa çarpık veriler üzerinde nispeten daha iyi performans gösterebileceğini işaret edebilir.

Diğer dönüşümler oldukça benzer AUC değerleri sunmakta, bu da bu dönüşümlerin etkinliğini büyük örneklem boyutlarında koruduğunu göstermektedir (Tablo 1).

Tablo 2'de farklı örneklem boyutları ve dönüşüm teknikleri için çarpıklık katsayıları, orijinal veriler ve AUC değerleri sunulmuştur. Çarpıklık katsayıları, veri setinin asimetrikliğini gösterir ve pozitif veya negatif değerler alabilir. Örneklem büyüklüğünün n=50 olması durumunda, orijinal veri ve Robust Scale dönüşümü aynı AUC değerini vermiştir (0,4930), bu da dönüşümlerin bu boyutta anlamlı bir iyileşme sağlamadığını gösterir. Box-Cox ve Yeo-Johnson dönüşümleri de nispeten düşük AUC değerleri sunmaktadır, ancak logaritmik, karekök ve Inverse dönüşümleri belirsiz sonuçlar vermiştir, bu da bu dönüşümlerin küçük örneklem boyutlarında etkili olmadığını veya uygulanamadığını gösterir. Quantile dönüşümü ise (0,4612), diğer dönüşümlerden daha düşük bir AUC değeri sağlamasına rağmen Rank dönüşümünden (0,4298) daha iyi performans göstermektedir.

Örneklem büyüklüğünün n=50 olması durumunda, Quantile dönüşümü en yüksek AUC değerini (0,6012) elde etmiştir. Bu durum, Quantile dönüşümünün sola çarpık veriler üzerinde küçük ve orta boy örneklem setlerinde etkili olabileceğini gösterir. Log ve karekök dönüşümleri ise AUC değerlerinde sırasıyla 0,5929 ve 0,5959 ile iyi performans sergilemiştir. Bu durum, bu dönüşümlerin bu örneklem boyutunda etkin olduğunu göstermektedir. Inverse dönüşümü ise diğer dönüşümlerle karşılaştırıldığında daha düşük bir AUC değeri (0,4066) ile zayıf performans göstermektedir. Örneklem büyüklüğünün n=100 olması durumunda, Inverse dönüşümü en yüksek AUC değerini (0,6014) elde etmiştir, bu da bu dönüşümün orta boy örneklem setlerinde sola çarpık verilerde etkili olabileceğini gösterir.

Quantile dönüşümü yine makul bir AUC değeri (0,4810) sunmaktadır. Log ve karekök dönüşümleri oldukça düşük AUC değerleri (0,3990 ve 0,3924) ile diğer dönüşümlere kıyasla daha düşük performans göstermiştir. Örneklem büyüklüğünün n=250 olması durumunda, Inverse dönüşümü en yüksek AUC değerini (0,5719) elde etmiştir. Bu durum, Inverse dönüşümünün daha büyük örneklem boyutlarında etkili olduğunu işaret eder.

Log ve karekök dönüşümleri diğer dönüşümlerle karşılaştırıldığında nispeten düşük AUC değerleri sunmaktadır (sırasıyla 0,4280 ve 0,4372). Box-Cox ve Yeo-Johnson dönüşümleri de benzer AUC değerleri sunmaktadır, ancak bu dönüşümler Inverse dönüşümüne kıyasla daha düşük performans göstermiştir. Örneklem büyüklüğünün n=500 olması durumunda, Inverse dönüşümü (0,5264) ve Box-Cox dönüşümü (0,5219) en yüksek AUC değerlerini elde etmişlerdir. Bu durum, bu dönüşümlerin büyük veri setlerinde sola çarpık veriler üzerinde etkili olduğunu gösterir. Log ve karekök dönüşümleri ise diğer dönüşümlere kıyasla daha düşük AUC değerleri (0,4733 ve 0,4772) ile zayıf performans sergilemiştir. Genel olarak, bu büyüklükteki örneklemler üzerinde tüm dönüşümler benzer performans göstermiş, ancak Inverse ve Box-Cox dönüşümleri öne çıkmıştır (Tablo 2). Ayrıca, farklı örneklem büyüklüklerinde çarpık verilere uygulanan dönüşümler ve AUC değerleri Şekil 1, Şekil 2, Şekil 3, Şekil 4 ve Şekil 5'te verilmiştir.

