T.C.

AYDIN ADNAN MENDERES ÜNİVERSİTESİ

SAĞLIK BİLİMLERİ ENSTİTÜSÜ

BİYOİSTATİSTİK

DOKTORA PROGRAMI

YÜKSEK BOYUTLU SAĞKALIM VERİLERİNİN DENETİMLİ TEMEL BİLEŞENLER, CEZALI COX REGRESYON VE AŞIRI ÖĞRENME MAKİNELERİ YÖNTEMLERİ İLE KARŞILAŞTIRMALI ANALİZİ

FULDEN CANTAŞ TÜRKİŞ

DOKTORA TEZİ

DANIŞMAN

PROF. DR. İMRAN KURT ÖMÜRLÜ

AYDIN–2022

KABUL VE ONAY

T.C. Aydın Adnan Menderes Üniversitesi Sağlık Bilimleri Enstitüsü Biyoistatistik Doktora Programı çerçevesinde Fulden CANTAŞ TÜRKİŞ tarafından hazırlanan "Yüksek Boyutlu Sağkalım Verilerinin Denetimli Temel Bileşenler, Cezalı Cox Regresyon ve Aşırı Öğrenme Makineleri Yöntemleri ile Karşılaştırmalı Analizi" başlıklı tez, aşağıdaki jüri tarafından Doktora Tezi olarak kabul edilmiştir.

Tez Savunma Tarihi: 08 / 06 / 2022

Üye (T.D.)	Prof. Dr. İmran KURT ÖMÜRLÜ	Aydın Adnan Menderes Üniversitesi	……
Üye	Prof. Dr. Mevlüt TÜRE	Aydın Adnan Menderes Üniversitesi	……
Üye	Prof. Dr. Kevser Setenay DİNÇER ÖNER	Eskişehir Osmangazi Üniversitesi	……
Üye	Prof. Dr. Gökay BOZKURT	Aydın Adnan Menderes Üniversitesi	……
Üye	Prof. Dr. Canan BAYDEMİR	Kocaeli Üniversitesi	……

T.D.: Tez danışmanı

ONAY:

Bu tez Aydın Adnan Menderes Üniversitesi Lisansüstü Eğitim-Öğretim ve Sınav Yönetmeliğinin ilgili maddeleri uyarınca yukarıdaki jüri tarafından uygun görülmüş ve Sağlık Bilimleri Enstitüsünün ……..……..… tarih ve …… sayılı oturumunda alınan …… nolu Yönetim Kurulu kararıyla kabul edilmiştir.

Prof. Dr. Süleyman AYPAK

Enstitü Müdürü V.

TEŞEKKÜR

Öğrenim sürecim boyunca akademik tecrübesi, bilgi birikimi ve yol göstericiliğiyle desteklerini hiçbir zaman esirgemeyen, sabır ve azim gerektiren bu süreçte her açıdan yanımda olduğunu hissettiren değerli hocam Sayın Prof. Dr. İmran KURT ÖMÜRLÜ'ye;

Bilgi, birikim ve tecrübesiyle öğrenim sürecim boyunca çalışmalarımı destekleyen değerli hocam Sayın Prof. Dr. Mevlüt TÜRE' ye;

Akademik anlamdaki gelişim sürecini paylaştığım ve ihtiyaç duyduğum zamanlarda yardımlarını esirgemeyen değerli çalışma arkadaşım Arş. Gör. Dr. Hakan ÖZTÜRK'e;

Öğrenim sürecimiz boyunca birbirimize destek olduğumuz, tezimin yazım kontrolünde de değerli vaktini ayırarak desteğini esirgemeyen, azmini her zaman alkışladığım arkadaşım Buğra VAROL'a sonsuz ve kalpten teşekkürlerimi borç bilirim.

Bu tez çalışmasını, her şeyde olduğu gibi eğitimim için de hem maddi hem manevi anlamda büyük fedakarlık gösteren, aldığım kararlarda hep yanımda olan canım annem Güler CANTAŞ ve canım babam Haydar CANTAŞ'a; çalışkanlığını örnek alarak aynı yolda ardından yürüdüğüm değerli ablam Ayten CANTAŞ BAĞDAŞ'a; her anımda yanımda olarak beni hep destekleyen değerli eşim Can TÜRKİŞ'e ithaf ediyorum. İyi ki varsınız…

İÇİNDEKİLER

SİMGELER VE KISALTMALAR DİZİNİ

p: Bağımsız değişken sayısı

n: Birim sayısı

x: Bağımsız değişkenler vektörü

E: Eğitim seti

$\mathbf{T_i}$: Gerçek sağkalım süresi

$\mathbf{C_i}$: Sansür süresi

δ: Sansür göstergesi

$\boldsymbol{\tau_i}$: Gözlenen sağkalım süresi

β: Regresyon kaysayıları vektörü

$\mathbf{x_i}$: i. birimin bağımsız değişkenler vektörü

$\mathbf{h_0(t)}$: Temel hazard fonksiyonu

$\mathbf{h_i(t)}$**:** i. birime ilişkin orantısal hazard fonksiyonu

$\mathbf{R(t_i)}$**:** t_i zamanında risk altında olan birimlerin kümesi

$\mathbf{pl(\beta)}$**:** Kısmi logaritmik olabilirlik fonksiyonu

$\mathbf{pl_{cezalı}(\beta)}$**:** Cezalı kısmi logaritmik olabilirlik fonksiyonu

$\mathbf{p_\lambda(|\beta_j|)}$**:** Ceza terimi

λ: Ceza belirleme parametresi

$\mathbf{L_0}$: L_0-ceza parametresi

$\mathbf{L_1}$: Lasso-ceza parametresi

$\mathbf{L_2}$: Ridge-ceza parametresi

ξ: Negatif olmayan ceza parametresi

argmin: Fonksiyonu minimum yapan değerler

argmaks: Fonksiyonu maksimum yapan değerler

$\mathbf{lim}$: Limit

$\boldsymbol{\beta^T}$: β katsayısının transpozu

$\mathbf{Q}$: Köşegen matris

$\mathbf{U}$: nxm boyutlu temel bileşen matrisi

$\mathbf{D}$: Tekil değerleri içeren matris

$\mathbf{V}$: Sütunları denetimli temel bileşenleri içeren matris

$\mathbf{V^T}$: V matrisinin transpozu

$\mathbf{u_i}$: i. denetimli temel bileşen

$\mathbf{y}$: Bağımlı değişken

$\mathbf{s}$: Standartlaştırılmış regresyon katsayılarının vektörü

$\boldsymbol{\theta}$: Eşik değer

$\bar{\mathbf{y}}$: y'nin ortalaması

$\mathbf{l_j(\boldsymbol{\beta})}$: Olabilirlik fonksiyonu

$\mathbf{K}$: Çekirdek matrisi

$\mathbf{pl^{O.T}_{cezalı}}$: Cezalı-olabilirlik tabanlı kısmi logaritmik olabilirlik fonksiyonu

$\mathbf{u}$: Gradyan vektörü

$\mathbf{L}$: Negatif kısmi logaritmik olabilirlik fonksiyonu

$\mathbf{\hat{h}}$: En küçük kareler tahmincisi

$\mathbf{H^{\dagger}}$: Moore-Penrose genelleştirilmiş tersi

$\mathbf{pl^{ELM}_{cezalı}(\boldsymbol{\beta}, h_0)}$: ELMCox modelinin cezalı logaritmik olabilirlik fonksiyonu

$\boldsymbol{\eta}$**:** Öteleme terimi

$\mathbf{S(t|x)}$**:** Sağkalım fonksiyonu

$\mathbf{\hat{G}(t)}$**:** Sansür dağılımının Kaplan-Meier tahmini

$\mathbf{p_0}$: Gerçek sınıf değerleri ile modelin sınıf tahmin değerleri arasında gözlenen uyum oranı

$\mathbf{p_c}$: Gerçek sınıf değerleri ile modelin sınıf tahmin değerleri arasında tesadüfen uyum olması beklenen birimlerin oranı

AUC : Eğri altında kalan alan

AUPR : Pozitif tanımlama oranı-duyarlılık eğrisi altında kalan alan

BS: Brier skoru

CoxBoost : Boosting tabanlı Cox regresyon analizi

Cox-L_1 : L_1-cezalı Cox regresyon analizi

Cox-L_2 : L_2-cezalı Cox regresyon analizi

DTBA : Denetimli temel bileşenler analizi

ELM : Aşırı öğrenme makineleri

ELMBJEN : Topluluk öğrenme ve Buckley-James tahmincili aşırı öğrenme makineleri

ELMCox : Cezalı Cox modelli aşırı öğrenme makineleri

ELMCoxBAR : Parçalı uyarlanabilir ridge cezalı Cox modelli aşırı öğrenme makineleri

ELMCoxBoost : Olabilirlik tabanlı boosting aşırı öğrenme makineleri

ELMCoxEN : Topluluk öğrenme ve cezalı Cox modelli aşırı öğrenme makineleri

ELMmBoost : Model tabanlı boosting aşırı öğrenme makineleri

GN: Gerçek negatif

GP: Gerçek pozitif

IBS : İntegrali alınmış Brier skoru

Kappa : Cohen'in kappa katsayısı

MKK : Matthews korelasyon katsayısı

RBF : Radyal tabanlı fonksiyon

ROC : Alıcı işlem karakteristiği eğrisi

RSFC : C-indekse göre bölünmüş rastgele sağkalım ormanları

RSFL : Log-rank istatistiğine göre bölünmüş rastgele sağkalım ormanları

RSFM : Maksimum sıra istatistiğine göre bölünmüş rastgele sağkalım ormanları

SELM : Aşırı öğrenme makineleri tabanlı sağkalım analizi

SLFN : Tek katmanlı, ileri beslemeli yapay sinir ağı

TBA : Temel bileşenler analizi

YN: Yanlış negatif

YP: Yanlış pozitif

ŞEKİLLER DİZİNİ

TABLOLAR DİZİNİ

ÖZET

YÜKSEK BOYUTLU SAĞKALIM VERİLERİNİN DENETİMLİ TEMEL BİLEŞENLER, CEZALI COX REGRESYON VE AŞIRI ÖĞRENME MAKİNELERİ YÖNTEMLERİ İLE KARŞILAŞTIRMALI ANALİZİ

Cantaş Türkiş F. Aydın Adnan Menderes Üniversitesi, Sağlık Bilimleri Enstitüsü, Biyoistatistik Programı, Doktora Tezi, Aydın, 2022.

Amaç: Bu çalışmanın amacı farklı sansür oranlarına göre türetilen yüksek boyutlu sağkalım verilerinde aşırı öğrenme makineleri tabanlı sağkalım modelleri, denetimli temel bileşenler analizi ile L_2-cezalı Cox regresyon modellerinin sağkalım süresi ve kısa dönem sağkalım durumu tahminindeki performanslarının karşılaştırılması, birbirlerine benzerlik ve birbirlerinden farklılıklarının belirlenmesidir.

Gereç ve Yöntem: n=200 birim ve aralarındaki korelasyon düzeyi -0,7 ile 0,7 arasında değişen p=1000 gen ifade değeri içeren yüksek boyutlu sağkalım veri setleri rastgele türetilmiştir. Türetilen veri setleri 70:30 oranında eğitim ve test setlerine rastgele ayrılmıştır. Eğitim setleri kullanılarak aşırı öğrenme makineleri tabanlı sağkalım, denetimli temel bileşenler ve L_2-cezalı Cox regresyon modelleri eğitilmiştir. 1000 döngü ile gerçekleştirilen simülasyon sonunda modellerin test setinde sağkalım süresi ve kısa dönem sağkalım tahminlerine ilişkin performanslarının belirlenmesi için C-indeks, integrali alınmış Brier skoru, duyarlılık, özgüllük, doğruluk, negatif tanımlama, pozitif tanımlama oranları ile pozitif tanımlama oranı-duyarlılık eğrisi altında kalan alan, alıcı işlem karakteristiği eğrisi altında kalan alan, F_1 skoru, Cohen'in kappa katsayısı ve Matthews korelasyon katsayısı performans ölçütleri hesaplanmıştır.

Bulgular: Simülasyon bulguları incelendiğinde, çalışmada kullanılan sağkalım modellerinin performanslarının birbirine yakın olduğu belirlenmiştir. Sağkalım modellerinin hem sağkalım süresi hem de kısa dönem sağkalım tahminine ilişkin performanslarının sansür oranındaki artış ile düşüş eğiliminde olduğu gözlenmiştir. Uygulanan aşamalı kümeleme analizine göre, değişen sansür oranlarına göre birbirine yakın performans gösteren yöntemlerin aynı kümede yer aldığı tespit edilmiştir. Tüm senaryolarda olabilirlik tabanlı boosting aşırı öğrenme makineleri ve L_2-cezalı Cox regresyon analizi yöntemlerinin birbirine en yakın performans gösteren yöntemler

olduğu, model tabanlı boosting aşırı öğrenme makineleri yönteminin ise diğer tüm yöntemlerden daha uzak ve düşük bir performans gösterdiği dikkat çekmiştir.

Sonuç: Sonuç olarak, sağkalım verilerindeki sansür oranının yüksek olması sağkalım modellerinin performanslarını olumsuz etkilemektedir. Modellerin yüksek boyutlu sağkalım verilerinin analizindeki performansları birbirine yakın olduğundan, denetimli temel bileşenler analizi gibi boyut indirgeme yöntemlerinin ve cezalı modellerin yerine yüksek boyutlu sağkalım verilerini özellikle doğrudan analiz edebilen aşırı öğrenme makineleri tabanlı sağkalım modellerinin kullanışlı ve diğer yöntemlere tercih edilebilir olduğu ortaya konmuştur.

Anahtar kelimeler: Aşırı öğrenme makineleri, Cezalı Cox regresyon analizi, Denetimli temel bileşenler, Sağkalım, Simülasyon

ABSTRACT

COMPARATIVE ANALYSIS OF HIGH DIMENSIONAL SURVIVAL DATA WITH SUPERVISED PRINCIPAL COMPONENTS, PENALIZED COX REGRESSION, AND EXTREME LEARNING MACHINES METHODS

Cantaş Türkiş F. Aydın Adnan Menderes University, Health Sciences Institute, Biostatistics Program, Doctorate Thesis, Aydın, 2022.

Objective: The goal of the study is to compare the performances of extreme learning machines-based survival, supervised principal components analysis, and L_2-penalized Cox regression methods and determine similarity and differences among the models in the prediction of survival time and short-term survival in high dimensional survival datasets generated by varying censoring rates.

Material and Methods: Gene expression survival datasets containing n=200 units and p=1000 gene expression levels whose correlation levels were changing between -0.7 and 0.7 were randomly generated. Simulated datasets were then randomly divided into training and test sets in a 70:30 ratio. Extreme learning machines-based survival, supervised principal components, and L_2-penalized Cox regression models were trained in a training set. At the end of the 1000 times repetitive simulation, Harrell's concordance index value, integrated Brier score, sensitivity, specificity, accuracy rates, and negative predictive value, positive predictive value, the area under precision-recall, area under the curve, F_1 score, Cohen's kappa coefficient, and Matthew's correlation coefficient were calculated to reveal the performances of the methods.

Results: When the simulation results were examined, it was determined that the survival models' performances were close to each other. It was also observed that the performances of survival models concerning the prediction of both survival time and short-term survival tend to decrease by increasing the censoring rate. According to the applied hierarchical clustering analysis, it was determined that the methods that perform close to each other according to the varying censoring rates were in the same cluster. It was noted that in all scenarios, an extreme learning machine Cox model with likelihood-based boosting and L_2-penalized Cox methods were the methods that showed the closest performance to each other. In contrast, an extreme

learning machine Cox model with a gradient-based boosting method showed far lower performance than other methods.

Conclusion: To conclude, the high rate of censoring in survival data adversely affects the performance of the survival models. Since the performances of the models in the analysis of high-dimensional survival data were close to one another, it was revealed that extreme learning machines-based survival models, which can directly analyze high-dimensional survival data, were useful and can be preferred instead of dimension reduction methods such as supervised principal component analysis, and penalized models.

Keywords: Extreme learning machines, Penalized Cox regression model, Simulation, Supervised principal components, Survival

1. GİRİŞ

Gelişen teknoloji ile birlikte birçok alanda olduğu gibi sağlık alanında da veri toplamak ve toplanan verileri saklamak daha da kolay hale gelmiştir. Bu durum, veri boyutundaki artışı da beraberinde getirmiştir. Veri boyutundaki artış ile birlikte değişken sayısının gözlem sayısından fazla olması durumu yüksek boyutlu veri kavramını ortaya çıkarmıştır.

Verinin yüksek boyutlu olması, incelenen birimler hakkında çok fazla bilgi elde edilmesini mümkün kılmaktadır. Örneğin, bir hasta için on binler ya da yüz binlerce gen ifade değeri içeren veri seti; hastalıkların biyolojik süreçlerini, tümör türlerini ya da genler arasındaki etkileşimin araştırılmasını sağlamaktadır. Günümüzde, yüksek boyutlu gen ifade veri setlerinde regresyon, sınıflandırma ya da sağkalım problemleri çözümlenebilmektedir. Buna karşılık, yüksek boyutlu verilerden istatistiksel analizler yardımıyla bilgi elde etme süreci, çoklu doğrusal bağlantı, uzun analiz süresi ve sonuçların yorumlanmasının güçlüğü gibi birtakım zorlukları da barındırmaktadır. Ayrıca klasik sağkalım analizi yöntemleri, teorik yapıları nedeniyle yüksek boyutlu verilerin analizinde doğrudan kullanılamamaktadır. Bu nedenle yeni analiz yöntemlerinin ortaya konması ya da varolan analiz yöntemlerinin geliştirilmesi gibi yaklaşımlar araştırmacıların odak noktası haline gelmiştir. Buna bağlı olarak, klasik sağkalım analizi modellerinde bazı değişiklikler yapılarak L_2-cezalı Cox regresyon (Cox-L_2) gibi modeller geliştirilmiş, denetimli temel bileşenler analizi (DTBA) gibi boyut indirgeme yöntemleri öne sürülmüş ve klasik yöntemler ile yeni yaklaşımlar birleştirilerek aşırı öğrenme makineleri tabanlı sağkalım (SELM) modelleri gibi yöntemler ortaya konmuştur (Bair ve Tibshirani, 2004; Verweij ve Van Houwelingen, 1994; Wang ve Li, 2019; Wang ve diğerleri., 2018; Wang ve Zhou, 2018). Yüksek boyutlu sağkalım verilerinin bu yöntemlerle analiz edilmesi; sağkalımı etkileyen risk faktörlerinin belirlenmesi, kanser gibi önemli hastalıkların erken teşhis edilmesi ya da 5-yıllık, 10-yıllık sağkalım tahmini gibi kısa ve uzun dönem risklerin belirlenmesi açısından büyük önem taşımaktadır (Delen ve diğerleri., 2005; Dhillon ve Singh, 2020; Gitto ve diğerleri., 2022; Li ve diğerleri., 2021; Lou ve diğerleri., 2022; Yu ve diğerleri., 2022). Literatürde, SELM modellerinin yüksek boyutlu sağkalım verilerinin analizinde kullanıldığı ve başka yöntemlerle karşılaştırıldığı çalışmalar incelendiğinde; (Wang ve Zhou, 2018)'nun topluluk öğrenme ve Buckley-James tahmincili ELM (ELMBJEN, ensemble of ELM with Buckley-James estimator), topluluk öğrenme ve cezalı Cox modelli ELM (ELMCoxEN, the ensemble of ELM with penalized Cox model), olabilirlik tabanlı

boosting ELM (ELMCoxBoost, ELM with likelihood-based boosting) ve model tabanlı boosting ELM (ELMmBoost, ELM with model-based boosting) yöntemlerini log-rank istatistiğine göre bölünmüş rastgele sağkalım ormanları (RSFL, random survival forests with log-rank split), maksimum sıra istatistiğine göre bölünmüş rastgele sağkalım ormanları (RSFM, RSF with maximally selected rank statistic splitting), C-indekse göre bölünmüş rastgele sağkalım ormanları (RSFC, RSF with C-index splitting), L_1-cezalı Cox regresyon (Cox-L_1) ve Cox-L_2 yöntemlerini yüksek boyutlu gerçek sağkalım veri setlerinde karşılaştırdıkları görülmüştür. (Wang ve Li, 2019)'nin çalışmasında ise, parçalı uyarlanabilir ridge ELMCox (ELMCoxBAR, Cox regression via broken adaptive ridge) modeli ileri sürülmüş ve ELMCoxBAR modelinin performansı Cox-L_1, Cox-L_2, RSF, RSFL, RSFM ve boosting tabanlı Cox regresyon (CoxBoost) modelleri ile karşılaştırılmıştır.

Tüm bağımsız değişkenleri kullanarak yüksek boyutlu sağkalım verilerini analiz eden yöntemler dışında, bu tür verileri analiz etmenin bir başka yolu da sağkalım ile ilişkili bağımsız değişkenleri belirleyerek analize dahil etmektir. Bu amaçla (Bair ve Tibshirani, 2004) tarafından geliştirilmiş DTBA yöntemi; yüksek boyutlu sağkalım verilerinde sağkalım ile ilişkili genlerin belirlenerek veri boyutunun azaltılmasını sağlayan bir yöntem olarak göze çarpmaktadır. Çalışmalarında da DTBA'nın yüksek boyutlu gerçek ve türettikleri gen ifade veri setlerinde sağkalım tahmin performansının başarılı olduğunu belirtmişlerdir. Literatürde DTBA'nın alternatif yöntemlerle simülasyon ya da gerçek gen ifade verilerinde karşılaştırıldığı başka çalışmalar da vardır (Aktürk Hayat ve diğerleri., 2016; Bair ve diğerleri., 2006; Türe ve Kurt Ömürlü, 2018). Fakat SELM yöntemleri ve DTBA'nın simülasyonla karşılaştırıldığı çalışma literatürde henüz yer almamaktadır. Ulusal ve uluslararası literatürde; değişen sansür oranlarına göre türetilen yüksek boyutlu sağkalım verilerinin analizinde SELM modelleri, DTBA ve Cox-L_2 sağkalım analizi yöntemlerinin simülasyon ile karşılaştırıldığı, birbirine benzer performans gösteren yöntemlerin belirlendiği başka bir çalışma olmaması çalışmamızı özgün kılmaktadır.