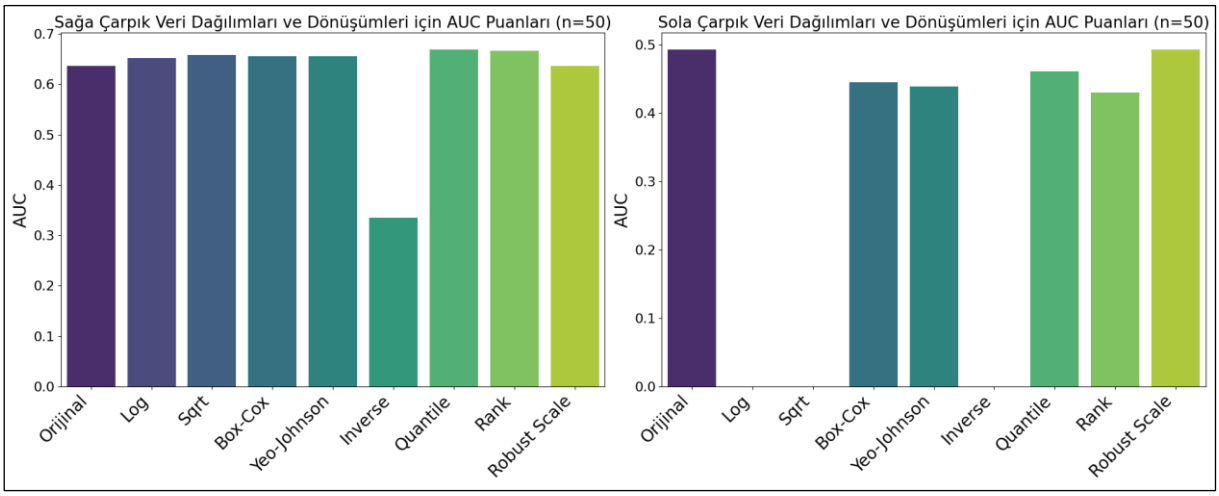
TABLO 1: Farklı örneklem büyüklüklerinde sağa çarpık verilere uygulanan dönüşümlere göre çarpıklık katsayıları ile AUC değerleri.

Örneklem boyutu	Dönüşümler	Çarpıklık katsayısı	AUC
n=50	Orijinal	0,5775	0,6369
	Log	0,2313	0,6520
	Karekök	0,0452	0,6583
	Box Cox	0,0324	0,6554
	Yeo Johnson	0,0265	0,6556
	Inverse	0,1651	0,3346
	Quantile	0,0000	**0,6694**
	Rank	0,0000	**0,6668**
	Robust Scale	0,5775	0,6369
n=75	Orijinal	0,8688	0,6893
	Log	0,3171	**0,6709**
	Karekök	0,1050	0,6584
	Box Cox	0,0369	0,6618
	Yeo Johnson	0,0346	0,6617
	Inverse	0,2001	0,3520
	Quantile	0,0000	0,6577
	Rank	0,0000	0,6495
	Robust Scale	0,8688	**0,6893**
n=100	Orijinal	0,9111	0,5521
	Log	0,4040	0,5606
	Karekök	0,1351	**0,5712**
	Box Cox	0,0753	**0,5664**
	Yeo Johnson	0,0747	**0,5664**
	Inverse	0,0380	0,4318
	Quantile	0,0000	0,5615
	Rank	0,0000	0,5629
	Robust Scale	0,9111	0,5521
n=250	Orijinal	1,1907	0,4794
	Log	0,3972	0,4794
	Karekök	0,0169	0,4809
	Box Cox	0,0335	0,4810
	Yeo Johnson	0,0288	0,4810
	Inverse	0,1928	**0,5208**
	Quantile	0,0000	0,4854
	Rank	0,0000	0,4805
	Robust Scale	1,1907	0,4794
n=500	Orijinal	1,0957	0,5137
	Log	0,3114	0,5108
	Karekök	0,0516	**0,5166**
	Box Cox	0,0327	**0,5122**
	Yeo Johnson	0,0282	**0,5123**
	Inverse	0,2582	0,4900
	Quantile	0,0000	0,5118
	Rank	0,0000	0,5068
	Robust Scale	1,0957	**0,5137**

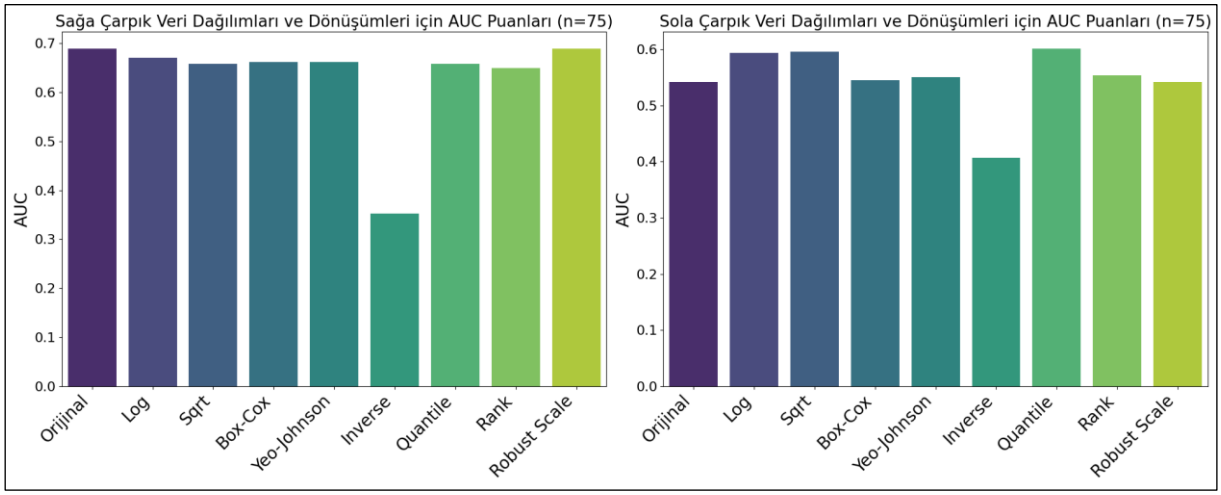
TABLO 2: Farklı örneklem büyüklüklerinde sola çarpık verilere uygulanan dönüşümlere göre çarpıklık katsayıları ile AUC değerleri.

Örneklem boyutu	Dönüşümler	Çarpıklık katsayısı	AUC
n=50	Orijinal	-1,2294	0,4930
	Log	-*	-*
	Karekök	-*	-*
	Box Cox	-0,20546	0,4451
	Yeo Johnson	-0,0370	0,4390
	Inverse	-*	-*
	Quantile	0,0000	0,4612
	Rank	0,0000	0,4298
	Robust Scale	-1,2294	**0,4930**
n=75	Orijinal	-0,3359	0,5420
	Log	5,6336	**0,5929**
	Karekök	4,9645	**0,5959**
	Box Cox	-0,1713	0,5445
	Yeo Johnson	0,0245	0,5500
	Inverse	-5,5481	0,4066
	Quantile	0,0000	**0,6012**
	Rank	0,0000	0,5531
	Robust Scale	-0,3359	0,5420
n=100	Orijinal	-0,9621	0,4791
	Log	5,3921	0,3990
	Karekök	4,6358	0,3924
	Box Cox	-0,1825	0,4749
	Yeo Johnson	-0,0244	0,4711
	Inverse	-5,3473	**0,6014**
	Quantile	0,0000	0,4810
	Rank	0,0000	0,4873
	Robust Scale	-0,9621	0,4791
n=250	Orijinal	-0,8191	0,4804
	Log	7,3843	0,4280
	Karekök	6,0755	0,4372
	Box Cox	-0,2070	0,4756
	Yeo Johnson	-0,0277	0,4763
	Inverse	-7,3034	**0,5719**
	Quantile	0,0000	0,4856
	Rank	0,0000	0,4734
	Robust Scale	-0,8191	0,4805
n=500	Orijinal	-0,9576	0,5228
	Log	8,7603	0,4733
	Karekök	6,5193	0,4772
	Box Cox	-0,1900	**0,5219**
	Yeo Johnson	-0,0252	0,5195
	Inverse	-8,5463	**0,5264**
	Quantile	0,0000	0,5167
	Rank	0,0000	0,5235
	Robust Scale	-0,9576	**0,5228**