1.1. Tezin Amacı

Bu çalışmada, değişen sansür oranlarına göre türetilen yüksek boyutlu sağkalım verilerinin sağkalım süresi ve kısa dönem sağkalım durumu tahminine ilişkin analizinde DTBA, Cox-L_2, ELMCox, ELMCoxBAR, ELMCoxBoost, ELMCoxEN ve ELMmBoost yöntemlerinin C-indeks, integrali alınmış Brier skoru (IBS), duyarlılık, özgüllük, doğruluk

oranları ile negatif tanımlama oranı (NTO), pozitif tanımlama oranı (PTO), PTO-duyarlılık eğrisi altında kalan alan (AUPR), alıcı işlem karakteristiği eğrisi (ROC) altında kalan alan (AUC), F_1 skoru, Cohen'in kappa katsayısı (Kappa) ve Matthews korelasyon katsayısı (MKK) kriterlerine göre değerlendirilmesi amaçlandı. Bu amaçlar doğrultusunda çalışmanın araştırma hipotezleri aşağıda verildi:

- Yüksek boyutlu sağkalım verilerindeki sansür oranı SELM, DTBA ve Cox-L_2 yöntemlerinin tahmin performanslarını etkilemektedir.
- Değişen sansür oranına göre SELM yöntemlerinin tahmin performansları DTBA ve Cox-L_2 yöntemlerinden daha yüksektir.

2. GENEL BİLGİLER

2.1. Cezalı Kısmi Logaritmik Olabilirlik Fonksiyonu

p sayıda bağımsız değişken seti, $\mathbf{x} = (x_1, x_2, x_3, \dots, x_p)$ bağımsız değişkenler vektörü, $\mathbf{x_i} = (x_{1i}, x_{2i}, x_{3i}, \dots, x_{pi})$ i. birimin bağımsız değişkenler vektörü olmak üzere $n * (p + 2)$ boyutlu E eğitim seti Eşitlik (1)'deki gibi tanımlansın:

$$E = (\tau_i, \delta_i, \mathbf{x}_i), \ i = 1, 2, \dots, n \tag{1}$$

Sağdan sansürlü bir veri setinde T_i gerçek sağkalım süresi, C_i sansür süresi ve δ_i ise sansür göstergesi olmak üzere gözlenen sağkalım süresi τ_i Eşitlik (2)'deki gibi tanımlansın:

$$\tau_i = \min(T_i, C_i) \tag{2}$$

$\boldsymbol{\beta} = (\beta_1, \beta_2, \beta_3, \dots, \beta_p)^T$ regresyon katsayısı vektörü, t ise T'nin herhangi bir değeri olmak üzere; i. birime ilişkin orantısal hazard fonksiyonu $h_i(t)$, $h_0(t)$ temel hazard fonksiyonu ile exp $(\mathbf{x}_i\boldsymbol{\beta})$ hazard oranının çarpımı şeklinde ifade edilir:

$$h_i(t) = h_0(t) * \exp(\mathbf{x}_i\boldsymbol{\beta}) \tag{3}$$

Böylece, Cox regresyon modelindeki herhangi bir (i, k) çiftine ilişkin $h_i(t)$ ve $h_k(t)$ hazard oranları orantısaldır. Burada $i \neq k$ ve k=1, 2, 3, … , n olarak tanımlıdır (Wang ve Li, 2019).

$R(t_i) = \{k | t_k \geq t_i\}$, t_i zamanında risk altında olan birimlerin kümesi olmak üzere; Cox regresyon modeline ilişkin $\boldsymbol{\beta}$ parametreleri, Eşitlik (4) ile ifade edilen kısmi logaritmik olabilirlik fonksiyonu maksimize edilerek tahmin edilir (Verweij ve Van Houwelingen, 1994).

$$pl(\boldsymbol{\beta}) = \sum_{i=1}^{n} \delta_i \mathbf{x}_i \boldsymbol{\beta} - \sum_{i=1}^{n} \delta_i \log\left[\sum_{j \in R(\tau_i)} \exp(\mathbf{x}_j \boldsymbol{\beta})\right] \tag{4}$$

Yarı-parametrik bir yöntem olması ve regresyon katsayılarının kolaylıkla yorumlanabilmesi ile Cox regresyon modeli, düşük boyutlu sağkalım verilerinin analizinde sık kullanılan bir yaklaşımdır. Ancak p>n olan sağkalım verilerinde veri boyutunun yüksek olması ve değişkenler arasındaki yüksek doğrusal bağlantıdan dolayı Cox regresyon modeli doğrudan uygulanamamaktadır (Lai ve diğerleri., 2013). Bu tür verilerin modellenmesi için kısmi cezalı logaritmik olabilirlik tabanlı Cox regresyon analizi yaklaşımları ileri sürülmüştür. Cezalı kısmi logaritmik olabilirlik fonksiyonu ($\mathrm{pl}_{\text{cezalı}}(\boldsymbol{\beta})$); $\mathrm{p}_\lambda(|\beta_j|)$ ceza terimi olmak üzere aşağıdaki eşitlik ile hesaplanır:

$$\mathrm{pl}_{\text{cezalı}}(\boldsymbol{\beta}) = \mathrm{pl}(\boldsymbol{\beta}) - \sum_{j=1}^{p} \mathrm{p}_\lambda(|\beta_j|) \tag{5}$$

Cezalı Cox regresyon modeline ilişkin $\boldsymbol{\beta}$ parametreleri, $\mathrm{pl}_{\text{cezalı}}(\boldsymbol{\beta})$ fonksiyonu maksimize edilerek tahmin edilir.

2.1.1. Parçalı Uyarlanabilir Ridge-Cezalı Cox Regresyon Modeli

CoxBAR modeli, bir Cox ridge tahmincisi ile başlayıp ardından iteratif olarak tekrar ağırlıklandırılmış ridge regresyon uygulayan, L_0-cezalı bir regresyona yakınsamayı amaçlayan cezalı Cox regresyon modellerinden biridir.

λ_n ve ξ_n negatif olmayan ceza belirleme parametreleri olmak üzere, CoxBAR modelinin β katsayı tahmini ilk olarak Cox-L_2 regresyon tahmini ile başlamaktadır (Verweij ve Van Houwelingen, 1994):

$$\hat{\beta}^{(0)} = \arg\min_{\beta} \left\{ -2\mathrm{pl}(\boldsymbol{\beta}) + \xi_n \sum_{j=1}^{p_n} \beta_j^2 \right\} \tag{6}$$

Yukarıdaki eşitlikteki $\boldsymbol{\beta}$ katsayısının tahmini, Cox-L_2 regresyon tahmincisi ile iteratif olarak yeniden ağırlıklandırılarak aşağıdaki eşitlikle güncellenir:

$$\hat{\beta}^{(k)} = \arg\min_{\beta} \left\{ -2\mathrm{pl}(\boldsymbol{\beta}) + \lambda_n \sum_{j=1}^{p_n} \frac{\beta_j^2}{\left(\hat{\beta}_j^{(k-1)}\right)^2} \right\}, \quad k \geq 1 \tag{7}$$

CoxBAR modelinin $\boldsymbol{\beta}$ parametre tahmini aşağıdaki gibi bulunur:

$$\widehat{\boldsymbol{\beta}} = \lim_{k\to\infty} \hat{\beta}^{(k)} \tag{8}$$

λ_n her iterasyonda sabitlenmiş olsa da, bir önceki iterasyondan elde edilen ridge regresyon tahminlerinin kareleri ile ters ağırlıklandırılır. Sonuç olarak, gerçek değerleri sıfır olan katsayılar bir sonraki iterasyonda daha büyük ceza değerine sahip olurken, gerçek değeri sıfırdan farklı olan katsayılar sabit bir değere yakınsar.

Cezalı modellerde ceza parametresinin değeri çapraz geçerlilik, Akaike bilgi kriteri ya da Bayes bilgi kriterine göre belirlenir. Ancak CoxBAR modelinde hesaplama kısıtlılığı oluşturabilecek bu kriterlerden farklı olarak $\lambda = \mathrm{In}(n)$ değeri ile L_0-cezalı Cox regresyon modelinin amaç fonksiyonu Bayes bilgi kriteri değerine eşittir. Ayrıca $\lambda = \mathrm{In}(n)$ olduğunda CoxBAR tahmin modelinin performansı ξ_n değerinden etkilenmemektedir (Kawaguchi ve diğerleri., 2017; Wang ve Li, 2019).

2.1.2. L_2-Cezalı Cox Regresyon Analizi

Cox-L_2 modeli, p>n olan sağkalım verilerinin analizinde ileri sürülen cezalı Cox regresyon analizi yöntemlerinden biridir. Burada ceza terimi L_2, ridge ceza terimi olarak da bilinmektedir. Cox-L_2 modeli ile tahmin edilen regresyon katsayılarının değeri sıfır civarındadır. Bu da, tüm bağımsız değişkenlerin modelde bulunduğunu gösterir. Ayrıca, Cox-L_2 modeline ilişkin tahmin değerleri λ parametresinin seçimine göre değişeceğinden, Cox-L_2 yöntemi tahminleri yanlıdır ancak klasik yöntemlerden daha kararlı sonuçlar ortaya koyar. λ parametresinin en uygun değeri, hata kareler ortalamasını en küçük yapan değerdir (Lai ve diğerleri., 2013; Van Houwelingen ve diğerleri., 2006; Verweij ve Van Houwelingen, 1994).

Cox regresyon modeline L_2-ceza parametresi eklendiğinde kısmi logaritmik olabilirlik fonksiyonu aşağıdaki gibi elde edilir (Perperoglou, 2014; Van Houwelingen ve diğerleri., 2006; Verweij ve Van Houwelingen, 1994):

$$\mathrm{pl}_{\text{cezalı}}(\boldsymbol{\beta}, h_0) = \mathrm{pl}(\boldsymbol{\beta}, h_0) - \frac{1}{2}\lambda\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{\beta} \qquad \lambda \geq 0 \tag{9}$$

λ'nın sabit pozitif bir değeri için $\boldsymbol{\beta}$'ya göre $\mathrm{pl}_{\mathrm{cezalı}}(\boldsymbol{\beta}, \mathrm{h}_0)$ fonksiyonunun maksimum yapılması için aşağıdaki eşitlikler kullanılarak Newton-Raphson algoritması uygulanır (Van Houwelingen ve diğerleri., 2006):

$$\frac{\partial \mathrm{pl}_{\mathrm{cezalı}}(\boldsymbol{\beta},\mathrm{h}_0)}{\partial \boldsymbol{\beta}} = \sum_{\mathrm{i}=1}^{\mathrm{n}} [\delta_{\mathrm{i}} - \exp(\mathbf{x}_{\mathrm{i}}\boldsymbol{\beta})\mathrm{H}_0(\mathrm{t}_{\mathrm{i}})]\mathbf{x}_{\mathrm{i}} - \lambda\boldsymbol{\beta} \quad \textbf{(10)}$$

$$\frac{\partial^2 \mathrm{pl}_{\mathrm{cezalı}}(\boldsymbol{\beta},\mathrm{h}_0)}{\partial \boldsymbol{\beta}^2} = -\left[\sum_{\mathrm{i}=1}^{\mathrm{n}} \exp(\mathbf{x}_{\mathbf{i}}\boldsymbol{\beta})\, \mathrm{H}_0(\mathrm{t}_{\mathrm{i}})\mathbf{x}_{\mathrm{i}}\mathbf{x}_{\mathrm{i}}^{\mathrm{T}} + \lambda\mathbf{I}_{\mathrm{p}}\right] = -\left(\mathbf{x}^{\mathrm{T}}\mathbf{Q}\mathbf{x} + \lambda\mathbf{I}_{\mathrm{p}}\right) \quad (11)$$

Burada $\mathrm{H}_0(\mathrm{t}_{\mathrm{i}}) = \sum_{\mathrm{s}\leq\mathrm{t}} \mathrm{h}_0(\mathrm{s})$ ve $\mathbf{Q} = \text{köşegen}(\exp(\mathbf{x}_{\mathbf{i}}\boldsymbol{\beta})\, \mathrm{H}_0(\mathrm{t}_{\mathrm{i}}))$.

Newton-Raphson prosedürüne göre $\boldsymbol{\beta}$'nın güncellenmesi ikinci kısmi türevin tersinin hesaplanmasını gerektirir (Van Houwelingen ve diğerleri., 2006):

$$\left(\mathbf{x}^{\mathrm{T}}\mathbf{Q}\mathbf{x} + \lambda\mathbf{I}_{\mathrm{p}}\right)^{-1} = \frac{1}{\lambda}\left[\mathbf{I}_{\mathrm{p}} - \mathbf{x}^{\mathrm{T}}(\mathbf{x}\mathbf{x}^{\mathrm{T}} + \lambda\mathbf{Q}^{-1})^{-1}\mathbf{x}\right]. \quad (12)$$

2.2. Denetimli Temel Bileşenler Analizi

Yüksek boyutlu sağkalım verilerindeki biyolojik çeşitlilikten dolayı, veri setindeki tüm genlere TBA uygulanması durumunda elde edilen temel bileşenlerin sağkalım ile ilişkili olacağının garantisi yoktur. Bu durumu çözmek için DTBA yöntemi (Bair ve Tibshirani, 2004) tarafından ileri sürülmüştür. DTBA yönteminde, TBA yönteminden farklı olarak veri setindeki tüm genler yerine bağımlı değişken ile en güçlü ilişkiye sahip olan genlerin alt kümesi kullanılarak denetimli temel bileşenler tahmin edilir (Bair ve diğerleri., 2006; Türe ve Kurt Ömürlü, 2018). Gen seçimi aşamasında bağımlı değişken de göz önünde tutulduğu için DTBA, denetimli bir yöntemdir (Chen ve diğerleri., 2008).

n birim, p değişkenden oluşan $\mathbf{x}_{\mathrm{nxp}}$ bağımsız değişkenler matrisi ve $\mathbf{y}$ ise (nx1) boyutlu bağımlı değişken vektörü olmak üzere DTBA modelinin uygulama adımları aşağıdaki gibidir (Bair ve diğerleri., 2006; Chen ve diğerleri., 2008):

1. Her bir bağımsız değişken için standartlaştırılmış regresyon katsayıları hesaplanır.

2. Çapraz geçerlilik ile hesaplanan θ eşik değerinden mutlak değerce büyük olan katsayıya sahip olan bağımsız değişkenlerden indirgenmiş veri matrisi oluşturulur.
3. İndirgenmiş veri matrisinden birinci ya da ilk birkaç denetimli temel bileşen hesaplanır.
4. Hesaplanan temel bileşen(ler) ile bağımlı değişken tahmini yapılır.

DTBA modelinin temeli tekil değer ayrışmasına dayanır. Bağımsız değişkenlerin sıfır ortalamaya göre merkezileştirildiği varsayılsın. $\mathbf{U}, \mathbf{D}$ ve $\mathbf{V}$ sırasıyla nxm, mxm, mxp boyutlu matris ve $\mathbf{x}$'in rankı m=minimum(n-1,p) olmak üzere $\mathbf{x}$'in tekil değer ayrışması aşağıdaki gibi ifade edilir (Bair ve diğerleri., 2006; Bair ve Tibshirani, 2004; Witten ve Tibshirani, 2010):

$$\mathbf{x} = \mathbf{U}\mathbf{D}\mathbf{V}^{\mathrm{T}} \tag{13}$$

Burada $\mathbf{D}$, d_j tekil değerleri içeren köşegen matrisi ifade etmektedir. Ayrıca $\mathbf{U}$ matrisinin sütunları $u_1, u_2, \dots, u_m$ denetimli temel bileşenleri içermektedir öyle ki tekil değerlerin $d_1 \geq d_2 \geq \cdots \geq d_m \geq 0$ olarak sıralı olduğu varsayılır.

s, her bir bağımsız değişkenin $\mathbf{y}$ bağımlı değişkeni üzerine etkilerinin ayrı ayrı ölçülmesi amacıyla hesaplanan standartlaştırılmış regresyon katsayılarının p boyutlu vektörü olmak üzere aşağıdaki gibi hesaplanır:

$$s_j = \frac{\mathbf{x}_j^{\mathrm{T}}\mathbf{y}}{\sqrt{\mathbf{x}_j^{\mathrm{T}}\mathbf{x}_j}} \tag{14}$$

C_θ, $|s_j| > \theta$ olan indislerin bir kümesi ve $\mathbf{x}_\theta$, $\mathbf{x}$'in C_θ'ya ilişkin sütunlarından oluşan matris olmak üzere $\mathbf{x}_\theta$'nın tekil değer ayrışması aşağıdaki gibi hesaplanır (Bair ve diğerleri., 2006):

$$\mathbf{x}_\theta = \mathbf{U}_\theta \mathbf{D}_\theta \mathbf{V}_\theta^{\mathrm{T}}. \tag{15}$$

Burada $\mathbf{U}_\theta = \mathbf{u}_{\theta,1}, \mathbf{u}_{\theta,2}, \dots, \mathbf{u}_{\theta,m}$ olarak gösterilir ve $\mathbf{u}_{\theta,1}, \mathbf{u}_{\theta,2}, \dots, \mathbf{u}_{\theta,m}$ $\mathbf{x}$'in sırasıyla birinci, ikinci, …, m. denetimli temel bileşenidir.

$\mathbf{u}_{\theta,1}$ tahmincili tek değişkenli doğrusal regresyon modeli aşağıdaki gibi gösterilir (Bair ve diğerleri., 2006):

$$\hat{\mathbf{y}}^{\mathrm{dtb},\theta} = \bar{\mathbf{y}} + \hat{\gamma} * \mathbf{u}_{\theta,1} \tag{16}$$

$\mathbf{u}_{\theta,1}$, $\mathbf{x}_\theta$'nın sol tekil vektörü olduğundan sıfır ortalamalı birim uzunluğa sahiptir. Böylece $\hat{\gamma} = \mathbf{u}_\theta^{\mathrm{T}}\mathbf{y}$ ve kesim noktası ise $\bar{\mathbf{y}}$, $\mathbf{y}$'nin ortalamasıdır. θ'nın en iyi değerinin hesaplanması için logaritmik olabilirlik ya da kısmi logaritmik olabilirlik oranı istatistiğinin çapraz geçerliliği kullanılır (Bair ve diğerleri., 2006; Bair ve Tibshirani, 2004).

Eşitlik (15)'e göre $\mathbf{U}_\theta$ aşağıdaki gibi yazılabilir:

$$\begin{aligned}\mathbf{U}_\theta &= \mathbf{x}_\theta \mathbf{V}_\theta \mathbf{D}_\theta^{-1} \\ &= \mathbf{x}_\theta \mathbf{W}_\theta\end{aligned} \tag{17}$$

olarak yazılır. Bu sebeple, örneğin; $\mathbf{u}_{\theta,1}$, $\mathbf{x}_\theta$'nın sütunlarının doğrusal birleşimidir: $\mathbf{u}_{\theta,1} = \mathbf{x}_\theta \mathbf{w}_{\theta,1}$. Böylece doğrusal model regresyon tahmini, $\mathbf{x}_\theta$'daki tüm tahminciler kullanılarak oluşturulan sınırlandırılmış doğrusal regresyon model tahmini olarak görülebilir (Bair ve diğerleri., 2006):

$$\begin{aligned}\hat{\mathbf{y}}^{\mathrm{dtb},\theta} &= \bar{\mathbf{y}} + \hat{\gamma} * \mathbf{x}_\theta \mathbf{w}_{\theta,1} \\ &= \bar{\mathbf{y}} + \mathbf{x}_\theta \widehat{\boldsymbol{\beta}}_\theta\end{aligned} \tag{18}$$

Burada $\widehat{\boldsymbol{\beta}}_\theta = \hat{\gamma}\mathbf{w}_{\theta,1}$ ile verilmiştir. $\mathbf{w}_{\theta,1}$, C_θ dışındaki ilişkili bağımsız değişkenleri göstermek üzere, elde edilen tahmin tüm p sayıda bağımsız değişken için doğrusaldır.

DTBA, yüksek boyutlu sağkalım verilerine uygulanabilir. Bu durumda, Eşitlik (14) ile belirtilen standartlaştırılmış regresyon katsayıları yerine skor istatistiği; Eşitlik (16) ile belirtilen tek değişkenli doğrusal regresyon modeli yerine Cox regresyon modeli kullanılır.

$\mathrm{l_j}(\beta)$, $\mathbf{x_j}$ bağımsız değişkeni ve $\mathbf{y}$ bağımlı değişkenine ilişkin verinin logaritmik olabilirlik ya da kısmi olabilirlik fonksiyonu olsun. Ayrıca $\mathrm{U_j}(\beta_0) = \mathrm{dl}/\mathrm{d}\beta|_{\beta=\beta_0}$ ve $\mathrm{l_j}(\beta_0) = -\mathrm{d}^2\mathrm{l_j}/\mathrm{d}\beta^2|_{\beta=\beta_0}$ olsun. j. tahminci için skor istatistiği aşağıdaki gibi hesaplanır (Bair ve diğerleri., 2006):

$$\mathrm{s_j} = \frac{\mathrm{U_j}(0)^2}{\mathrm{I_j}(0)} \tag{19}$$

Hesaplanan bu istatistik, Eşitlik (14) ile gösterilen standartlaştırılmış regresyon katsayılarına eşittir.

2.3. Aşırı Öğrenme Makineleri (ELM)

ELM, genelleştirilmiş tek katmanlı ileri beslemeli yapay sinir ağı (SLFN) yapısına sahip bir makine öğrenme yöntemidir (Huang, 2014; HuangChen ve diğerleri., 2006; Huang ve diğerleri., 2004; HuangZhu ve diğerleri., 2006; Kasun ve diğerleri., 2016).