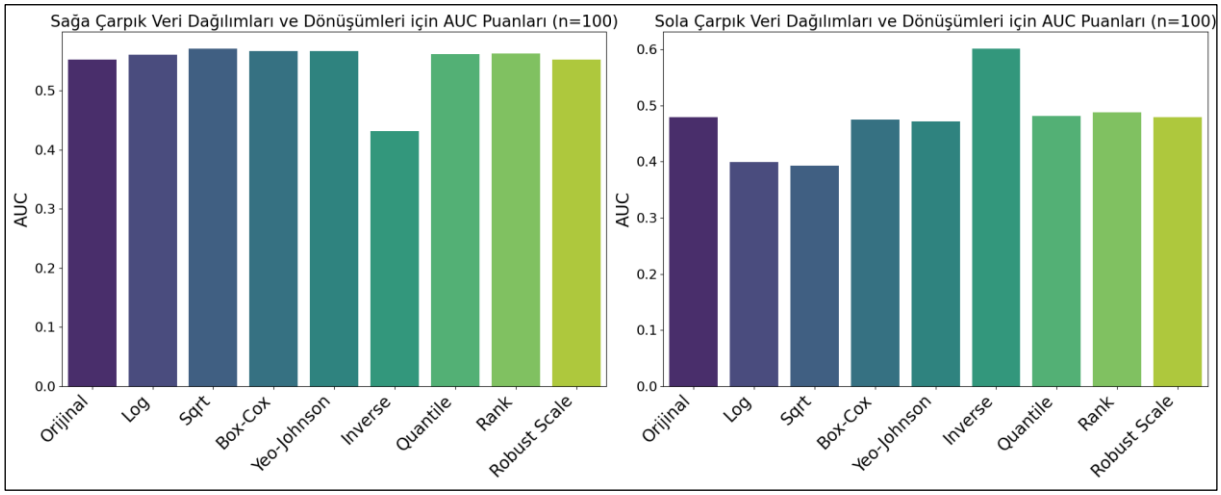
*Matematiksel olarak tanımsız veya geçersiz sonuçlar üretildiğinden veri elde edilemedi.



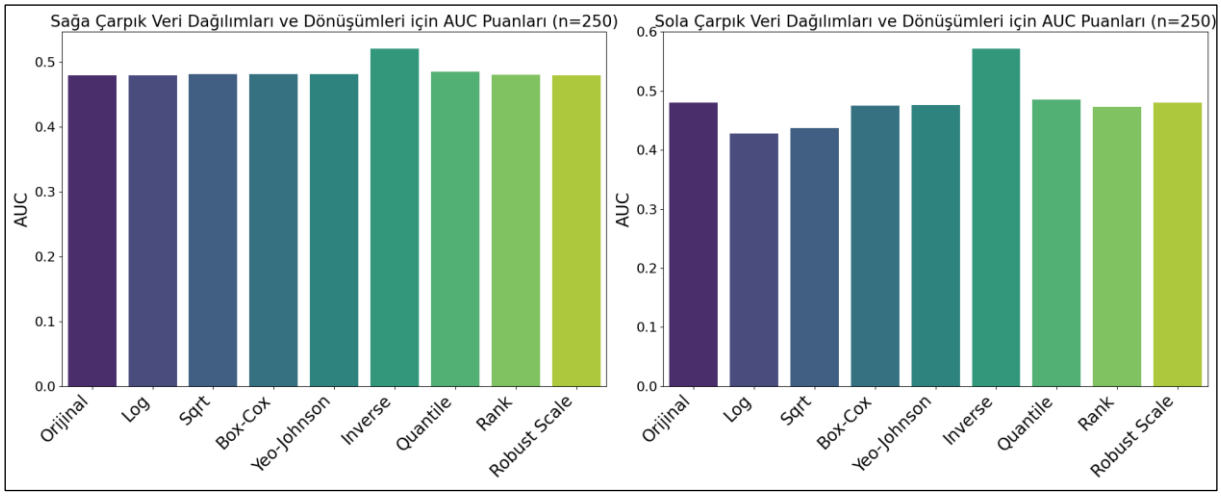
ŞEKİL 1: Farklı örneklem büyüklüklerinde çarpık verilere uygulanan dönüşümler ve AUC değerleri (n=50).