ELM teorisine göre w_i gizli katman ağırlıkları ve b_i bias değerleri rastgele oluşturulur ve bu değerler eğitim setinden bağımsızdır ve bu parametrelerin ayarlanmasına gerek yoktur (Kasun ve diğerleri., 2016). Sadece eğitim hatasını minimum yapan geri yayılımlı öğrenme algoritmasından farklı olarak, ELM algoritması hem eğitim hatasını hem de çıktı ağırlıklarını minimum yapmaktadır (Kasun ve diğerleri., 2016; Rumelhart ve diğerleri., 1986). Çıktı ağırlıklarının minimum olması ise genelleştirme performansının daha iyi olmasını sağlamaktadır (Huang ve diğerleri., 2011). Bu teorinin uygulanabilirliği evrensel yakınsama teoremine dayanmaktadır. Evrensel yakınsama teoremine göre; sabit olmayan, parçalı sürekli bir g fonksiyonu için f(x) sürekli çıktı fonksiyonu, uyarlanabilir g gizli düğümlerini içeren SLFN yapıları ile tahmin edilecekse, böyle bir SLFN yapısına ilişkin gizli düğüm parametrelerinin ayarlanmasına gerek yoktur. Kısaca, f(x) sürekli çıktı fonksiyonu ve L gizli düğüm sayısı olmak üzere rastgele türetilmiş bir $\{w_i, b_i\}_{i=1}^{L}$ dizisi için

$$\lim_{L\to\infty} \|f(x) - f_L(x)\| = \lim_{L\to\infty} \left\| f(x) - \sum_{i=1}^{L} g(x, w_i, b_i)\beta_i \right\| = 0 \qquad (20)$$

eşitliği, $\|f(x) - f_L(x)\|$ normunu en küçük yapan $\boldsymbol{\beta}$ değerleri için tanımlıdır (HuangChen ve diğerleri., 2006).

n gözlem sayısı, p bağımsız değişken sayısı, $y_i \in R^m$ her gözleme ilişkin bağımlı değişken değeri (Regresyon problemlerinde m=1, sınıflandırma problemlerinde m≥2) olmak üzere eğitim seti $\{(x_i, y_i)|x_i \in R^p, y_i \in R^m\}_{i=1}^{n}$ olsun. Buna göre g(.) aktivasyon fonksiyonu; $w_i \in R^p$, giriş ve çıktı katmanı arasında yer alan, herhangi bir sürekli olasılık dağılımından rasgele türetilmiş giriş ağırlıkları vektörü; b_i, i. gizli katmana ilişkin bias ve $\boldsymbol{\beta} = [\beta_1, \dots, \beta_i, \dots, \beta_L]^T$ çıktı ağırlıkları vektörü ve $\mathbf{h}(x) = [h_1(x),\ h_2(x), \dots, h_L(x)]^T$ ise gizli

katmanları belirtmek üzere bir SLFN aşağıdaki eşitlikle gösterilir (Şekil 1) (HuangZhu ve diğerleri., 2006; Wang ve Li, 2019):

$$f_L(x) = \sum_{i=1}^{L} g(x, w_i, b_i)\, \beta_i = \mathbf{h}(x)\boldsymbol{\beta} \tag{21}$$

$\mathbf{H} = [\mathbf{h}(x_1), \mathbf{h}(x_2), \ldots, \mathbf{h}(x_n)]^T$ olmak üzere $\boldsymbol{\beta}$ çıktı ağırlıklarının çözümü eğitim hatasını minimum yapan değere eşittir (Lu ve diğerleri., 2019):

$$\hat{\beta} = \text{argmin}\|\mathbf{y} - \mathbf{H}\boldsymbol{\beta}\| \tag{22}$$

Çıktı ağırlıklarının optimal tahmini Moore-Penrose genelleştirilmiş tersi ($\mathbf{H}^{\dagger}$) ile aşağıdaki gibi temsil edilebilir:

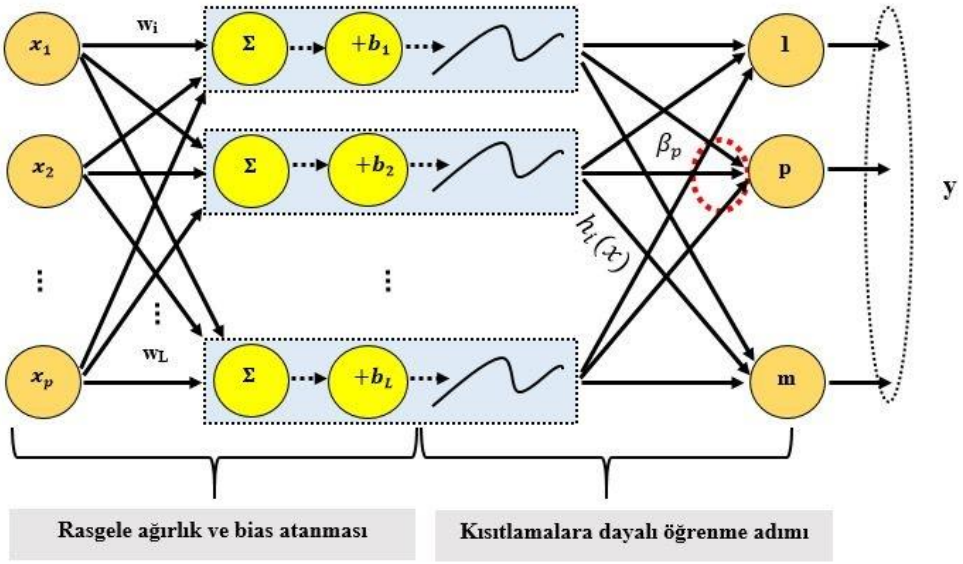
$$\hat{\beta} = \mathbf{H}^{\dagger} y \tag{23}$$

$\mathbf{H}^{\dagger}$'nın çözümü için dik izdüşüm kullanılabilir:

$$\mathbf{H}^{\dagger} = (\mathbf{H}^T\mathbf{H})^{-1}\mathbf{H}^T \tag{24}$$

Ridge regresyon teorisine göre daha kararlı ve daha iyi genelleştirilebilir sonuçlar elde edilebilmesi için $\mathbf{H}^T\mathbf{H}$ matrisinin köşegenlerine pozitif bir değer eklenebilir (Huang ve diğerleri., 2011; Wang ve Li, 2019). Böylece $\mathbf{I}_{nxn}$ birim matris olmak üzere çıktı ağırlıklarının kapalı formdaki çözümü aşağıdaki eşitlikle gösterilir (Kasun ve diğerleri., 2016):

$$\widehat{\boldsymbol{\beta}} = \mathbf{H}^y \left(\frac{I}{C} + \mathbf{H}\mathbf{H}^y\right)^{-1} \mathbf{y} \tag{25}$$



Şekil 1. ELM yapısı

H fonksiyonu bilinmiyor ya da kapalı formda ise ELM yapısı pozitif tanımlı herhangi bir çekirdek dönüşümü (K) ile de oluşturulabilir (Lu ve diğerleri., 2019).

$$\mathbf{K}^{\mathrm{T}}(\mathbf{x_i})\boldsymbol{\beta} = \mathrm{y_i} - \varepsilon_\mathrm{i}, \quad \mathrm{i} = 1,2,\dots,\mathrm{n} \tag{26}$$

kısıtı ile çekirdek ELM ya da çekirdek fonksiyonlu SLFN yapısı optimizasyon formunda aşağıdaki gibi tanımlanır:

$$\min_{\beta} \frac{1}{2}\|\boldsymbol{\beta}\|_2^2 + \frac{\mathrm{C}}{2}\sum_{\mathrm{i}=1}^{\mathrm{n}} \varepsilon_\mathrm{i}^2 \tag{27}$$

Burada ε_i, i. eğitim hatası, C ise $\boldsymbol{\beta}$ ve ε arasındaki denge parametresidir.

Uygun çekirdek dönüşümü ile, çözülememiş doğrusal bir problem çekirdek uzayında çözülebilir bir doğrusal probleme dönüştürülebilir. Karush-Kuhn-Tucker teoremine (Fletcher, 2013) göre ve ardından uygulanan α_i Lagrange çarpanı ile $\boldsymbol{\beta}$ çıktı ağırlıklarının çözümü sağlanır (Huang, 2014):

$$\min_{\beta,\alpha,\varepsilon} \left\{ \mathrm{L} = \frac{1}{2}\|\boldsymbol{\beta}\|_2^2 + \frac{\mathrm{C}}{2}\sum_{\mathrm{i}=1}^{\mathrm{n}} \varepsilon_\mathrm{i}^2 - \sum_{\mathrm{i}=1}^{\mathrm{n}} \alpha_\mathrm{i}(\mathbf{K}^{\mathrm{T}}(\mathbf{x_i})\boldsymbol{\beta} - \mathrm{y_i} + \varepsilon_\mathrm{i}) \right\} \tag{28}$$

Kısmi türevler sıfıra eşitlendikten sonra Karush-Kuhn-Tucker koşulları aşağıdaki gibi yazılabilir:

$$\frac{\partial L}{\partial \beta_j} = 0, \;\; j = 1, \dots, L \;\Rightarrow \boldsymbol{\beta} = \mathbf{K}^T \boldsymbol{\alpha} \tag{29}$$

$$\frac{\partial L}{\partial \varepsilon_i} = 0, \;\; i = 1, \dots, n \;\Rightarrow \boldsymbol{\alpha} = C\boldsymbol{\varepsilon} \tag{30}$$

$$\frac{\partial L}{\partial \alpha_i} = 0, \;\; i = 1, \dots, n \;\Rightarrow \mathbf{K}\boldsymbol{\beta} - \mathbf{y} + \boldsymbol{\varepsilon} = \mathbf{0} \tag{31}$$

Burada $\boldsymbol{\phi} = [\varphi(x_1), \varphi(x_2), \dots, \varphi(x_n)]^T$ çekirdek çıktı fonksiyonudur. Buna göre çıktı fonksiyonu aşağıdaki şekilde ifade edilebilir:

$$f(x) = \boldsymbol{\varphi}^T(x)\boldsymbol{\phi}^T\boldsymbol{\alpha} = \boldsymbol{\varphi}^T(x)\boldsymbol{\phi}^T(I_n/C + \boldsymbol{\phi}\boldsymbol{\phi}^T)^{-1}\mathbf{y} \tag{32}$$

Burada $\mathbf{I}_n$, n-boyutlu birim matristir. Ridge teorisine göre $\mathbf{I}_n/C$ çıktı ağırlıklarının modelin genelleştirme performansını artırmaktadır (Fill ve Fishkind, 2000; Hoerl ve Kennard, 1970). Çekirdek matrisi ise aşağıdaki gibi tanımlıdır:

$$\mathbf{K} = \boldsymbol{\phi}\boldsymbol{\phi}^T : K_{i,j} = \boldsymbol{\varphi}(\mathbf{x_i}) * \boldsymbol{\varphi}(\mathbf{x_j}) = k(\mathbf{x_i}, \mathbf{x_j}) \tag{33}$$

Böylece çıktı fonksiyonunun kompakt formülü aşağıdaki eşitlikle gösterilir:

$$f(x) = [k(x, x_1), k(x, x_2), \dots, \; k(x, x_n)]\boldsymbol{\alpha} \tag{34}$$

Burada $\boldsymbol{\alpha} = (\mathbf{I}_n/C + \mathbf{K})^{-1}\mathbf{y}$, kullanılan çekirdek fonksiyonu ile elde edilen çıktı ağırlıklarıdır (Huang, 2014; Lu ve diğerleri., 2019).

2.3.1. Cezalı Cox Modelli Aşırı Öğrenme Makineleri (ELMCox)

ELMCox modelinde; cezalı Cox regresyon modelinde yer alan bağımsız değişkenlerin doğrusal birleşimi yerine ELM sinir ağı modelinin doğrusal olmayan çıktı fonksiyonu yer almaktadır. Buna göre, ELMCox modeline ilişkin kısmi logaritmik olabilirlik fonksiyonu aşağıdaki forma dönüşür:

$$pl_{cezalı}^{ELM}(\boldsymbol{\beta}, h_0) = \sum_{i=1}^{n} \delta_i f(x) - \sum_{i=1}^{n} \delta_i \log\left[\sum_{j \in R(\tau_i)} \exp\left(f(x)\right)\right] - \frac{1}{2}\lambda \boldsymbol{\beta}^T \boldsymbol{\beta} \qquad \lambda \geq 0 \qquad (35)$$

Eşitlik (35)'ten $\boldsymbol{\beta}$ katsayıları tahmin edilmektedir. ELMCox yöntemi, cezalı Cox yönteminin doğrusal olmayan bir uzantısı olarak kabul edilebilir (Dhillon ve Singh, 2020; Wang ve Li, 2019; Wang ve diğerleri., 2018; Wang ve Zhou, 2018). ELMCox modelinde bir birime ait hazard oranı aşağıdaki gibi hesaplanır (Dhillon ve Singh, 2020):

$$h_i(t) = h_0(t) * \exp\left(f(x_i)\right) \qquad (36)$$

2.3.2. Topluluk Öğrenme ve Cezalı Cox Modelli Aşırı Öğrenme Makineleri (ELMCoxEN)

ELMCoxEN modeli, ELMCox modeli ile elde edilen tahmin değerlerinin daha kararlı sonuçlar vermesi amacıyla ileri sürülmüştür. Bu modelde, rasgele orman algoritması tabanlı topluluk öğrenme yönteminden yararlanılmıştır (Dhillon ve Singh, 2020; Wang ve Zhou, 2018). Böylece, p sayıda bağımsız değişken arasından m sayıda bağımsız değişken içeren bootstrap örneklem rasgele oluşturulur. Regresyon problemlerinde genellikle $m = p/3$ iken sınıflandırma problemlerinde ise $m = \sqrt{p}$ şeklinde belirlenir (N. Altman ve Krzywinski, 2017; Breiman, 2001; James ve diğerleri., 2013; Wang ve diğerleri., 2018). Belirlenen örneklem sayısı kadar ELMCoxEN modeli kurulur. Regresyon problemlerinde her örneklemden elde edilen tahminlerin ortalaması alınırken sınıflandırma problemlerinde ise oylama yönteminden yararlanılır (Breiman, 2001, 2004).

2.3.3. Parçalı Uyarlanabilir Ridge Cezalı Cox Modelli Aşırı Öğrenme Makineleri (ELMCoxBAR)

Orantısal hazard ve regresyon katsayılarının doğrusallığı varsayımları cezalı Cox modellerinin potansiyel kısıtlılıklarıdır. Çünkü gerekli varsayımlar sağlanmadığında Cox regresyon modeli kurulması uygun değildir. Böyle durumlarda ELM gibi parametrik olmayan ve lineer olmayan sinir ağı modelleri kullanışlı olabilmektedir. ELMCoxBAR modelinde BAR ceza terimli ELMCox modeli oluşturulur. Bu model, Eşitlik (4)'te yer alan bağımsız değişkenlerin doğrusal birleşimleri ile sinir ağının çıktı katmanı fonksiyonunun yer

değiştirilmesiyle oluşturulur (Dhillon ve Singh, 2020; Wang ve Li, 2019; Wang ve Zhou, 2018; Yang ve diğerleri., 2021). Çıktı fonksiyonu Eşitlik (33)'teki çekirdek fonksiyonu olmak üzere ELMCoxBAR algoritmasına ilişkin uygulama adımları aşağıdaki gibidir (Wang ve Li, 2019):

(a) Uygun bir çekirdek fonksiyonu belirlenir ve $\mathbf{K}_{nxn}$ çekirdek matrisi hesaplanır.

(b) Eşitlik (4)'teki doğrusal fonksiyon ile $\mathbf{K}_{nxn}\boldsymbol{\beta}$ yer değiştirir. Böylece ELMCoxBAR modelinin kısmi logaritmik olabilirlik fonksiyonu aşağıdaki gibi elde edilir:

$$\mathrm{pl}(\boldsymbol{\beta}) = \sum_{i=1}^{n} \delta_i \mathbf{K}_{nxn}\boldsymbol{\beta} - \sum_{i=1}^{n} \delta_i \log\left[\sum_{j\in R(\tau_i)} \exp(\mathbf{K}_{nxn}\boldsymbol{\beta})\right] - 0.5\mathrm{In}(n) \sum_{j=1}^{p} \frac{\beta_j^2}{\hat{\beta}_j^2} \qquad (37)$$

(c) $\mathrm{pl}(\boldsymbol{\beta})$ eşitliğini maksimum yapacak $\boldsymbol{\beta}$ değerleri iteratif olarak Newton-Raphson algoritması ile hesaplanır. Burada $\hat{\beta}^{(0)}$ başlangıç değeri olarak ridge ceza terimi kullanılır. Ridge ceza terimi en optimal değere eşit olmak zorunda değildir. $\mathrm{In}(n)$ değerine eşit olacak şekilde seçilebilir.

(d) $\boldsymbol{\beta}$ değerleri istenen değere yakınsayana kadar ya da iterasyon sayısı belirlenen değere ulaşana kadar (b) ve (c) adımları tekrarlanır.

(e) Yeni bir birim için $\mathbf{x}$ bağımsız değişkenleri ile yapılan hazard oranı tahmini aşağıdaki gibi hesaplanır:

$$h_i(t) = h_0(t)\exp\left(f(\mathbf{x})\right) = h_0(t)\exp\left(\sum_{j=1}^{n} K\left(\mathbf{x}, \mathbf{x}_j\right)\beta_j\right) \qquad (38)$$

2.3.4. Olabilirlik Tabanlı Boosting Aşırı Öğrenme Makineleri (ELMCoxBoost)

ELMCoxBoost yönteminde, ELM yönteminin sonuçlarını güçlendirmek için sağkalım verilerine olabilirlik tabanlı boosting algoritması uygulanır (Wang ve Zhou, 2018).

CoxBoost algoritması, ceza teriminin değerini gerçek ve tahmin değerleri arasındaki farkın fonksiyonu şeklinde tanımlanmasını sağlayan kayıp fonksiyonu olarak negatif L_2-cezalı kısmi logaritmik olabilirlik fonksiyonunu kullanır (Binder ve Schumacher, 2009):

$$\mathrm{pl}_{\mathrm{cezalı}}(\boldsymbol{\beta}) = \mathrm{pl}(\boldsymbol{\beta}) - 0.5\lambda\beta^{T}\mathbf{I}\boldsymbol{\beta} \qquad (39)$$

Burada **I**, pxp boyutlu birim matrisi ifade etmektedir. Parametre tahminlerinin iteratif olarak güncellenmesinin devamı için yukarıdaki eşitliğe bir öteleme terimi $\hat{\eta} = \mathbf{x}^{T}\widehat{\boldsymbol{\beta}}$ dahil edildikten sonra elde edilen olabilirlik tabanlı-cezalı kısmi logaritmik olabilirlik fonksiyonu aşağıdaki gibidir (Binder ve Schumacher, 2009):

$$\mathrm{pl}^{\mathrm{O.T}}_{\mathrm{cezalı}}\left(\boldsymbol{\beta}\middle|\widehat{\boldsymbol{\beta}}\right) = \sum_{i=1}^{n} \delta_i \left[\hat{\eta}_i + \mathbf{x}_i\boldsymbol{\beta} - \log\left(\sum_{j\in R(\tau_i)} \exp\{\hat{\eta}_j + \mathbf{x}_j\boldsymbol{\beta}\}\right)\right] - \frac{\lambda}{2}\boldsymbol{\beta}^{T}\mathbf{I}\boldsymbol{\beta} \tag{40}$$

Her bir boosting iterasyonunda, cezalı kısmi logaritmik olabilirlik fonksiyonunu maksimum yapan değer, olası güncellemelerin hesaplanmasında kullanılır. Olabilirlik tabanlı boosting algoritması, p>n durumunda bağımsız değişkenler üzerinden ve kısıtlanmış p adet kısmi logaritmik olabilirlik fonksiyonunda $\left(\mathrm{pl}(\beta_j)\right)$ uygulanır. Her boosting iterasyonunda, kısıtlanmış kısmi logaritmik olabilirlik fonksiyonları $\hat{\beta}_j$ değerine doğru yer değiştirilerek kısıtlanmış, olabilirlik tabanlı-cezalı kısmi logaritmik olabilirlik fonksiyonları elde edilir (De Bin, 2016):

$$\mathrm{pl}^{\mathrm{O.T}}_{\mathrm{cezalı}}\left(\beta_j\middle|\widehat{\boldsymbol{\beta}}\right) = \frac{\partial \mathrm{pl}^{\mathrm{O.T}}_{\mathrm{cezalı}}\left(\beta_j\middle|\widehat{\boldsymbol{\beta}}\right)}{\partial(\beta_j)} \tag{41}$$

Yukarıdaki kısmi türev çözümünü maksimum yapan değerler aday güncellemelerdir ve cezalı kısmi logaritmik olabilirlik fonksiyonunu maksimum yapan değer öteleme terimine eklenir. Aşağıda olabilirlik tabanlı boosting algoritmasının adımları özetlenmiştir (Binder ve diğerleri., 2009; Binder ve diğerleri., 2013; De Bin, 2016):

1. Başlangıç değeri $\widehat{\boldsymbol{\beta}} = (0, \ldots, 0)$ olarak atanır.
2. Kısıtlanmış maksimum kısmi olabilirlik tahmininin sıfır civarındaki birinci dereceden yakınsaması $\left(\mathrm{pl}^{\mathrm{O.T}}_{\beta_j}\left(0\middle|\widehat{\boldsymbol{\beta}}\right)\right)$ tarafından olası güncellemeler hesaplanır:

$$\hat{b}_j^{\mathrm{O.T}} = \frac{\mathrm{pl}^{\mathrm{O.T}}_{\beta_j}\left(0\middle|\widehat{\boldsymbol{\beta}}\right)}{-\mathrm{pl}^{\mathrm{O.T}}_{\beta_j\beta_j}\left(0\middle|\widehat{\boldsymbol{\beta}}\right)} \tag{42}$$

3. En iyi güncelleme değeri seçilir:

$$j^* = \text{argmin}_{1\le j\le p} pl^{O.T}_{\beta_j}\left(0|\widehat{\boldsymbol{\beta}}\right)^2 / \left[-pl^{O.T}_{\beta_j\beta_j}\left(0|\widehat{\boldsymbol{\beta}}\right)\right] \quad (43)$$

4. Tahmin değeri $\hat{\beta}_{j*} = \hat{\beta}_{j^*} + \hat{b}^{O.T}_{j^*}$ olarak güncellenir.

2. ve 3. adımlar iterasyon sayısı kadar tekrarlanır. Burada $pl^{O.T}_{\beta_j\beta_j}\left(\beta_j|\widehat{\boldsymbol{\beta}}\right) = \frac{\partial^2 pl^{O.T}_{cezalı}\left(\beta_j|\widehat{\boldsymbol{\beta}}\right)}{\partial\beta_j^2}$ olarak ifade edilir.

2.3.5. Model Tabanlı Boosting Aşırı Öğrenme Makineleri (ELMmBoost)

ELMmBoost yönteminde ELM yönteminin sonuçlarını güçlendirmek için sağkalım verilerine model tabanlı boosting (mBoost) algoritması uygulanır (Dhillon ve Singh, 2020; Wang ve Zhou, 2018). mBoost algoritması, $\mathbf{u}$ gradyan vektörü, $L(y, F(x))$ negatif kısmi logaritmik olabilirlik fonksiyonu ve $\hat{h}(u, x_j)$ en küçük kareler tahmincisi olmak üzere $(\mathbf{x}^T\mathbf{x})^{-1}\mathbf{x}^T\mathbf{u}$ ifadesinde fonksiyonel olarak azalan gradyan algoritmasının doğrudan bir uygulamasıdır (De Bin, 2016). mBoost algoritmasının uygulama adımları aşağıda verilmiştir:

1. Başlangıç değeri $\widehat{\boldsymbol{\beta}} = (0, \ldots, 0)$ olarak atanır.
2. Negatif gradyan vektörü hesaplanır:

$$\mathbf{u}^{(i)} = \delta^{(i)} - \sum_{l\in R^{(i)}} \delta^{(l)} \frac{\exp\{\mathbf{x}^{(l)T}\widehat{\boldsymbol{\beta}}\}}{\sum_{k\in R^{(l)}} \exp\{\mathbf{x}^{(k)T}\widehat{\boldsymbol{\beta}}\}} \quad (44)$$

3. Negatif gradyan vektörüne en küçük kareler tahmincisi uygulanarak olası güncellemeler hesaplanır:

$$\hat{b}_j = \left(\mathbf{x}_j^T\mathbf{x}_j\right)^{-1}\mathbf{x}_j^T\mathbf{u} \quad (45)$$

Burada $\hat{b}_j$, zayıf tahmincidir.