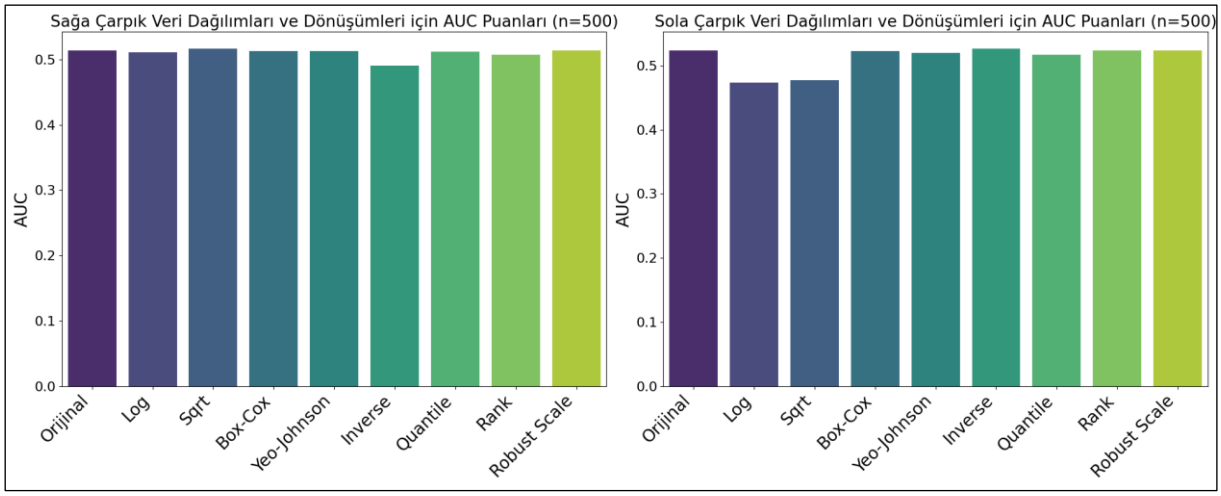
ŞEKİL 2: Farklı örneklem büyüklüklerinde çarpık verilere uygulanan dönüşümler ve AUC değerleri (n=75).



ŞEKİL 3: Farklı örneklem büyüklüklerinde çarpık verilere uygulanan dönüşümler ve AUC değerleri (n=100).



ŞEKİL 4: Farklı örneklem büyüklüklerinde çarpık verilere uygulanan dönüşümler ve AUC değerleri (n=250).



ŞEKİL 5: Farklı örneklem büyüklüklerinde çarpık verilere uygulanan dönüşümler ve AUC değerleri (n=500).