4. Kayıp fonksiyonunu minimum yapan en iyi güncelleme değeri seçilir:

$$j^* = \text{argmin}_j \sum_{i=1}^{n} \left(\mathbf{u}^{(i)} - \mathbf{x}_j^{(i)}\hat{b}_j\right)^2 \quad (46)$$

5. Tahmin değeri güncellenir:

$$\hat{\beta}_{j^*} = \hat{\beta}_{j^*} + v\hat{b}_{j^*} \quad (47)$$

2. – 5. adımlar iterasyon sayısı kadar tekrarlanır. mBoost algoritması $\mathrm{pl}(\boldsymbol{\beta})$'nın gradyan vektörünü $\mathrm{F}(\mathrm{x_j}, \boldsymbol{\beta})$'ya göre hesaplar. Bu ise kısmi logaritmik olabilirlik fonksiyonunun eğiminin yerel olarak en dik olduğu noktaya karşılık gelir (Bühlmann ve Hothorn, 2007; De Bin, 2016; Hothorn ve diğerleri., 2010; Ridgeway, 2020).

2.4. Model Değerlendirme Ölçütleri

Çalışmamızdaki sağkalım modellerinin sağkalım süresi tahmin performansları için C-indeks ve IBS hesaplanırken; kısa dönem sağkalım durumu tahmin performansları için duyarlılık, özgüllük, doğruluk oranları ile NTO, PTO, F_1 skoru, AUC, AUPR, kappa ve MKK ölçütleri hesaplandı.

C-İndeks: C-indeks, rastgele seçilen iki birimden prognostik skoru daha yüksek olanının diğer birimden daha uzun yaşama olasılığını hesaplayan bir uyum iyiliği ölçütüdür (Harrell ve diğerleri., 1982; Harrell Jr ve diğerleri., 1996). Ölüm olayına kadar geçen sürenin tahmininde C-indeks, en az bir tanesi ölmüş olan tüm olası gözlem çiftleri dikkate alınarak hesaplanmaktadır. Eğer bir gözlem çiftini oluşturan birimlerden birine ilişkin sağkalım süresi tahmin değeri diğer birimden daha büyük, ve bu birim gerçekte de daha uzun yaşamış ise, bu gözlem çiftine ilişkin elde edilen tahmin değerlerinin gerçek durumla uyumlu olduğu söylenir. Eğer bir gözlem çiftinde yer alan birimlerin ikisi de aynı zamanda öldüyse ya da biri ölmüş, diğeri sansürlü ise bu çiftin uyumsuz olduğu, ve C-indeksin hesaplanmasında kullanılamayacağı söylenir (Harrell Jr ve diğerleri., 1996). C-indeks, [0,1] aralığında değer alır. C-indeks değerinin 1'e yakın olması modelin tahmin performansının yüksek olduğunu göstermektedir (Harrell ve diğerleri., 1982; Harrell Jr ve diğerleri., 1996; Wang ve Li, 2019).

C-indeks, uyumlu gözlem çiftlerinin karşılaştırılabilir tüm gözlem çiftlerinin sayısına oranlanması ile aşağıdaki gibi hesaplanır:

$$\mathrm{C-İndeks} = \frac{\sum_{i,j} I(\tilde{T}_i > \tilde{T}_j) * I(\eta_i > \eta_j) * \Delta_j}{\sum_{i,j} I(\tilde{T}_i > \tilde{T}_j) * \Delta_j} \tag{48}$$

Burada i ve j örneklemdeki gözlem çiftlerini, $\tilde{T}_i$ ve $\tilde{T}_j$ bu gözlem çiftlerine ilişkin sağkalım sürelerini, η_i ve η_j ise tahmin değerlerini belirtmektedir. Δ_j ise, karşılaştırılan sağkalım süresi

çiftleri arasında değeri küçük olan sağkalım süresine sahip olan gözlem sansürlü ise, bu gözlem çiftini hesaplamanın dışında bırakan çarpım faktörüdür; yani Δ_j=0'dır (Harrell ve diğerleri., 1982; Schmidt ve diğerleri., 2008).

İntegrali Alınmış Brier Skoru: BS, zamana bağlı bir fonksiyondur ve tahmin edilen sağkalım fonksiyonu $\hat{S}(t|\mathbf{x})$ ve test verisi arasında, sansürün ters olasılığı ile ağırlıklandırılmış hata kareler ortalaması olarak kabul edilir. Test verisindeki birim sayısı n, indikatör fonksiyonu I ve sansür dağılımının Kaplan-Meier tahmini ise $\widehat{G}(t)$ olmak üzere BS aşağıdaki eşitlikle hesaplanır (Graf ve diğerleri., 1999; Kronek ve Reddy, 2008):

$$BS(t) = \frac{1}{n}\sum_{i=1}^{n}\left\{\left[0 - \hat{S}(t|\mathbf{x_i})\right]^2 I(\tau_i \leq t, \delta_i = 1)\left(1/\widehat{G}(\tau_i)\right) + \left[1 - \hat{S}(t|\mathbf{x_i})\right]^2 I(\tau_i > t)(1/\widehat{G}(t))\right\} \quad (49)$$

IBS ise modelin tüm zamanlardaki tahmin değerlerinin genel bir ölçüsüdür ve aşağıdaki gibi hesaplanır (Graf ve diğerleri., 1999; Kronek ve Reddy, 2008):

$$IBS = \frac{1}{\max(\tau_i)}\int_0^{\max(\tau_i)} BS(t)dt \quad (50)$$

IBS, [0,1] aralığında değer alır. IBS değerinin 0'a yakın olması modelin tahmin performansının yüksek olduğunu göstermektedir (Wang ve diğerleri., 2018).

Duyarlılık Oranı: Duyarlılık oranı, bir model tarafından doğru olarak tahmin edilen gerçek pozitiflerin oranıdır (D. G. Altman ve Bland, 1994). Gerçek ve tahmin değerlerinin birbiri ile karşılaştırılması sonucu gerçek pozitif (GP), yalancı negatif (YN), yalancı pozitif (YP) ve gerçek negatif (GN) frekans değerlerinden oluşan karar matrisi $M = \begin{pmatrix} GP & YP \\ YN & GN \end{pmatrix}$ gösterilmek üzere duyarlılık oranı aşağıdaki gibi hesaplanır (Chicco ve Jurman, 2020):

$$\text{Duyarlılık oranı} = \frac{GP}{GP+YN} \quad (51)$$

Özgüllük Oranı: Özgüllük oranı, bir model tarafından doğru olarak tahmin edilen gerçek negatiflerin oranıdır (N. Altman ve Krzywinski, 2017):

$$\text{Özgüllük oranı} = \frac{GN}{GN+YP} \quad (52)$$

Doğruluk Oranı: Doğruluk oranı, bir model tarafından doğru olarak tahmin edilen tüm birimlerin, veri setinde yer alan tüm birimlerin sayısına oranıdır (McNeil ve Adelstein, 1976):

$$\text{Doğruluk oranı} = \frac{\text{GP+GN}}{\text{GP+GN+YP+YN}} \tag{53}$$

Negatif Tanımlama Oranı: NTO, modelin doğru olarak tahmin ettiği negatif birimlerin, modelin negatif olarak tahmin ettiği tüm birimlere oranıdır (McNeil ve Adelstein, 1976):

$$\text{NTO} = \frac{\text{GN}}{\text{GN+YN}} \tag{54}$$

Pozitif Tanımlama Oranı: PTO, modelin doğru olarak tahmin ettiği pozitif birimlerin, modelin pozitif olarak tahmin ettiği tüm birimlere oranıdır (McNeil ve Adelstein, 1976):

$$\text{PTO} = \frac{\text{GP}}{\text{GP+YP}} \tag{55}$$

F_1 Skoru: F_1 skoru, duyarlılık oranı ve PTO'nun harmonik ortalaması alınarak hesaplanan bir model değerlendirme ölçütüdür. F_1 skoru [0,1] aralığında değer alır ve 1'e yaklaştıkça modelin performansı artar (Chicco ve Jurman, 2020):

$$F_1 \text{ skoru} = \frac{2}{(\text{Duyarlılık oranı})^{-1}+(\text{PTO})^{-1}} = 2 * \frac{\text{Duyarlılık oranı}*\text{PTO}}{\text{Duyarlılık oranı+PTO}} \tag{56}$$

ROC Eğrisi Altında Kalan Alan: AUC; farklı kesim noktaları için yapılan sınıflandırma tahminlerine ilişkin hesaplanan duyarlılık oranı ve yalancı pozitiflik oranlarının birleştirilmesi ile elde edilen ROC eğrisi altında kalan alandır. AUC, (0,1) arasında değer alır. AUC değerinin 1'e yakın olması modelin tahmin performansının yüksek olduğunu belirtir (Fawcett, 2006; Hanley ve McNeil, 1982). ROC eğrisi $y = f(x)$ biçiminde matematiksel bir fonksiyon ile ifade edilmek üzere, AUC aşağıdaki gibi hesaplanır (Krzanowski ve Hand, 2009):

$$\text{AUC} = \int_0^1 f(x)dx \tag{57}$$

Pozitif Tanımlama Oranı-Duyarlılık Eğrisi Altında Kalan Alan: AUPR, farklı kesim noktaları için yapılan sınıflandırma tahminlerine ilişkin hesaplanan PTO ve duyarlılık oranının birleştirilmesi ile elde edilen eğri altında kalan alandır. AUPR, (0,1) arasında değer alır. AUPR değerinin 1'e yakın olması modelin tahmin performansının yüksek olduğunu belirtir (Davis ve Goadrich, 2006; Keilwagen ve diğerleri., 2014).

Cohen'in Kappa Katsayısı: Kappa katsayısı, isimsel ölçekli gözlenen ve beklenen değerler arasındaki uyum oranının bir fonksiyonudur. Kappa katsayısı, [0,1] aralığında değer alır. Kappa katsayısının 1'e eşit olması, model tarafından tahmin edilen sınıf değerleri ile gerçek sınıf değerleri arasında mükemmel uyum olduğunu, yani modelin sınıflandırma performansının mükemmel olduğunu; 0'a eşit olması ise modelin tahmin performasının çok kötü olduğunu belirtir (Cohen, 1960; Warrens, 2015). p_0, gerçek sınıf değerleri ile modelin sınıf tahmin değerleri arasında gözlenen uyum oranı ve p_c, gerçek sınıf değerleri ile modelin sınıf tahmin değerleri arasında tesadüfen uyum olması beklenen birimlerin oranı olmak üzere kappa katsayısı aşağıdaki gibi hesaplanır (Cohen, 1960):

$$\text{Kappa} = \frac{p_o - p_c}{1 - p_c} \tag{58}$$

Matthews Korelasyon Katsayısı: MKK; dengesiz sınıf dağılımından etkilenmeyen, gerçek ve tahmin edilen değerler arasındaki Pearson çarpım moment korelasyon katsayısıdır (Powers, 2020). MKK, [-1, +1] aralığında değer alır. MKK'nin -1'e yakın değer alması modelin sınıflandırma tahmin performansının düşük; +1'e yakın değer alması ise modelin sınıflandırma tahmin performansının yüksek olduğunu gösterir (Chicco ve Jurman, 2020). MKK, aşağıdaki eşitlikle hesaplanır (Chicco ve diğerleri., 2021):

$$\text{MKK} = \frac{\text{GP} * \text{GN} - \text{YP} * \text{YN}}{\sqrt{(\text{GP}+\text{YP}) * (\text{GP}+\text{YN}) * (\text{GN}+\text{YP}) * (\text{GN}+\text{YN})}} \tag{59}$$

3. GEREÇ VE YÖNTEM

Bu çalışmanın uygulama bölümü yüksek boyutlu gen ifade verilerinde sağkalım süresi ve kısa dönem sağkalım durumu tahmini olmak üzere iki kısımdan oluşmaktadır. Bu amaçla n=200 birim ve p=1000 gen ifadesinden oluşan yüksek boyutlu sağkalım verileri türetildi. Türetilen veriler Cox-L_2, DTBA, ELMCox, ELMCoxBAR, ELMCoxBoost, ELMCoxEN ve ELMmBoost yöntemleri ile analiz edildi. 1000 döngü ile gerçekleştirilen simülasyon sonunda elde edilen sonuçlara göre sağkalım modelleri karşılaştırıldı.

3.1. Simülasyon Algoritması

Bu çalışmada gerçekleştirilen simülasyon algoritması iki aşamadan oluşmaktadır. İlk aşamada sağkalım süresi tahmini, ikinci aşamada ise kısa dönem sağkalım durumu tahmini gerçekleştirilmiştir. Bu amaçla yazılan simülasyon algoritmaları, sırasıyla "Simülasyon Algoritması – I" ve "Simülasyon Algoritması – II" başlıkları altında açıklanmıştır. Simülasyon algoritmasına ilişkin akış şeması ise Şekil 2'de belirtilmiştir.

3.1.1. Simülasyon Algoritması – I

1. Örneklem büyüklüğü (n=200) ve gen ifade sayısı (p=1000) için sabit bir değer atandı.
2. $\boldsymbol{\mu} \sim \text{Uniform}(0,1)$ ve $\boldsymbol{\sigma} \sim \text{Uniform}(0,1)$ dağılımından ortalama ve standart sapma değerleri türetildi.
3. Uniform dağılımdan türetilen ortalama ve standart sapma değerleri ile korelasyon düzeyi -0,7 ve +0,7 değerleri arasında rastgele değişmek üzere $\mathbf{X} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ bağımsız değişkenleri çok değişkenli normal dağılımdan türetildi.
4. İlk 200 gen ifade değerinin toplamı μ_T olmak üzere, sağkalım süresi değerleri $\mathbf{T} \sim N(\mu_T; 0{,}01)$ dağılımından türetildi.
5. Sağkalım süresine ilişkin değerler, en büyük değer 120 ay olacak şekilde standartlaştırıldı.
6. Sansür oranı %10, %30, %50 ve %70 olacak şekilde sansür süresi değerleri $\mathbf{C} \sim \text{Uniform}(0, \mathbf{T}/\text{sansür oranı})$ dağılımından türetildi.
7. δ_i sansür durumu aşağıdaki tanımına göre ifade edildi:

$$\delta_i = \begin{cases} 0, & T_i \leq C_i \\ 1, & T_i > C_i \end{cases} \qquad i: 1,2, \dots, n$$

8. Veri seti, 70:30 oranında eğitim-test seti olarak rasgele ikiye ayrıldı.
9. Eğitim setinde Cox-L_2, DTBA, ELMCox, ELMCoxBAR, ELMCoxBoost, ELMCoxEN ve ELMmBoost yöntemleri ile sağkalım süresi tahmin modelleri oluşturuldu.
10. Oluşturulan modellerin test setinde sağkalım süresi tahminine ilişkin performanslarının belirlenmesi için C-indeks ve IBS performans ölçütleri hesaplandı.
11. 2. – 10. adımlar 1000 kez tekrar edildi.
12. Elde edilen sonuçlara aşamalı kümeleme analizi uygulanarak dendrogram grafikleri çizildi. Böylece birbirine yakın ve uzak performans gösteren yöntemler belirlendi.

3.1.2. Simülasyon Algoritması – II

1. Simülasyon algoritması – I'in 8. adımında elde edilen eğitim ve test setlerindeki sağkalım süresi değişkeni; sağkalım süresi <60 ay olan birimler "kısa dönem sağkalım", ≥60 ay olan birimler ise "uzun dönem sağkalım" grubu olacak şekilde iki sınıfa ayrıldı.
2. 9. adımda elde edilen sağkalım süresi tahmin değerlerinden kısa dönem sağkalım olasılıkları tahmin edildi.
3. Oluşturulan modellerin test setinde kısa dönem sağkalım durumu tahminine ilişkin performanslarının belirlenmesi için duyarlılık, özgüllük, doğruluk oranı, NTO, PTO, AUPR, AUC, F_1 skoru, kappa katsayısı ve MKK performans ölçütleri hesaplandı.
4. 1. – 3. adımlar 1000 kez tekrar edildi.
5. Elde edilen sonuçlara aşamalı kümeleme analizi uygulanarak dendrogram grafikleri çizildi. Böylece birbirine yakın ve uzak performans gösteren yöntemler belirlendi.

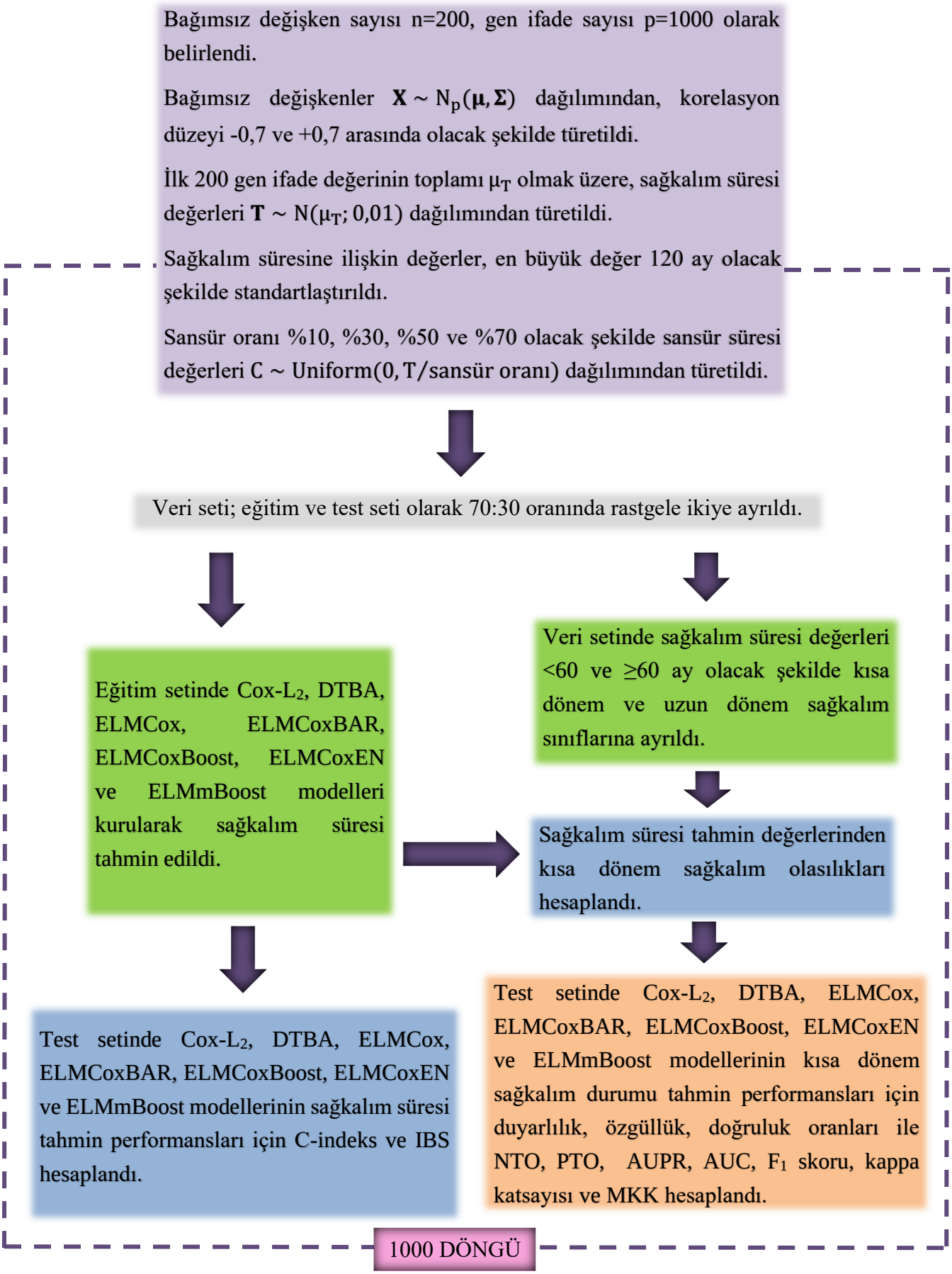
3.2. Tahmin Modellerine İlişkin Parametreler

Çalışmada kullanılan tüm SELM modellerinde denemeler sonucu belirlenen c=0,5 parametre değeri ile doğrusal çekirdek fonksiyonu kullanılmıştır. Bunun yanı sıra, ELMCox ve ELMCoxEN modellerinde ceza terimi olarak L_2 parametresi kullanılmıştır. ELMCoxEN modelinde, bootstrap yeniden örnekleme yöntemi ile seçilecek temel model sayısı 100, her modele seçilecek bağımsız değişken sayısı ise p/3 olarak belirlenmiştir. ELMCox, ELMCoxEN

ve Cox-L_2 modellerinde, hata kareler ortalamasını en küçük yapan λ parametresi çapraz geçerlik ile hesaplanmıştır.

3.3. Kullanılan Programlar

Bu çalışmada yer alan uygulamalar R programlama dilinin 4.0.5 versiyonu kullanılarak gerçekleştirildi. Veri türetimi, analizi ve sonuçların kaydedilmesi için R programlama dilinde haven, survival, superpc, data.table, SurvELM, ELMSurv, mboost, Coxboost, glmnet, Metrics, dplyr, survcomp, stats, PMCMR, tsutils, DescTools, dynsurv, survAUC, openxlsx, raster, caret, GMCM, MASS, Matrix, MBESS, evolqg, mltools, pROC, PRROC, matrixcalc, lqmm, truncnorm, EnvStats ve qlcMatrix paketleri kullanıldı.



Şekil 2. Simülasyon algoritması akış şeması

4. BULGULAR

Cox-L_2, DTBA, ELMCox, ELMCoxBAR, ELMCoxBoost, ELMCoxEn ve ELMmBoost yöntemlerinin performanslarının karşılaştırılması için hesaplanan C-indeks, IBS, duyarlılık, özgüllük, doğruluk oranı, NTO, PTO, AUPR, AUC, F_1 skoru, kappa katsayısı ve MKK değerlerinin normal dağılıma uygunluğu Kolmogorov-Smirnov analizi ile test edildi. Performans ölçütlerinin dağılımı normal dağılıma uygunluk göstermediği için modellerin performanslarına ilişkin tanımlayıcı istatistikler medyan (25. – 75. Persantil) şeklinde verildi.

n=200 birim arasından sansür oranı %10, %30, %50 ve %70 olacak şekilde türetilen veri setleri ile oluşturulan Cox-L_2, DTBA, ELMCox, ELMCoxBAR, ELMCoxBoost, ELMCoxEn ve ELMmBoost modellerinin sağkalım süresi ve kısa dönem sağkalım durumu tahmini performanslarına ilişkin bulgular sırasıyla Bölüm 4.1. ve Bölüm 4.2.'de verilmiştir.

4.1. Sağkalım Süresi Tahminine İlişkin Bulgular

Değişen sansür oranına sahip veri setlerinde yöntemlerin sağkalım süresi tahminindeki C-indeks ve IBS performanslarına ilişkin bulgular Tablo 1 ve Şekil 3-4'te verilmiştir.

C-indeks değerlerinin medyan değişim aralığı sansür oranı %10 için 0,746-0,796; %30 için 0,739-0,798; %50 için 0,726-0,791; %70 için ise 0,708-0,784'tür. Cox-L_2, DTBA, ELMCox, ELMCoxBAR, ELMCoxBoost, ELMCoxEn ve ELMmBoost yöntemlerinin C-indeks değerlerine ilişkin medyan değişim aralığı ise sırasıyla 0,778-0,797; 0,773-0,777; 0,768-0,792; 0,762-0,787; 0,784-0,798; 0,761-0,778 ve 0,708-0,746'dır. Tüm yöntemlerin C-indeks değerleri değişen sansür oranlarına göre incelendiğinde, tüm durumlarda en yüksek C-indeks değerine sahip yöntemin ELMCoxBoost, en düşük C-indeks değerine sahip yöntemin ise ELMmBoost olduğu görülmektedir. Tüm yöntemlerin C-indeks bakımından birbirine çok yakın sonuçlar verdiği ve sansür oranı arttıkça ELMCox, ELMCoxBAR, ELMCoxEN ve ELMmBoost yöntemlerinin C-indeks performanslarının azaldığı gözlenmektedir (Şekil 3).

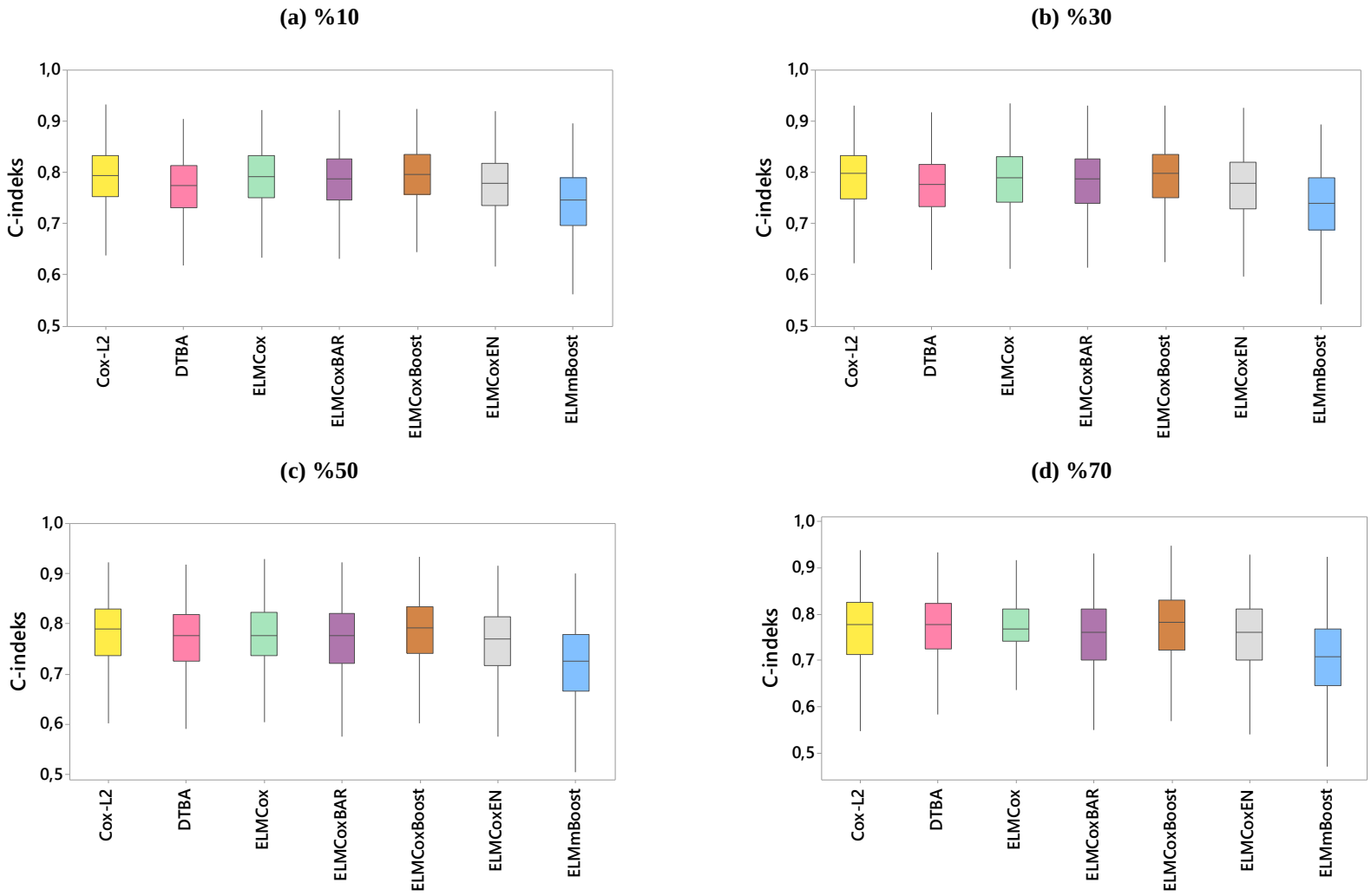
Sağkalım yöntemlerinin IBS değerlerinin medyan değişim aralığı sansür oranı %10 için 0,072-0,096; %30 için 0,077-0,103; %50 için 0,086-0,109; %70 için ise 0,101-0,120'dir. Cox-L_2, DTBA, ELMCox, ELMCoxBAR, ELMCoxBoost, ELMCoxEn ve ELMmBoost yöntemlerinin IBS değerlerine ilişkin medyan değişim aralığı ise sırasıyla 0,074-0,109; 0,096-

0,120; 0,074-0,111; 0,076-0,106; 0,072-0,101; 0,078-0,108 ve 0,087-0,116'dır. Tüm yöntemlerin IBS değerleri değişen sansür oranlarına göre incelendiğinde, tüm durumlarda en düşük IBS değerine sahip yöntemin ELMCoxBoost, en yüksek IBS değerine sahip yöntemin ise DTBA olduğu görülmektedir. Ayrıca, tüm yöntemlerin IBS bakımından birbirine çok yakın sonuçlar verdiği ve sansür oranı arttıkça sağkalım yöntemlerinin tümüne ilişkin IBS değerlerinin arttığı gözlenmektedir (Şekil 4).

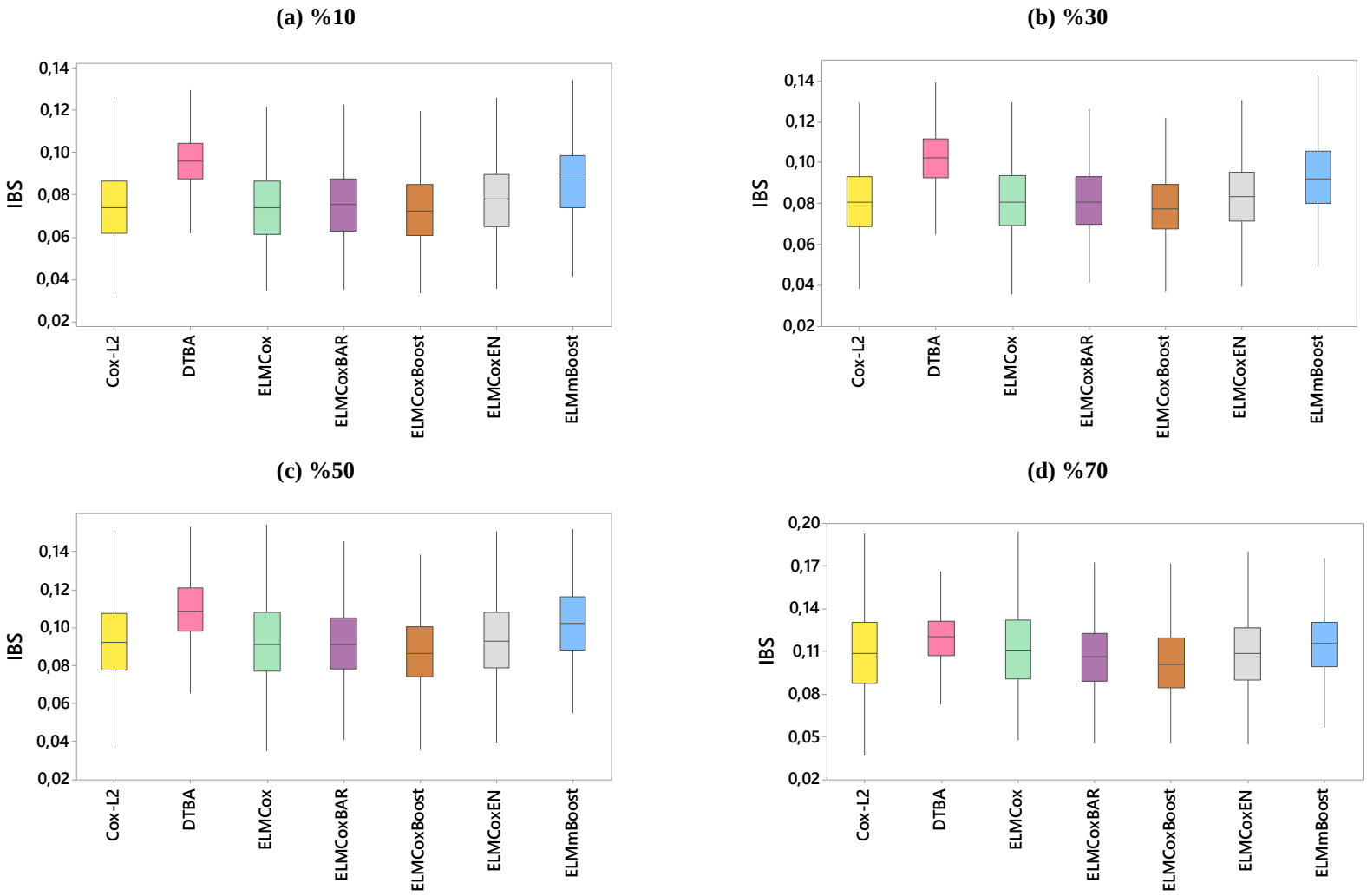
C-indeks ve IBS sonuçlarına göre sağkalım yöntemlerinin değişen sansür oranına göre performansları birbirine yakın olduğu için yöntemler arasındaki ilişkilerin belirlenebilmesi amacıyla C-indeks ve IBS sonuçları kullanılarak aşamalı kümeleme analizi yapılmış ve bu analiz sonucu elde edilen dendrogram grafikleri Şekil 5'te verilmiştir. Sansür oranı %10 ve %30 için ELMCoxBoost, Cox-L_2, ELMCox ve ELMCoxBAR modelleri benzer performans göstermektedir (Şekil 5a-b). Sansür oranı %50'ye çıkarıldığında ELMCox, Cox-L_2, ELMCoxBoost modellerinin aynı kümede yer aldıkları; benzer şekilde ELMCoxEN, ELMCoxBAR modellerinin de birbirine yakın performans göstererek bir küme içerisinde yer aldıkları görülmektedir (Şekil 5c). Sansür oranı %70 için ELMCox, ELMCoxEN ve ELMCoxBAR; ELMCoxBoost, Cox-L_2 ve DTBA yöntemleri birbirine yakın performans göstererek iki ayrı kümede toplanmıştır (Şekil 5d). Tüm senaryolarda ELMCoxBoost ve Cox-L_2 yöntemlerinin aynı küme içinde yer aldığı; ELMmBoost yönteminin ise diğer tüm yöntemlerden ayrıştığı görülmektedir (Şekil 5a-d).

Tablo 1. Değişen sansür oranlarına göre yöntemlerin C-indeks ve IBS değerlerine ilişkin tanımlayıcı istatistikler

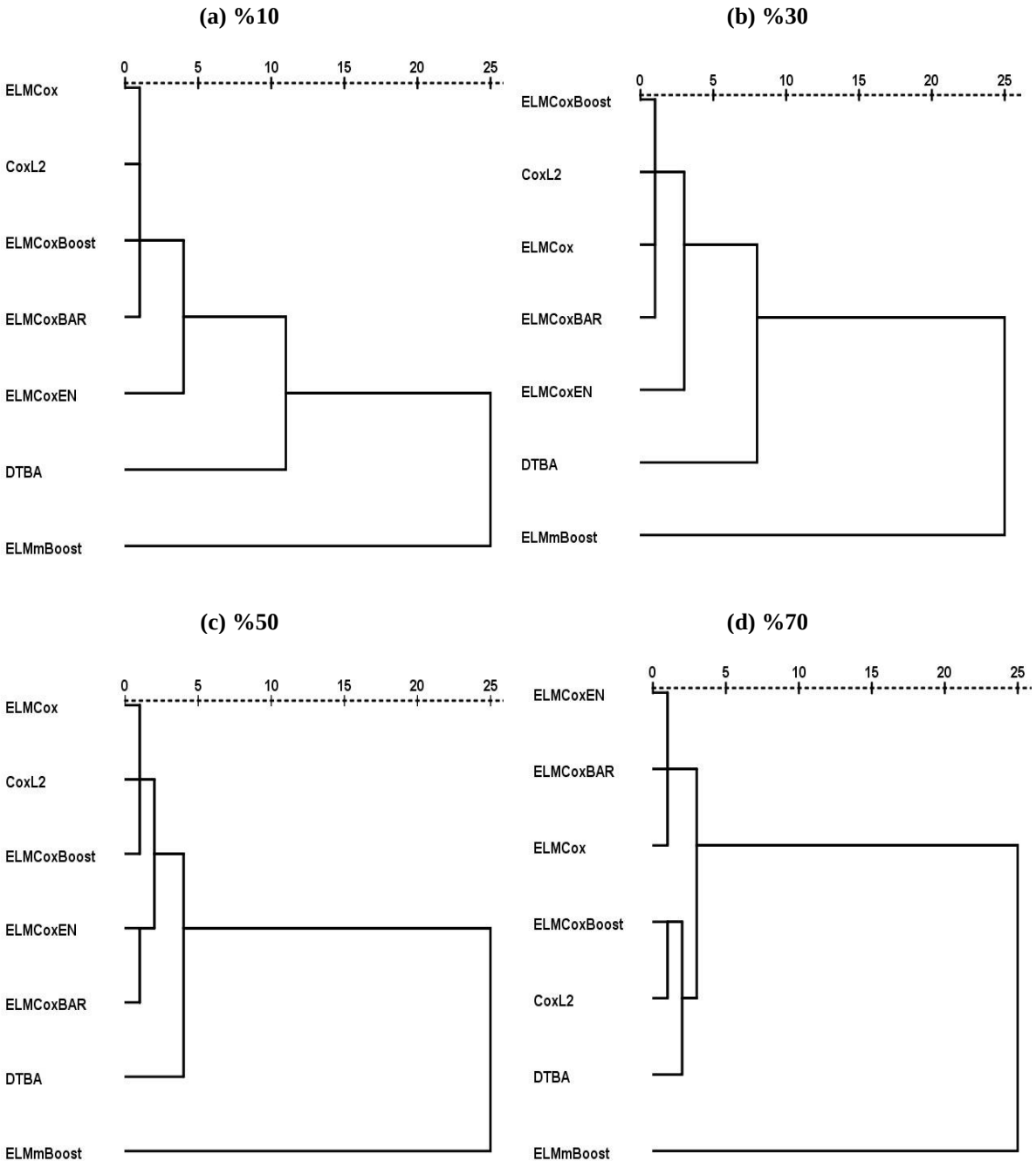
	Yöntem	Sansür Oranı			
		%10	%30	%50	%70
C-indeks	Cox-L_2	0,794 (0,753 - 0,833)	0,797 (0,747 - 0,832)	0,788 (0,737 - 0,830)	0,778 (0,713 - 0,826)
	DTBA	0,773 (0,731 - 0,814)	0,776 (0,732 - 0,815)	0,775 (0,725 - 0,817)	0,777 (0,726 - 0,823)
	ELMCox	0,792 (0,749 - 0,831)	0,790 (0,741 - 0,830)	0,785 (0,730 - 0,827)	0,768 (0,742 - 0,813)
	ELMCoxBAR	0,787 (0,747 - 0,826)	0,786 (0,739 - 0,825)	0,776 (0,721 - 0,819)	0,762 (0,700 - 0,811)
	ELMCoxBoost	0,796 (0,756 - 0,834)	0,798 (0,751 - 0,835)	0,791 (0,740 - 0,834)	0,784 (0,723 - 0,830)
	ELMCoxEN	0,778 (0,736 - 0,818)	0,778 (0,728 - 0,819)	0,770 (0,716 - 0,814)	0,761 (0,702 - 0,811)
	ELMmBoost	0,746 (0,696 - 0,788)	0,739 (0,688 - 0,788)	0,726 (0,666 - 0,779)	0,708 (0,645 - 0,769)
IBS	Cox-L_2	0,074 (0,062 - 0,087)	0,081 (0,069 - 0,093)	0,092 (0,078 - 0,108)	0,109 (0,088 - 0,130)
	DTBA	0,096 (0,087 - 0,105)	0,103 (0,093 - 0,112)	0,109 (0,098 - 0,121)	0,120 (0,107 - 0,131)
	ELMCox	0,074 (0,062 - 0,087)	0,080 (0,069 - 0,093)	0,091 (0,077 - 0,108)	0,111 (0,090 - 0,132)
	ELMCoxBAR	0,076 (0,063 - 0,088)	0,078 (0,067 - 0,089)	0,091 (0,078 - 0,105)	0,106 (0,089 - 0,123)
	ELMCoxBoost	0,072 (0,061 - 0,085)	0,077 (0,067 - 0,089)	0,086 (0,074 - 0,101)	0,101 (0,084 - 0,119)
	ELMCoxEN	0,078 (0,065 - 0,090)	0,084 (0,072 - 0,096)	0,093 (0,079 - 0,108)	0,108 (0,090 - 0,126)
	ELMmBoost	0,087 (0,074 - 0,099)	0,092 (0,080 - 0,106)	0,102 (0,088 - 0,116)	0,116 (0,099 - 0,130)



Şekil 3. Değişen sansür oranlarına göre sağkalım yöntemlerinin C-indeks değerlerine ilişkin kutu grafikleri



Şekil 4. Değişen sansür oranlarına göre sağkalım yöntemlerinin IBS değerlerine ilişkin kutu grafikleri



Şekil 5. Değişen sansür oranlarına göre modellerin sağkalım süresi tahmin performansları arasındaki ilişkileri gösteren dendrogram grafikleri

4.2. Kısa Dönem Sağkalım Durumu Tahminine İlişkin Bulgular

Sağkalım yöntemlerinin değişen sansür oranına sahip veri setlerinde kısa dönem sağkalım durumu tahminindeki duyarlılık, özgüllük, doğruluk oranı ile NTO ve PTO değerlerine ilişkin bulgular Tablo 2 ve Şekil 6-10'da verilmiştir.

Duyarlılık oranlarının medyan değişim aralığı sansür oranı %10 için 0,815-0,875; %30 için 0,821-0,872; %50 için 0,800-0,870; %70 için ise 0,773-0,897'dir. Cox-L_2, DTBA, ELMCox, ELMCoxBAR, ELMCoxBoost, ELMCoxEn ve ELMmBoost yöntemlerinin duyarlılık oranlarına ilişkin medyan değişim aralığı ise sırasıyla 0,833-0,875; 0,806-0,842; 0,865-0,897; 0,833-0,864; 0,850-0,875; 0,833-0,851 ve 0,773-0,821'dir (Tablo 2 ve Şekil 6).

Sansür oranı %10, %30, %50 ve %70 için özgüllük oranlarının medyan değişim aralığı sırasıyla 0,780-0,813; 0,769-0,815; 0,763-0,818 ve 0,722-0,800'dür. Cox-L_2, DTBA, ELMCox, ELMCoxBAR, ELMCoxBoost, ELMCoxEn ve ELMmBoost yöntemlerinin özgüllük oranlarına ilişkin medyan değişim aralığı ise sırasıyla 0,795-0,818; 0,781-0,800; 0,722-0,813; 0,784-0,808; 0,800-0,815; 0,778-0,807 ve 0,750-0,780'dir (Tablo 2 ve Şekil 7).

Doğruluk oranı değerlerinin medyan değişim aralığı sansür oranı %10, %30, %50 ve %70 için sırasıyla 0,783-0,833; 0,767-0,833; 0,767-0,817 ve 0,750-0,817'dir. Cox-L_2, DTBA, ELMCox, ELMCoxBAR, ELMCoxBoost, ELMCoxEn ve ELMmBoost yöntemlerinin doğruluk oranlarına ilişkin medyan değişim aralığı ise sırasıyla 0,800-0,833; 0,783-0,800; 0,783-0,817; 0,783-0,817; 0,817-0,833; 0,783,0,817 ve 0,750-0,783'tür (Tablo 2 ve Şekil 8).

NTO değerlerinin medyan değişim aralığı sansür oranı %10 için 0,806-0,867; %30 için 0,800-0,857; %50 için 0,800-0,857; %70 için ise 0,768-0,833'tür. Cox-L_2, DTBA, ELMCox, ELMCoxBAR, ELMCoxBoost, ELMCoxEn ve ELMmBoost yöntemlerinin NTO değerlerine ilişkin medyan değişim aralığı ise sırasıyla 0,826-0,867; 0,800-0,833; 0,816-0,857; 0,820-0,849; 0,833-0,863; 0,815-0,840 ve 0,768-0,806'dır (Tablo 2 ve Şekil 9).

PTO değerlerinin medyan değişim aralığı sansür oranı %10 için 0,786-0,827; %30 için 0,778-0,824; %50 için 0,771-0,821; %70 için ise 0,750-0,810'dur. Cox-L_2, DTBA, ELMCox, ELMCoxBAR, ELMCoxBoost, ELMCoxEn ve ELMmBoost yöntemlerinin PTO değerlerine ilişkin medyan değişim aralığı ise sırasıyla 0,810-0,821; 0,790-0,810; 0,750-0,822; 0,794-0,818; 0,810-0,827; 0,794-0,815 ve 0,763-0,786'dır (Tablo 2 ve Şekil 10).