TARTIŞMA

Çalışmamızda sağa ve sola çarpık veriler üzerinde uygulanan çeşitli dönüşüm yöntemlerinin AUC değerleri üzerindeki etkileri incelenmiştir. Özellikle, küçük örneklem boyutlarında Quantile dönüşümünün sağa çarpık verilerde genellikle yüksek AUC değerlerini sağladığını tespit etmiş olmamız, bu dönüşüm yönteminin çarpık veri setleri üzerinde güçlü bir normalleştirici etkiye sahip olduğunu ortaya koymaktadır. Bu bulgu, Zhang ve Castelló tarafından yapılan araştırmalarla uyumlu olup, çarpık dağılımlar üzerinde veri dönüşümlerinin önemini bir kez daha vurgulamaktadır.[2]

Öte yandan, log, karekök, Box-Cox ve Yeo-Johnson dönüşümleri, özellikle orta ve büyük örneklem boyutlarında benzer performans sergileyerek, çarpık verilerin dağılımını iyileştirmede etkili olabilecek yöntemler olarak öne çıkmıştır. Bu dönüşümler, Friedman, Hastie ve ark. tarafından geliştirilen modelleme yöntemleri ile uyumlu olup, verilerin daha iyi analiz edilmesini ve sonuçların güvenilirliğini artırmayı hedefleyen biyoistatistiksel çalışmalar için kritik öneme sahiptir.[8]

Inverse dönüşümünün küçük örneklem boyutlarında düşük AUC değerleri ile özellikle sağa çarpık veriler üzerinde zayıf performans sergilemesi, bu dönüşüm yönteminin genellikle daha büyük örneklem boyutlarında etkili olabileceğini göstermektedir. Hosmer ve Lemeshow tarafından yapılan çalışmalar, Inverse dönüşümünün belirli koşullarda daha iyi sonuçlar verebileceğini belirtmekte olup, çalışmamız bu bulgularla paralellik göstermektedir.[7]

Son olarak, Box-Cox ve Yeo-Johnson dönüşümlerinin genel olarak dengeli performans sergilemesi, bu yöntemlerin hem sağa hem de sola çarpık veriler üzerinde güvenilir bir seçenek olduğunu ortaya koymaktadır. Çalışmanın bulguları da Ünal'ın çalışması ile uyumlu olarak bulunmuştur.[19] McCullagh ve Nelder tarafından da önerilen bu dönüşüm yöntemleri, veri dağılımlarının homojenliğini artırmak ve analizlerin doğruluğunu yükseltmek amacıyla biyoistatistik alanında sıkça başvurulan teknikler arasında yer almaktadır.[20]

Bu bulgular, sağa ve sola çarpık verilerle çalışırken uygun dönüşüm yöntemlerinin seçiminin model performansını nasıl etkileyebileceğini göstermekte ve araştırmacılara pratik rehberlik sunmaktadır. Literatürde bu problemin üstesinden gelmek için Arslan ve ark. tarafından, veri setlerine çeşitli matematiksel dönüşümler uygulayan web tabanlı yazılım tasarlanmıştır.[21] Uygun dönüşüm yöntemlerinin seçimi, özellikle tıbbi araştırmaların klinik sonuçlar üzerindeki etkisini değerlendirme sürecinde kritik bir öneme sahiptir.

SONUÇ

Orijinal verilerle kıyaslandığında, sağa çarpık verilere uygulanan dönüşüm teknikleri arasında Quantile dönüşümü, özellikle küçük örneklem boyutlarında (örneğin, n=50 ve n=75) yüksek AUC değerlerine sahiptir. Logaritmik, karekök (Sqrt), Box-Cox ve Yeo-Johnson dönüşümleri ise sağa çarpık veriler üzerinde benzer performanslar sergilemiş, özellikle çarpıklık katsayılarını düşürerek AUC değerlerini artırmıştır. Inverse dönüşümü ise küçük örneklem boyutlarında genellikle düşük AUC değerleriyle etkisiz bir dönüşüm olarak tespit edilmiştir.

Sola çarpık verilerde ise Quantile dönüşümü, küçük örneklem boyutlarında etkili olurken, Inverse dönüşümü daha büyük örneklem boyutlarında (örneğin, n=250 ve n=500) en iyi AUC değerlerini vermiştir. Box-Cox ve Yeo-Johnson dönüşümleri ise sola çarpık verilerde çarpıklığı başarılı bir şekilde azaltarak, özellikle orta ve büyük örneklem boyutlarında daha dengeli AUC değerleri sağlamıştır. Ayrıca, çalışmadaki amacımız, farklı dönüşüm tekniklerinin çarpık veriler üzerindeki AUC tahmin performansını karşılaştırmalı olarak değerlendirmekti. Ancak, gerçek bir AUC değeri olmadığından, elde edilen simülasyon sonuçlarını belirli hata kriterleri (bias, MSE vb.) ile karşılaştıramamak, bu çalışmanın kısıtları arasında yer almaktadır.