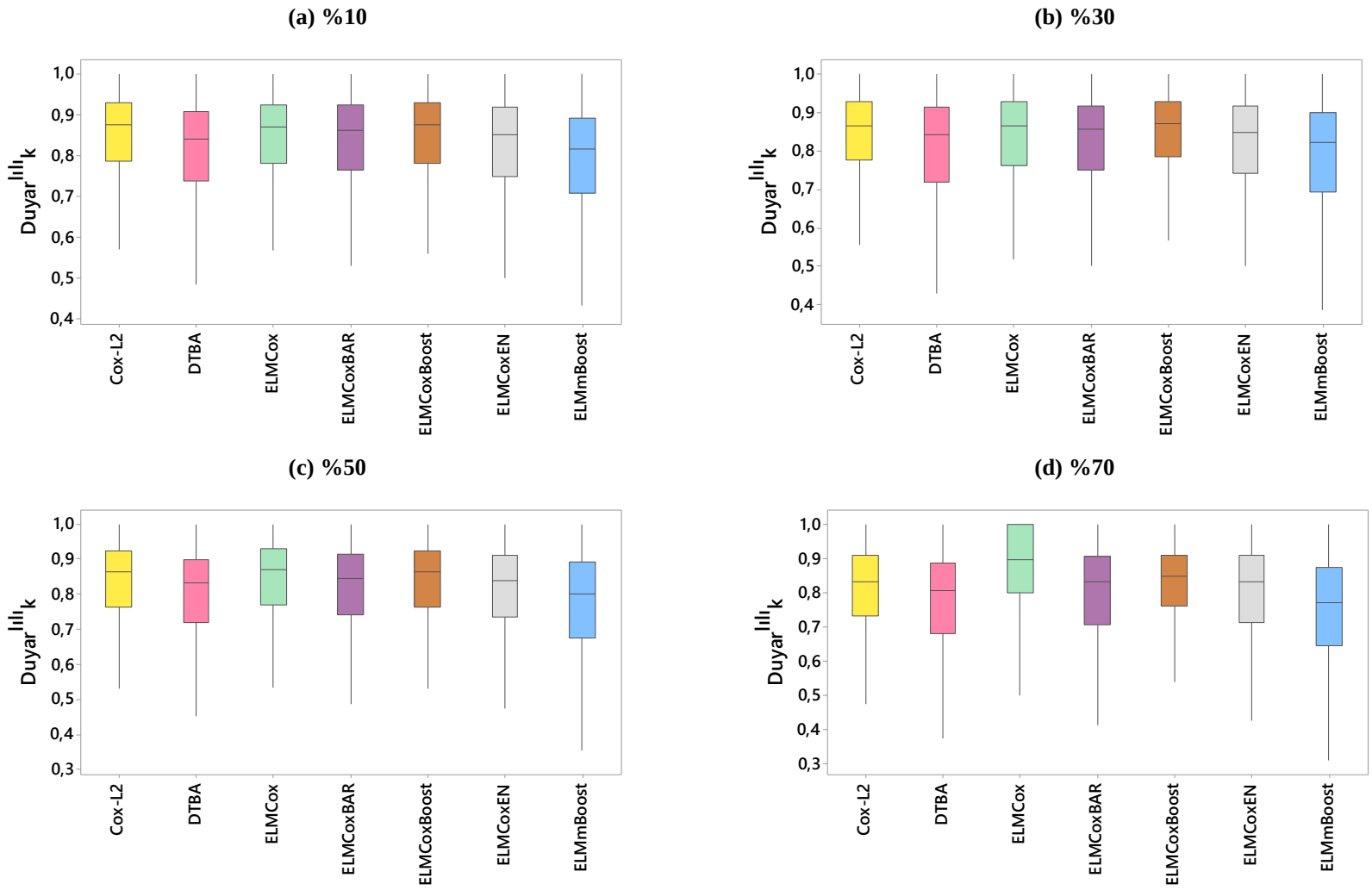
Tüm sağkalım yöntemlerinin kısa dönem sağkalım durumu tahminine ilişkin duyarlılık, özgüllük, doğruluk oranı ile NTO ve PTO değerleri değişen sansür oranına göre incelendiğinde yöntemlerin birbirine yakın performans gösterdikleri belirlenmiştir. Buna ek olarak, sansür oranı arttıkça duyarlılık oranı bakımından Cox-L_2, ELMCoxBAR, ELMCoxBoost ve ELMCoxEN; özgüllük oranı bakımından DTBA, ELMCox, ELMCoxBAR ve ELMmBoost; NTO bakımından DTBA, ELMCox ve ELMCoxBAR; PTO bakımından ise ELMCox, ELMCoxBoost, ELMCoxEN ve ELMmBoost yöntemlerinin performanslarının azaldığı görülmektedir. Sağkalım yöntemlerinin doğruluk oranı performansları ise değişen sansür oranına göre daha kararlı sonuçlar göstermiş olsa da gözlemlerin yarısından fazlası sansürlü olduğunda Cox-L_2, ELMCox, ELMCoxBAR, ELMCoxBoost ve ELMCoxEN yöntemlerinin performanslarının azaldığı belirlenmiştir (Tablo 2 ve Şekil 6-10).

Tablo 2. Değişen sansür oranlarına göre yöntemlerin duyarlılık, özgüllük, doğruluk oranları ile NTO ve PTO değerlerine ilişkin tanımlayıcı istatistikler

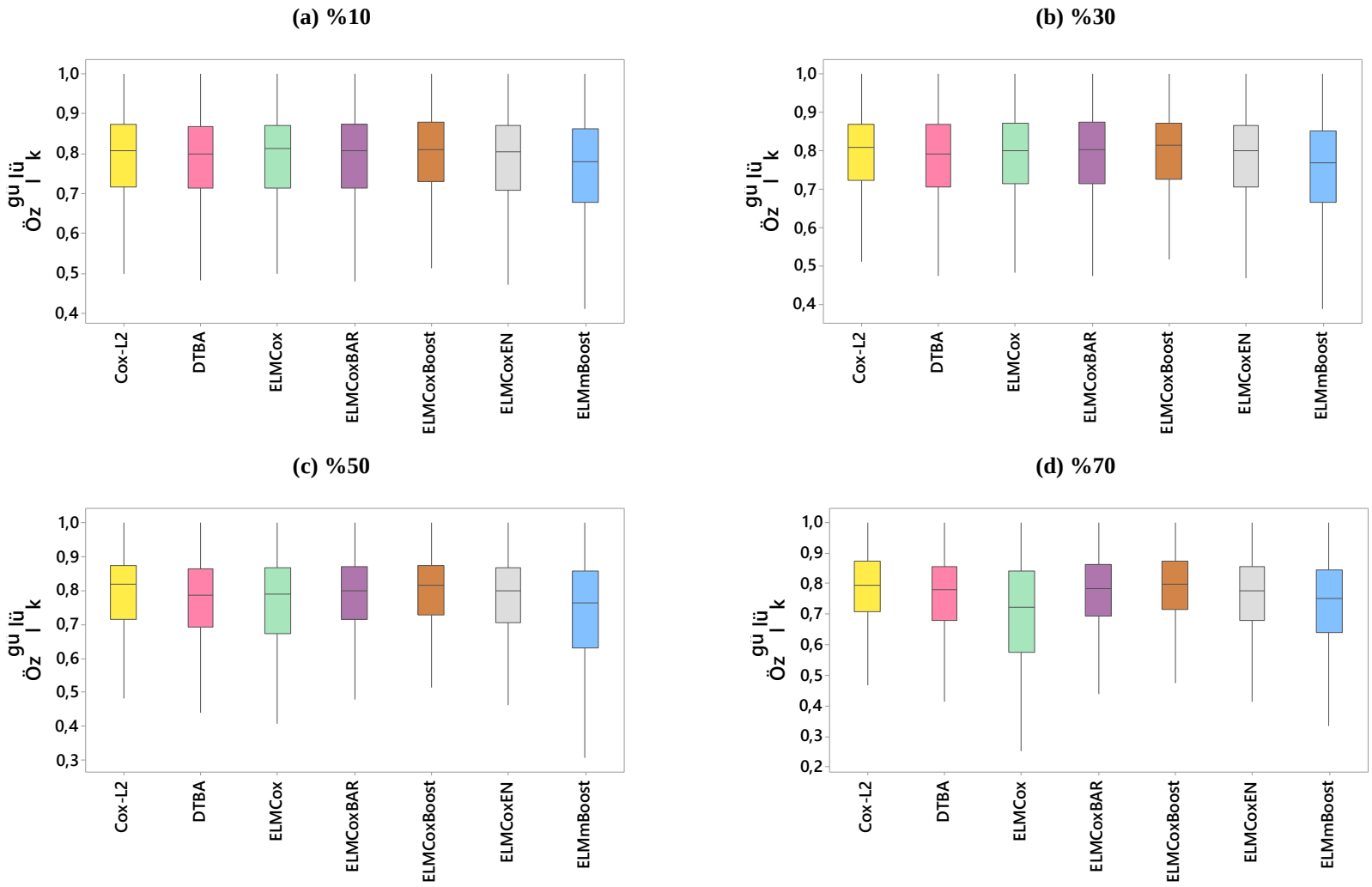
	Yöntem	Sansür Oranı			
		%10	%30	%50	%70
Duyarlılık	Cox-L_2	0,875 (0,786 - 0,931)	0,867 (0,778 - 0,929)	0,864 (0,765 - 0,923)	0,833 (0,733 - 0,909)
	DTBA	0,840 (0,738 - 0,909)	0,842 (0,719 - 0,914)	0,833 (0,720 - 0,900)	0,806 (0,682 - 0,889)
	ELMCox	0,871 (0,781 - 0,926)	0,865 (0,764 - 0,929)	0,870 (0,769 - 0,931)	0,897 (0,800 – 1,000)
	ELMCoxBAR	0,864 (0,767 - 0,926)	0,857 (0,750 - 0,919)	0,846 (0,742 - 0,913)	0,833 (0,708 - 0,906)
	ELMCoxBoost	0,875 (0,781 - 0,931)	0,872 (0,784 - 0,930)	0,865 (0,765 - 0,923)	0,850 (0,762 - 0,912)
	ELMCoxEN	0,851 (0,750 - 0,920)	0,849 (0,742 - 0,917)	0,840 (0,735 - 0,911)	0,833 (0,714 - 0,909)
	ELMmBoost	0,815 (0,708 - 0,893)	0,821 (0,695 - 0,900)	0,800 (0,677 - 0,892)	0,773 (0,646 - 0,875)
Özgüllük	Cox-L_2	0,808 (0,718 - 0,874)	0,809 (0,724 - 0,868)	0,818 (0,714 - 0,874)	0,795 (0,708 - 0,875)
	DTBA	0,800 (0,714 - 0,868)	0,793 (0,707 - 0,870)	0,788 (0,693 - 0,865)	0,781 (0,680 - 0,857)
	ELMCox	0,813 (0,714 - 0,872)	0,800 (0,714 - 0,871)	0,790 (0,674 - 0,867)	0,722 (0,574 - 0,840)
	ELMCoxBAR	0,808 (0,714 - 0,874)	0,802 (0,714 - 0,875)	0,800 (0,714 - 0,872)	0,784 (0,693 - 0,865)
	ELMCoxBoost	0,810 (0,731 - 0,880)	0,815 (0,727 - 0,871)	0,815 (0,727 - 0,875)	0,800 (0,714 - 0,875)
	ELMCoxEN	0,807 (0,710 - 0,871)	0,800 (0,706 - 0,867)	0,800 (0,706 - 0,869)	0,778 (0,679 - 0,857)
	ELMmBoost	0,780 (0,677 - 0,862)	0,769 (0,667 - 0,852)	0,763 (0,630 - 0,857)	0,750 (0,641 - 0,846)
Doğruluk	Cox-L_2	0,817 (0,767 - 0,867)	0,833 (0,767 - 0,867)	0,817 (0,767 - 0,867)	0,800 (0,750 - 0,850)
	DTBA	0,800 (0,750 - 0,850)	0,800 (0,750 - 0,850)	0,800 (0,733 - 0,833)	0,783 (0,717 - 0,817)
	ELMCox	0,817 (0,767 - 0,867)	0,817 (0,767 - 0,867)	0,800 (0,733 - 0,850)	0,783 (0,650 - 0,833)
	ELMCoxBAR	0,817 (0,767 - 0,867)	0,817 (0,767 - 0,867)	0,800 (0,750 - 0,850)	0,783 (0,733 - 0,846)
	ELMCoxBoost	0,833 (0,783 - 0,867)	0,833 (0,771 - 0,867)	0,817 (0,767 - 0,867)	0,817 (0,767 - 0,850)
	ELMCoxEN	0,817 (0,750 - 0,850)	0,817 (0,750 - 0,850)	0,800 (0,733 - 0,850)	0,783 (0,733 - 0,833)
	ELMmBoost	0,783 (0,717 - 0,833)	0,767 (0,717 - 0,833)	0,767 (0,700 - 0,817)	0,750 (0,683 - 0,800)

Tablo 2. Değişen sansür oranlarına göre yöntemlerin duyarlılık, özgüllük, doğruluk oranları ile NTO ve PTO değerlerine ilişkin tanımlayıcı istatistikler (Devam)

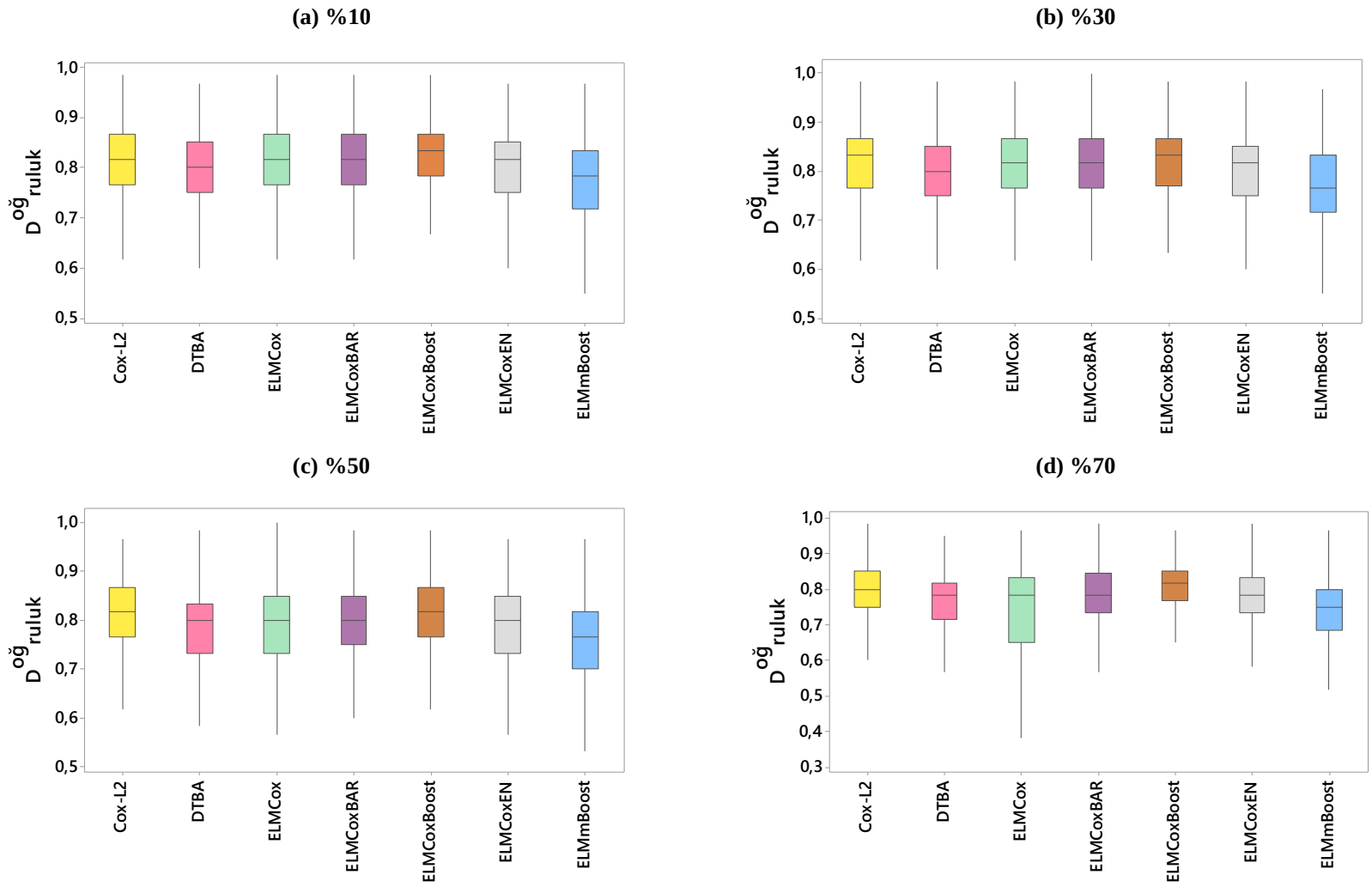
Yöntem		Sansür Oranı			
		%10	%30	%50	%70
NTO	Cox-L_2	0,867 (0,771 - 0,926)	0,855 (0,778 - 0,917)	0,857 (0,774 - 0,920)	0,826 (0,750 - 0,900)
	DTBA	0,833 (0,742 - 0,905)	0,830 (0,737 - 0,900)	0,826 (0,728 - 0,897)	0,800 (0,700 - 0,877)
	ELMCox	0,857 (0,772 - 0,926)	0,849 (0,768 - 0,909)	0,840 (0,781 - 0,906)	0,816 (0,788 - 0,875)
	ELMCoxBAR	0,849 (0,760 - 0,929)	0,842 (0,758 - 0,912)	0,839 (0,750 - 0,906)	0,820 (0,727 - 0,895)
	ELMCoxBoost	0,863 (0,781 - 0,923)	0,857 (0,774 - 0,920)	0,857 (0,767 - 0,920)	0,833 (0,757 - 0,903)
	ELMCoxEN	0,840 (0,756 - 0,913)	0,833 (0,758 - 0,909)	0,839 (0,746 - 0,904)	0,815 (0,727 - 0,897)
	ELMmBoost	0,806 (0,712 - 0,886)	0,800 (0,700 - 0,882)	0,800 (0,706 - 0,875)	0,768 (0,667 - 0,857)
PTO	Cox-L_2	0,818 (0,739 - 0,877)	0,821 (0,744 - 0,875)	0,818 (0,739 - 0,871)	0,810 (0,725 - 0,875)
	DTBA	0,810 (0,731 - 0,865)	0,810 (0,730 - 0,865)	0,794 (0,710 - 0,857)	0,790 (0,690 - 0,857)
	ELMCox	0,822 (0,742 - 0,875)	0,821 (0,739 - 0,872)	0,800 (0,692 - 0,868)	0,750 (0,600 - 0,849)
	ELMCoxBAR	0,818 (0,733 - 0,871)	0,818 (0,746 - 0,875)	0,806 (0,719 - 0,868)	0,794 (0,710 - 0,865)
	ELMCoxBoost	0,827 (0,750 - 0,875)	0,824 (0,750 - 0,875)	0,821 (0,733 - 0,875)	0,810 (0,733 - 0,875)
	ELMCoxEN	0,815 (0,724 - 0,872)	0,807 (0,731 - 0,871)	0,800 (0,711 - 0,862)	0,794 (0,704 - 0,862)
	ELMmBoost	0,786 (0,692 - 0,857)	0,778 (0,682 - 0,857)	0,771 (0,668 - 0,842)	0,763 (0,652 - 0,840)



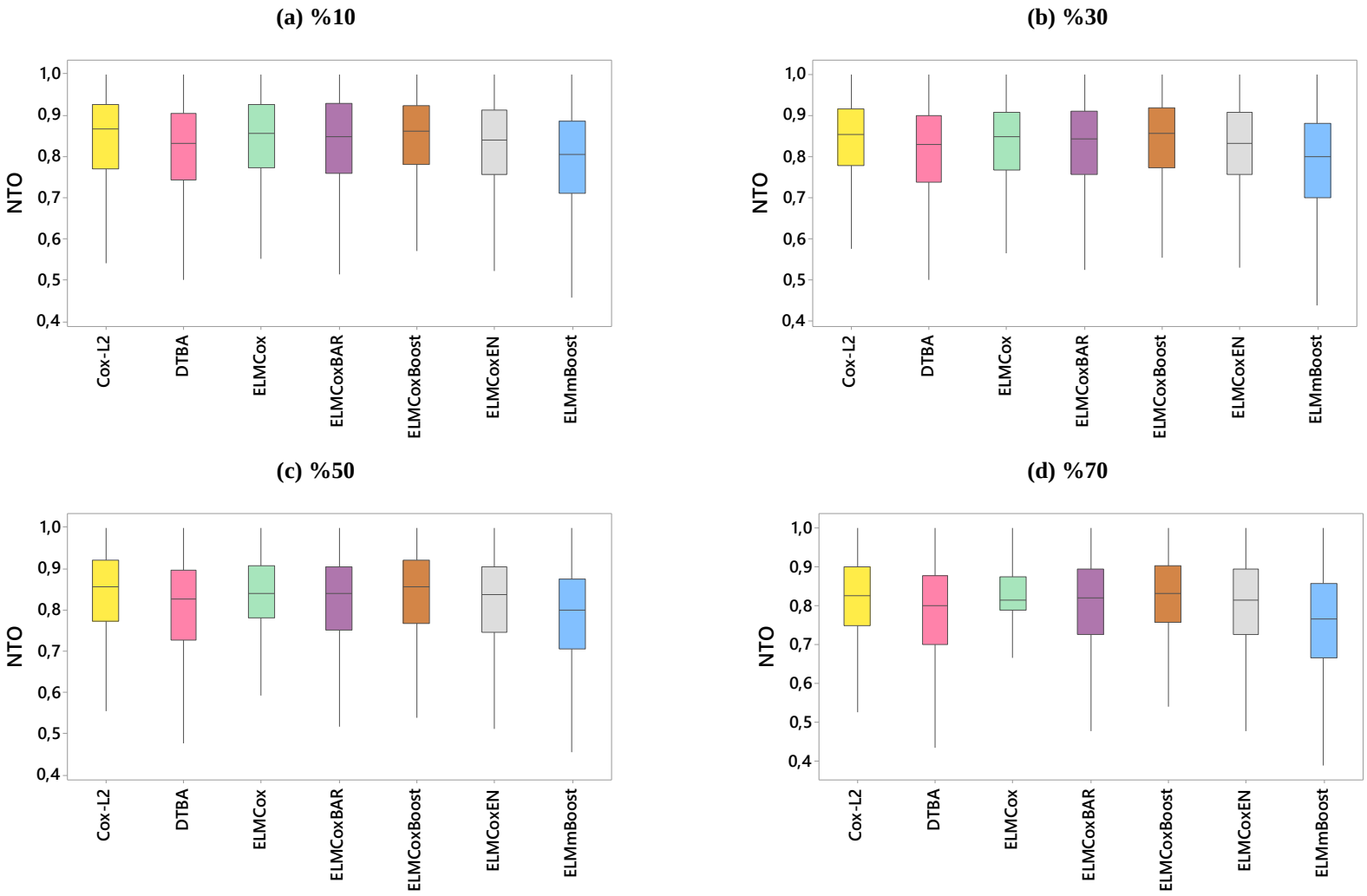
Şekil 6. Değişen sansür oranlarına göre sağkalım yöntemlerinin duyarlılık oranlarına ilişkin kutu grafikleri