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 tamamen yazarın kendi eseri olup başka hiçbir yazar katkısı alınmamıştır.

KAYNAKLAR

1. Altman DG, Bland JM. Diagnostic tests 3: receiver operating characteristic plots. BMJ. 1994;309(6948):188. [Crossref] [PubMed] [PMC]
2. Zhang Z, Castelló A. Principal components analysis in clinical studies. Ann Transl Med. 2017;5(17):351. [Crossref] [PubMed] [PMC]
3. Pearson K. LIII. On lines and planes of closest fit to systems of points in space. Philos Mag J Sci. 1901;2(11):559-72. [Crossref]
4. Osborne J. Improving your data transformations: applying the box-cox transformation. Pract Assess Res Eval. 2010;15(1):12. [Crossref]
5. Tukey JW. Exploratory Data Analysis. Vol. 2. 1st ed. Springer; 1977.
6. Box GE, Cox DR. An analysis of transformations. J R Stat Soc Series B Stat Methodol. 1964;26(2):211-43. [Crossref]
7. Hosmer D, Lemeshow S. Applied Logistic Regression. 2nd ed. New York, NY, US: Wiley; 2000. [Crossref] [PubMed]
8. Hastie T, Tibshirani R, Friedman J. The Elements of Statistical Learning: Data Mining, Inference, and Prediction. 2nd ed. Springer; 2017.
9. Fawcett T. An introduction to ROC analysis. Pattern Recognit Lett. 2006;27(8):861-74. [Crossref]
10. Zou KH, O'Malley AJ, Mauri L. Receiver-operating characteristic analysis for evaluating diagnostic tests and predictive models. Circulation. 2007;115(5):654-7. [Crossref] [PubMed]
11. Durrleman S, Simon R. Flexible regression models with cubic splines. Stat Med. 1989;8(5):551-61. [Crossref] [PubMed]
12. McCullagh P. Generalized Linear Models. 2nd ed. London: Routledge; 2019. [Crossref]
13. Pepe MS. The Statistical Evaluation of Medical Tests for Classification and Prediction. 1st ed. Oxford: Oxford University Press; 2003. [Crossref]
14. Yeo IK, Johnson RA. A new family of power transformations to improve normality or symmetry. Biometrika. 2000;87(4):954-9. [Crossref]
15. Hastie T, Tibshirani R, Friedman J. The Elements of Statistical Learning. 1st ed. Springer; 2001. [Crossref]
16. Conover WJ, Iman RL. Rank transformations as a bridge between parametric and nonparametric statistics. Am Stat. 1981;35(3):124-9. [Crossref]
17. Huber PJ, Ronchetti EM. Robust Statistics. 1st ed. Hoboken, N.J.: John Wiley & Sons; 1981. [Crossref]
18. Bamber D. The area above the ordinal dominance graph and the area below the receiver operating characteristic graph. J Math Psychol. 1975;12(4):387-415. [Crossref]
19. Ünal İ. Çarpık dağılımlı verilerde ROC eğrisi altında kalan alan tahmininde transformasyon etkili mi [Is the transformation useful to estimate the area under the ROC curve with skewed data]? Cukurova Med J. 2018;43(1):141-7. [Crossref]
20. McCullagh P, Nelder JA. Generalized Linear Models. 1st ed. Boca Raton, Fla: Chapman & Hall/CRC Monographs on Statistics and Applied Probability; 1989. [Crossref]
21. Arslan AK, Tunç Z, Çolak C. Veri dönüşümü için açık kaynak erişimli web tabanlı yazılım: veri dönüşüm yazılımı [Open source access web based software for data transformation: data transformation software]. Firat Univ Saglik Bilim Tip Derg. 2019;33(3):175-81. [Link]