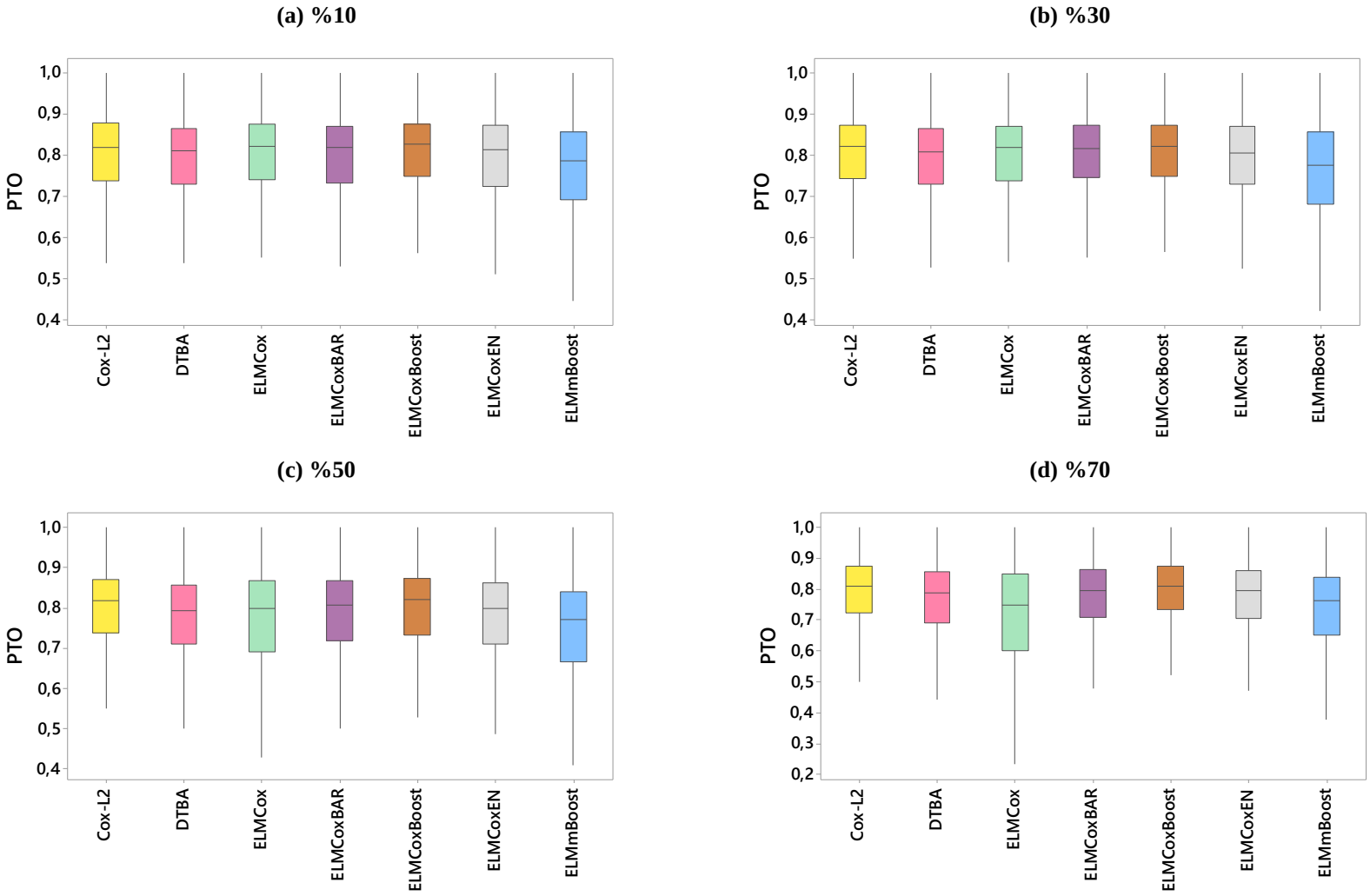
Şekil 7. Değişen sansür oranlarına göre sağkalım yöntemlerinin özgüllük oranlarına ilişkin kutu grafikleri



Şekil 8. Değişen sansür oranlarına göre sağkalım yöntemlerinin doğruluk oranlarına ilişkin kutu grafikleri



Şekil 9. Değişen sansür oranlarına göre sağkalım yöntemlerinin NTO değerlerine ilişkin kutu grafikleri



Şekil 10. Değişen sansür oranlarına göre sağkalım yöntemlerinin PTO değerlerine ilişkin kutu grafikleri

Sağkalım yöntemlerinin değişen sansür oranına sahip veri setlerinde kısa dönem sağkalım durumu tahminindeki AUPR, AUC, F_1 skoru, kappa katsayısı ve MKK değerlerine ilişkin bulgular Tablo 3 ve Şekil 11-15'te verilmiştir.

Sansür oranı %10, %30, %50 ve %70 için AUPR değerlerinin medyan değişim aralığı sırasıyla 0,839-0,895; 0,836-0,893; 0,817-0,889 ve 0,807-0,879'dur. Cox-L_2, DTBA, ELMCox, ELMCoxBAR, ELMCoxBoost, ELMCoxEn ve ELMmBoost yöntemlerinin AUPR değerlerine ilişkin medyan değişim aralığı ise sırasıyla 0,869-0,889; 0,835-0,866; 0,831-0,893; 0,861-0,890; 0,879-0,895; 0,858-0,880 ve 0,807-0,839'dur (Tablo 3 ve Şekil 11).

Sansür oranı %10, %30, %50 ve %70 için AUC değerlerinin medyan değişim aralığı sırasıyla 0,816-0,867; 0,808-0,865; 0,788-0,863 ve 0,776-0,851'dir. Cox-L_2, DTBA, ELMCox, ELMCoxBAR, ELMCoxBoost, ELMCoxEn ve ELMmBoost yöntemlerinin AUC değerlerine ilişkin medyan değişim aralığı ise sırasıyla 0,843-0,866; 0,816-0,845; 0,811-0,863; 0,829-0,858; 0,851-0,867; 0,829-0,850 ve 0,776-0,816'dır (Tablo 3 ve Şekil 12).

Sansür oranı %10, %30, %50 ve %70 için F_1 skoru değerlerinin medyan değişim aralığı sırasıyla 0,773-0,825; 0,772-0,824; 0,754-0,818 ve 0,739-0,808'dir. Cox-L_2, DTBA, ELMCox, ELMCoxBAR, ELMCoxBoost, ELMCoxEn ve ELMmBoost yöntemlerinin F_1 skoru değerlerine ilişkin medyan değişim aralığı ise sırasıyla 0,800-0,824; 0,769-0,800; 0,776-0,820; 0,787-0,815; 0,808-0,825; 0,783-0,809 ve 0,739-0,773'tür (Tablo 3 ve Şekil 13).

Tüm yöntemlerin AUPR, AUC ve F_1 skoru performansları değişen sansür oranına göre incelendiğinde; yöntemlerin birbirine yakın performans gösterdikleri ve sansür oranındaki artıştan olumsuz olarak etkilendikleri, sansür oranı arttıkça performanslarının da buna bağlı olarak azaldığı görülmektedir (Tablo 3 ve Şekil 11-13).

Sansür oranı %10, %30, %50 ve %70 için kappa değerlerinin medyan değişim aralığı sırasıyla 0,535-0,634; 0,526-0,633; 0,500-0,626 ve 0,470-0,600'dür. Cox-L_2, DTBA, ELMCox, ELMCoxBAR, ELMCoxBoost, ELMCoxEn ve ELMmBoost yöntemlerinin kappa değerlerine ilişkin medyan değişim aralığı ise sırasıyla 0,585-0,633; 0,532-0,599; 0,529-0,628; 0,561-0,622; 0,600-0,634; 0,558-0,602 ve 0,470-0,535'tir. Tüm yöntemlerin gerçek durum ile orta düzeyde uyumlu tahminlerde bulunduğu ve birbirine yakın performans gösterdiği görülmektedir. Buna karşılık, tüm senaryolarda gerçek durum ile en çok uyum gösteren tahminlerde bulunan modelin ELMCoxBoost, en az uyum gösteren tahminlerde bulunan modelin ise ELMmBoost olduğu belirlenmiştir. Ayrıca sansür oranı arttıkça yöntemlerin

tahmin değerleri ile gerçek durum arasındaki uyumun azaldığı da gözlenmektedir (Tablo 3 ve Şekil 14).

Sansür oranı %10, %30, %50 ve %70 için MKK değerlerinin medyan değişim aralığı sırasıyla 0,551-0,645; 0,541-0,642; 0,519-0,634 ve 0,489-0,609'dur. Cox-L_2, DTBA, ELMCox, ELMCoxBAR, ELMCoxBoost, ELMCoxEn ve ELMmBoost yöntemlerinin MKK değerlerine ilişkin medyan değişim aralığı ise sırasıyla 0,598-0,645; 0,546-0,606; 0,540-0,635; 0,576-0,633; 0,609-0,643; 0,571-0,613 ve 0,489-0,551'dir. Tüm yöntemlerin gerçek durum ile pozitif yönlü, orta düzeyde uyumlu tahminlerde bulunduğu ve birbirine yakın performans gösterdiği görülmektedir. Buna karşılık, gerçek durum ile en iyi uyum gösteren tahmin değerlerine sahip modelin ELMCoxBoost, en kötü uyum gösteren tahmin değerlerine sahip modelin ise ELMmBoost olduğu belirlenmiştir. Ayrıca sansür oranı arttıkça yöntemlerin gerçek durumu tahmin etmedeki uyumlarının azaldığı da gözlenmektedir (Tablo 3 ve Şekil 15).

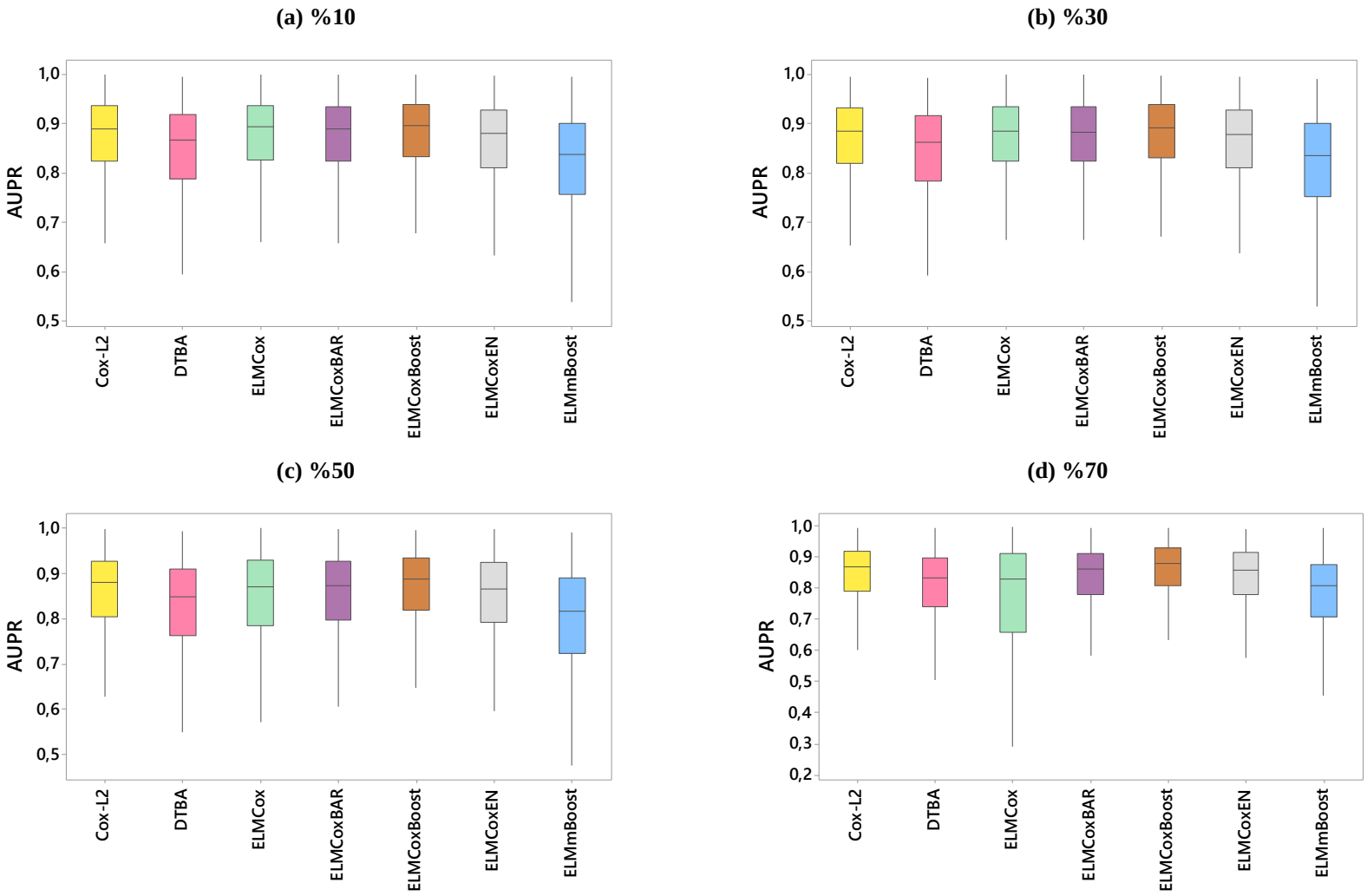
Tablo 2-3'te verilen sonuçlara göre sağkalım yöntemlerinin değişen sansür oranına göre performansları birbirine yakın olduğu için, yöntemler arasındaki ilişkilerin belirlenebilmesi amacıyla duyarlılık, özgüllük, doğruluk oranı, NTO, PTO, AUPR, AUC, F_1 skoru, kappa katsayısı ve MKK sonuçları kullanılarak aşamalı kümeleme analizi yapılmış ve bu analiz sonucu elde edilen dendrogramlar Şekil 16'da verilmiştir. Sansür oranı %10 için ELMCox, ELMCoxBAR, ELMCoxBoost ve Cox-L_2 yöntemlerinin aynı kümede yer aldıkları; benzer şekilde, DTBA ve ELMCoxEN yöntemlerinin de birbirine yakın performans göstererek bir küme oluşturdukları görülmektedir (Şekil 16a). Sansür oranı %30 için ise, tüm yöntemler arasından ELMCoxBoost ve Cox-L_2 yöntemlerinin birbirine yakın performans göstererek bir küme oluşturdukları belirlenmiştir. Buna ek olarak; ELMCox, ELMCoxBAR ve ELMCoxEN yöntemlerinin de birbirine yakın sonuçlar verdiği görülmüştür (Şekil 16b). Veri setindeki gözlemlerin %50'si sansürlü ise ELMCoxBoost ve Cox-L_2; ELMCoxEN, ELMCoxBAR, ELMCox ve DTBA yöntemlerinin birbirine benzer performans göstererek iki ayrı küme içerisinde yer aldıkları bulgusuna ulaşılmıştır (Şekil 16c). Sansür oranı %70'e çıkarıldığında ise ELMCoxBoost ve Cox-L_2 yöntemlerinin performanslarının birbirine yakın olduğu, bu yöntemlerin aynı küme içinde yer aldıkları belirlenmiştir. Ayrıca; DTBA, ELMCoxBAR ve ELMCoxEN yöntemlerinin de birbirine yakın performans gösterdikleri belirlenmiştir (Şekil 16d). Şekil 5'teki sonuçlara benzer şekilde; tüm senaryolarda ELMCoxBoost ve Cox-L_2 yöntemlerinin aynı küme içinde yer aldığı, ELMmBoost yönteminin ise diğer tüm yöntemlerden ayrıştığı görülmektedir (Şekil 16a-d).

Tablo 3. Değişen sansür oranlarına göre yöntemlerin AUPR, AUC, F_1 skoru, kappa katsayısı ve MKK değerlerine ilişkin tanımlayıcı istatistikler

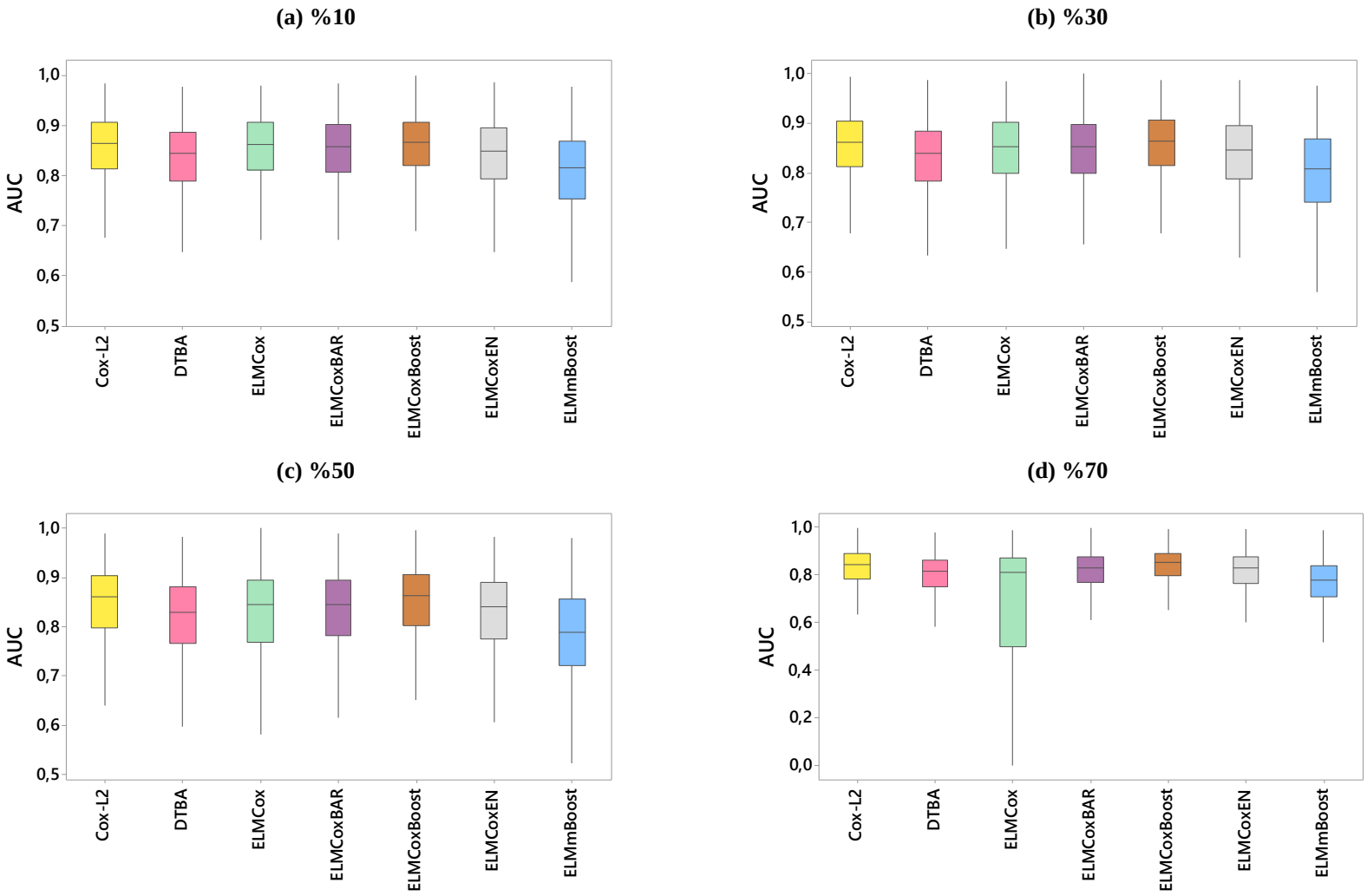
Yöntemler		Sansür Oranı			
		%10	%30	%50	%70
AUPR	Cox-L_2	0,889 (0,825 - 0,936)	0,886 (0,820 - 0,932)	0,880 (0,805 - 0,927)	0,869 (0,792 - 0,920)
	DTBA	0,866 (0,789 - 0,918)	0,862 (0,785 - 0,917)	0,849 (0,763 - 0,911)	0,835 (0,741 - 0,899)
	ELMCox	0,893 (0,827 - 0,938)	0,886 (0,824 - 0,935)	0,872 (0,785 - 0,930)	0,831 (0,658 - 0,911)
	ELMCoxBAR	0,890 (0,823 - 0,935)	0,883 (0,824 - 0,934)	0,874 (0,798 - 0,928)	0,861 (0,781 - 0,912)
	ELMCoxBoost	0,895 (0,833 - 0,939)	0,893 (0,833 - 0,940)	0,889 (0,818 - 0,935)	0,879 (0,809 - 0,929)
	ELMCoxEN	0,880 (0,810 - 0,928)	0,879 (0,811 - 0,927)	0,865 (0,793 - 0,924)	0,858 (0,779 - 0,915)
	ELMmBoost	0,839 (0,756 - 0,902)	0,836 (0,752 - 0,901)	0,817 (0,724 - 0,890)	0,807 (0,707 - 0,876)
AUC	Cox-L_2	0,866 (0,815 - 0,908)	0,862 (0,814 - 0,904)	0,861 (0,798 - 0,904)	0,843 (0,785 - 0,888)
	DTBA	0,845 (0,789 - 0,888)	0,839 (0,784 - 0,885)	0,830 (0,767 - 0,882)	0,816 (0,749 - 0,863)
	ELMCox	0,863 (0,812 - 0,907)	0,854 (0,800 - 0,903)	0,846 (0,769 - 0,895)	0,811 (0,500 - 0,872)
	ELMCoxBAR	0,858 (0,807 - 0,902)	0,853 (0,800 - 0,897)	0,845 (0,782 - 0,895)	0,829 (0,770 - 0,877)
	ELMCoxBoost	0,867 (0,820 - 0,908)	0,865 (0,814 - 0,907)	0,863 (0,803 - 0,905)	0,851 (0,795 - 0,891)
	ELMCoxEN	0,850 (0,795 - 0,896)	0,846 (0,789 - 0,895)	0,840 (0,776 - 0,890)	0,829 (0,764 - 0,877)
	ELMmBoost	0,816 (0,755 - 0,869)	0,808 (0,741 - 0,869)	0,788 (0,721 - 0,856)	0,776 (0,707 - 0,836)
F_1 skoru	Cox-L_2	0,824 (0,768 - 0,875)	0,821 (0,762 - 0,871)	0,815 (0,746 - 0,873)	0,800 (0,727 - 0,857)
	DTBA	0,800 (0,737 - 0,853)	0,800 (0,725 - 0,857)	0,787 (0,714 - 0,842)	0,769 (0,696 - 0,833)
	ELMCox	0,820 (0,755 - 0,872)	0,812 (0,746 - 0,872)	0,800 (0,727 - 0,862)	0,776 (0,696 - 0,840)
	ELMCoxBAR	0,815 (0,759 - 0,868)	0,811 (0,746 - 0,866)	0,800 (0,730 - 0,857)	0,787 (0,710 - 0,848)
	ELMCoxBoost	0,825 (0,767 - 0,877)	0,824 (0,765 - 0,875)	0,818 (0,750 - 0,872)	0,808 (0,744 - 0,862)
	ELMCoxEN	0,809 (0,745 - 0,857)	0,807 (0,742 - 0,861)	0,794 (0,722 - 0,852)	0,783 (0,711 - 0,846)
	ELMmBoost	0,773 (0,704 - 0,831)	0,772 (0,696 - 0,833)	0,754 (0,667 - 0,819)	0,739 (0,667 - 0,806)

Tablo 3. Değişen sansür oranlarına göre yöntemlerin AUPR, AUC, F_1 skoru, kappa katsayısı ve MKK değerlerine ilişkin tanımlayıcı istatistikler (Devam)

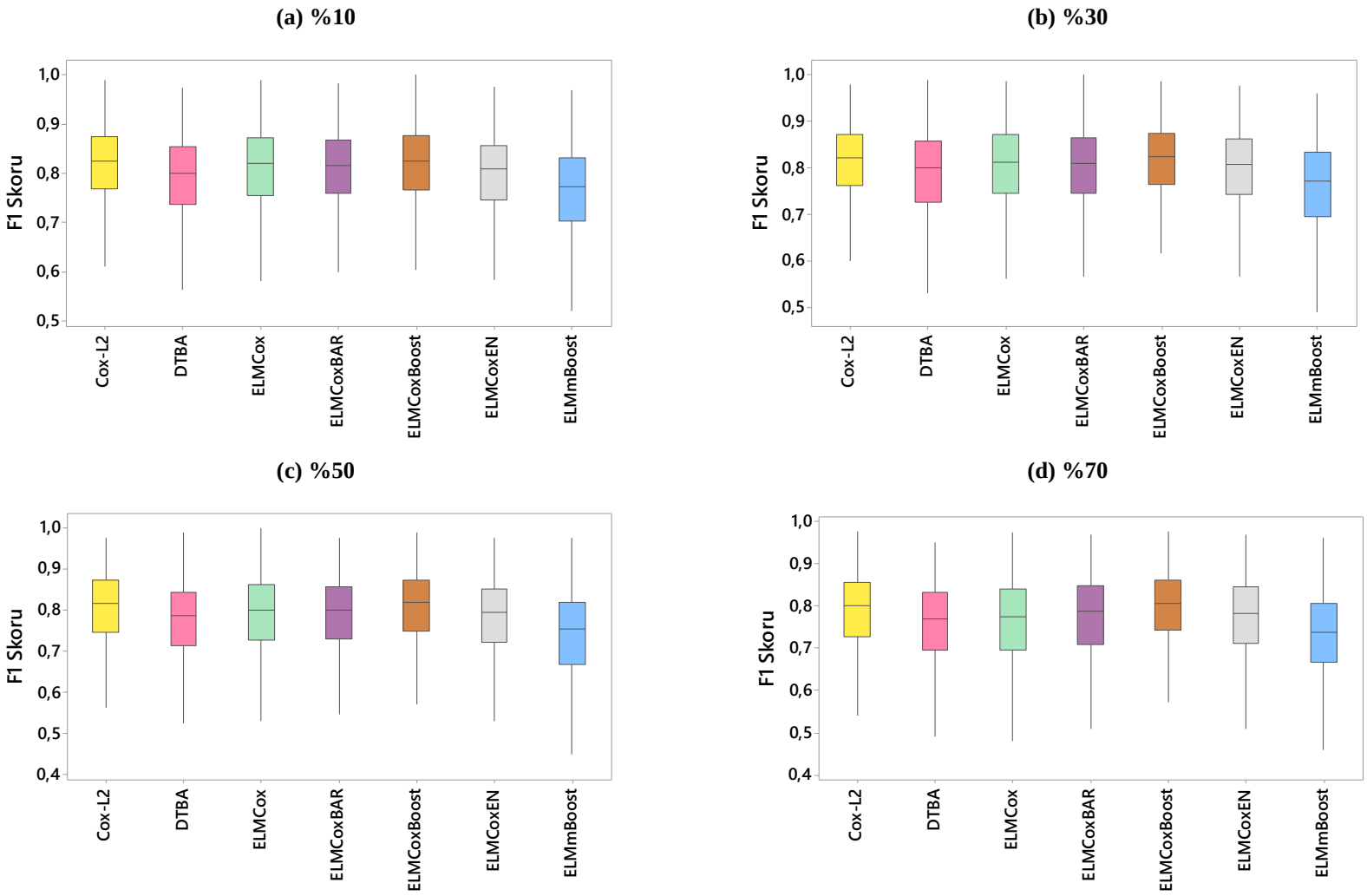
	Yöntemler	Sansür Oranı			
		%10	%30	%50	%70
Kappa	Cox-L_2	0,633 (0,531 - 0,729)	0,632 (0,525 - 0,718)	0,622 (0,503 - 0,723)	0,585 (0,481 - 0,690)
	DTBA	0,599 (0,491 - 0,681)	0,579 (0,478 - 0,675)	0,558 (0,457 - 0,661)	0,532 (0,418 - 0,629)
	ELMCox	0,628 (0,523 - 0,724)	0,615 (0,500 - 0,702)	0,590 (0,452 - 0,699)	0,529 (0,434 - 0,638)
	ELMCoxBAR	0,622 (0,514 - 0,701)	0,602 (0,503 - 0,700)	0,595 (0,476 - 0,697)	0,561 (0,458 - 0,661)
	ELMCoxBoost	0,634 (0,539 - 0,729)	0,633 (0,527 - 0,722)	0,626 (0,510 - 0,726)	0,600 (0,502 - 0,697)
	ELMCoxEN	0,602 (0,499 - 0,698)	0,598 (0,492 - 0,696)	0,585 (0,462 - 0,685)	0,558 (0,446 - 0,659)
	ELMmBoost	0,535 (0,430 - 0,634)	0,526 (0,406 - 0,633)	0,500 (0,383 - 0,611)	0,470 (0,352 - 0,571)
MKK	Cox-L_2	0,645 (0,545 - 0,733)	0,642 (0,535 - 0,724)	0,632 (0,523 - 0,731)	0,598 (0,500 - 0,694)
	DTBA	0,606 (0,504 - 0,686)	0,594 (0,497 - 0,690)	0,569 (0,473 - 0,667)	0,546 (0,443 - 0,636)
	ELMCox	0,635 (0,540 - 0,730)	0,628 (0,530 - 0,714)	0,600 (0,529 - 0,702)	0,540 (0,444 - 0,654)
	ELMCoxBAR	0,633 (0,534 - 0,713)	0,621 (0,525 - 0,706)	0,603 (0,495 - 0,702)	0,576 (0,476 - 0,667)
	ELMCoxBoost	0,643 (0,558 - 0,734)	0,642 (0,546 - 0,728)	0,634 (0,530 - 0,731)	0,609 (0,519 - 0,700)
	ELMCoxEN	0,613 (0,514 - 0,700)	0,605 (0,505 - 0,700)	0,597 (0,482 - 0,694)	0,571 (0,468 - 0,668)
	ELMmBoost	0,551 (0,455 - 0,644)	0,541 (0,432 - 0,649)	0,519 (0,407 - 0,624)	0,489 (0,380 - 0,589)



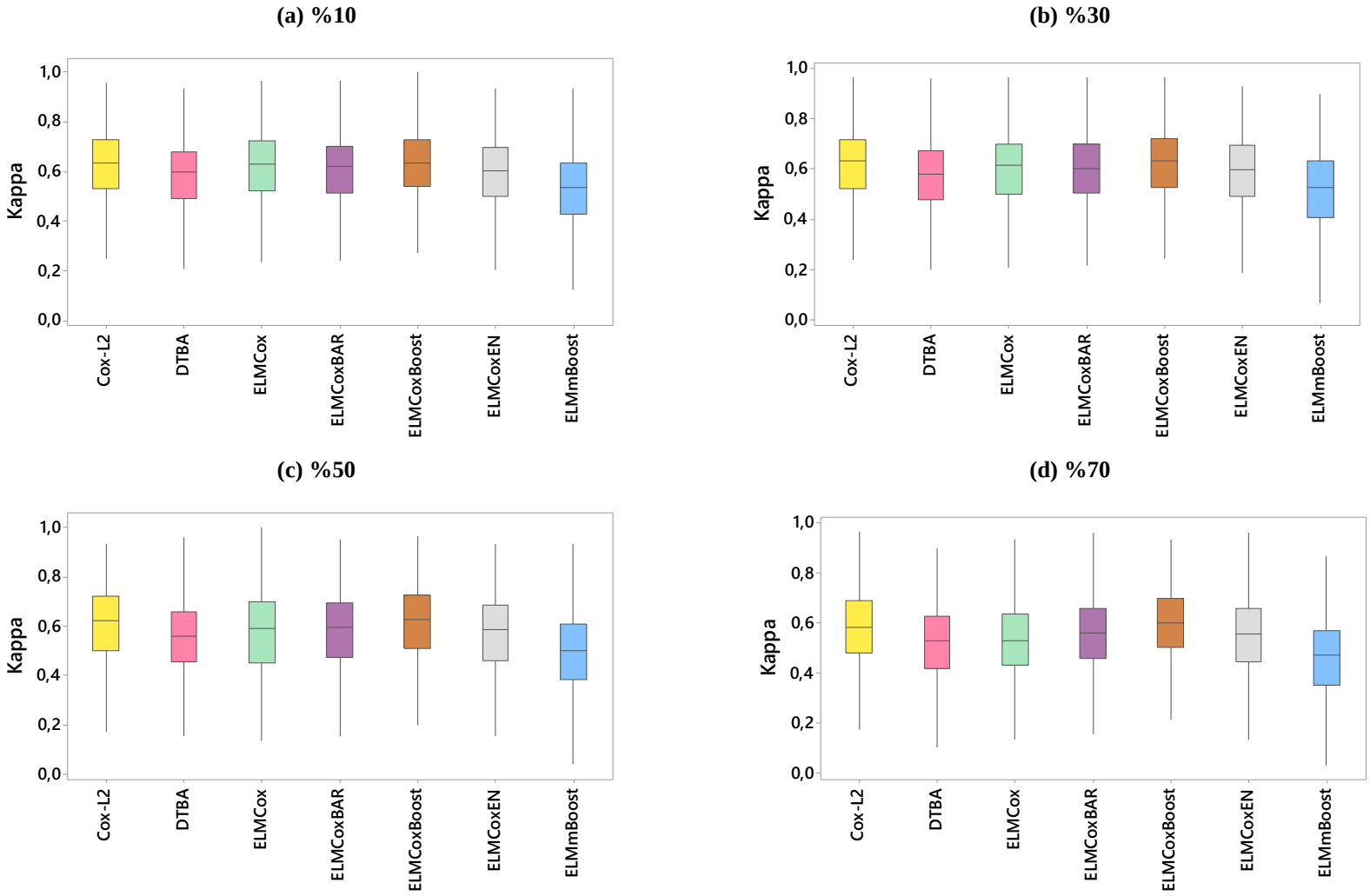
Şekil 11. Değişen sansür oranlarına göre sağkalım yöntemlerinin AUPR değerlerine ilişkin kutu grafikleri



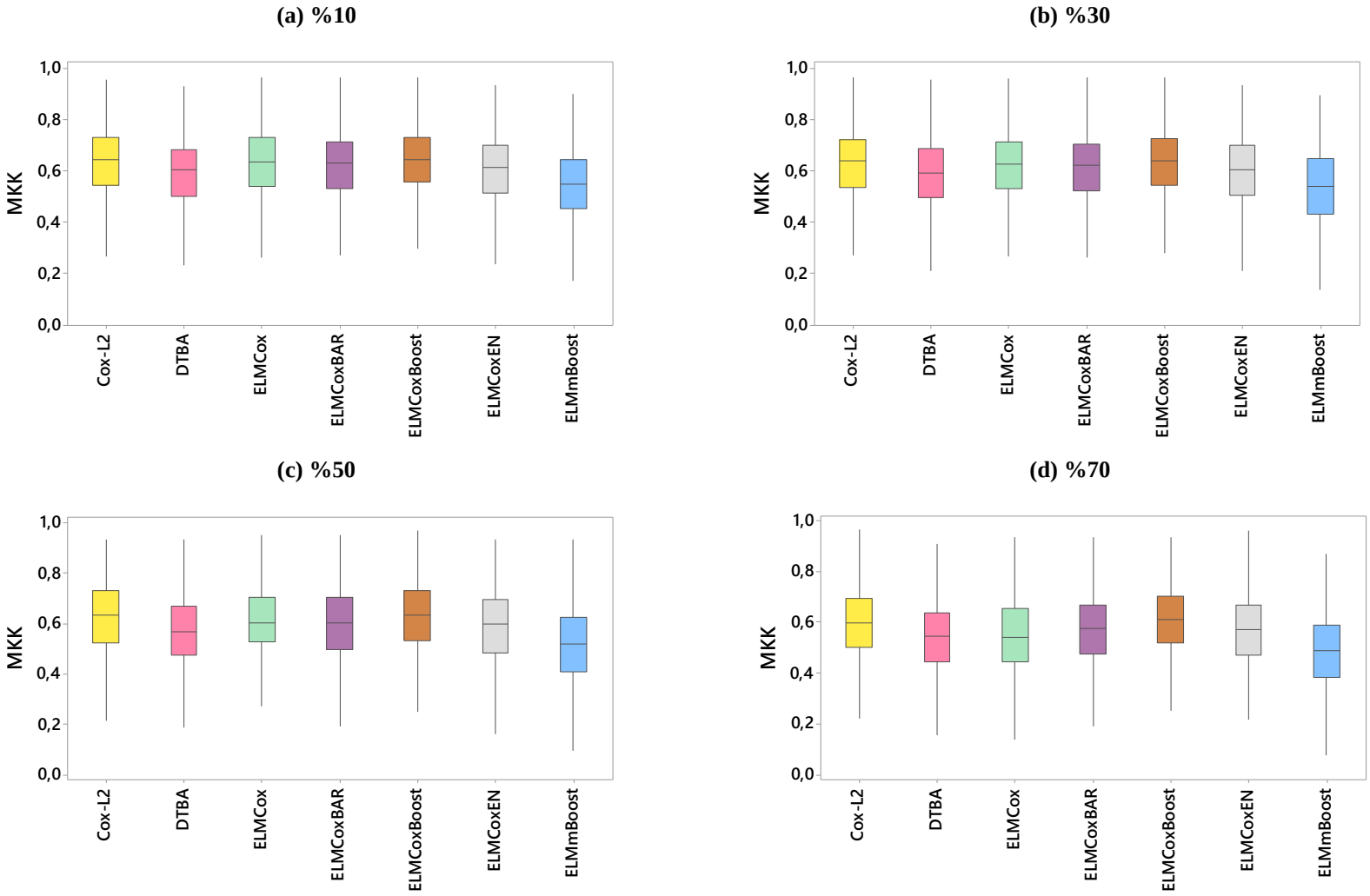
Şekil 12. Değişen sansür oranlarına göre sağkalım yöntemlerinin AUC değerlerine ilişkin kutu grafikleri



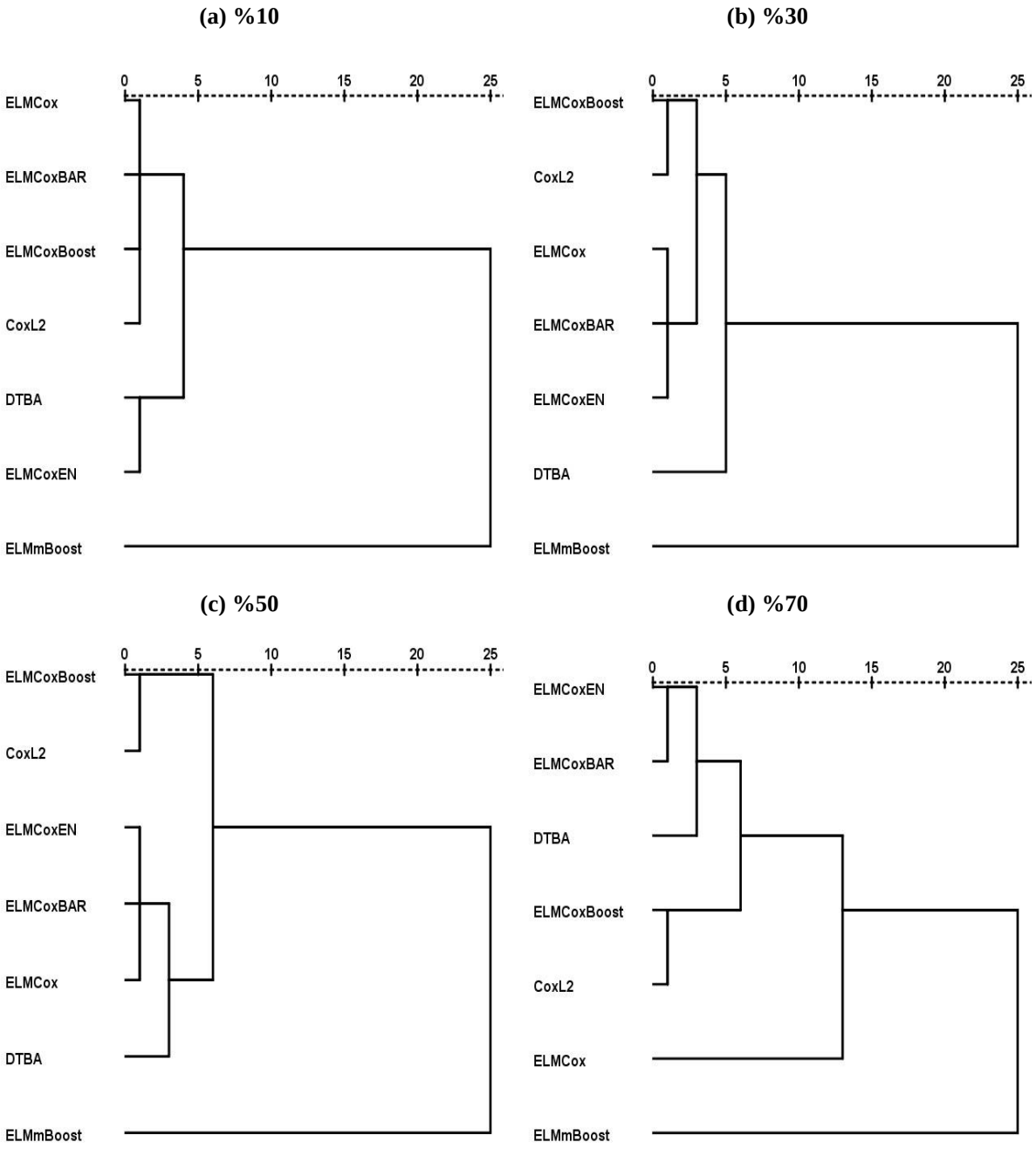
Şekil 13. Değişen sansür oranlarına göre sağkalım yöntemlerinin F_1 skoru değerlerine ilişkin kutu grafikleri



Şekil 14. Değişen sansür oranlarına göre sağkalım yöntemlerinin kappa değerlerine ilişkin kutu grafikleri



Şekil 15. Değişen sansür oranlarına göre sağkalım yöntemlerinin MKK değerlerine ilişkin kutu grafikleri



Şekil 16. Değişen sansür oranlarına göre modellerin kısa dönem sağkalım durumu tahmin performansları arasındaki ilişkileri gösteren dendrogram grafikleri

5. TARTIŞMA

Teknolojinin ilerlemesiyle birlikte kolaylaşan veri toplama ve saklama süreci, veri boyutunda artışa neden olmakla birlikte; bu verilerin analizi için özel yöntemlere gereksinim duyulmasını beraberinde getirmiştir. Sağlık alanındaki yüksek boyutlu veriler, hastalar hakkında çok fazla bilgi barındırdığı için; bu verilerin modellenmesi, kanser gibi çok önemli hastalıkların erken teşhisi, kısa ve uzun dönem risklerinin belirlenmesi açısından oldukça önemlidir. Yüksek boyutlu verilerin modellenmesi için, boyut yüksekliğinin neden olduğu çoklu doğrusal bağlantı, uzun işlem süresi, bulguların yorumlanma zorluğu gibi problemlerin ortadan kaldırılarak uygun olan analiz yönteminin belirlenmesi büyük önem arz etmektedir. Klasik sağkalım analizi yöntemleri p>n olduğunda doğrudan uygulanamadıklarından, yüksek boyutlu sağkalım verilerinin analizinde bu yöntemlerin kullanılabilmesi için veri boyutunun indirgenmesi gereklidir. Boyut indirgemeye dayalı analizler, verinin büyük bir kısmını temsil eden bileşenleri tahmin etmek ve sağkalım ile ilişkili bağımsız değişkenleri elde etmek amacıyla uygulanmaktadır. (Bair ve Tibshirani, 2004), n=100 birim ve p=5000 gen ifade değeri türeterek oluşturdukları yüksek boyutlu sağkalım verilerinin analizinde, DTBA yönteminin performansını denetimli ve denetimsiz yöntemlerle karşılaştırmışlardır. Elde ettikleri sonuçlara göre; yarı-denetimli bir yöntem olan DTBA yönteminin, denetimli ve denetimsiz olan yöntemlerden oldukça yüksek performans sergilediğini rapor etmişlerdir. 1000 döngü ile gerçekleştirilen çalışmamızda ise; n=200 ve p=1000 gen ifade değerinden oluşan yüksek boyutlu sağkalım verileri türetilmiştir. Türetilen verilerin DTBA, SELM ve Cox-L_2 yöntemleri ile analiz edilmesi sonucu, sansür oranı %70 olduğunda ELMCoxBoost ve Cox-L_2 yöntemlerinin sağkalım süresi tahminindeki performanslarının DTBA yöntemine yakın olduğu görülmüştür. Yöntemlerin kısa dönem sağkalım durumu tahminindeki performanslarına bakıldığında, sansür oranı %10 için ELMCoxEN; %50 için ELMCoxEN, ELMCoxBAR ve ELMCox; %70 için ELMCoxBAR ve ELMCoxEN yöntemlerinin DTBA yöntemine benzer sonuçlar verdikleri belirlenmiştir.

Yüksek boyutlu sağkalım verilerinde, sağkalım ile ilişkili değişkenler belirlenerek veri boyutu indirgenmesine rağmen değişken sayısı gözlem sayısından yine de fazla olabilir. Bu sebeple, yüksek boyutlu sağkalım verilerini boyut indirgemeden doğrudan analiz edebilen SELM yöntemleri önerilmiştir. SELM yöntemlerinin tercih edilmesinin sebeplerinden biri de

gerçek veri setlerini diğer yöntemlerden daha başarılı performansla analiz etmeleridir (Dhillon ve Singh, 2020; Sun ve diğerleri., 2018).

Yüksek boyutlu sağkalım verilerinin analizinde, uygun olan yöntemin belirlenmesini gerektiren bir diğer unsur ise sansür oranıdır. Araştırmalarda veri yapısına en uygun olan analiz yönteminin seçilmesi, analiz sürecinde büyük rol oynamaktadır. Sansür oranı, sağkalım verilerinin karakteristik bir özelliği olduğundan en başarılı analiz yöntemi de veriden veriye değişebilmektedir. (Wang ve Zhou, 2018); sansür oranı %27,1 ve %68,3 arasında değişen yüksek boyutlu altı gerçek sağkalım veri setinde ELMBJEN, ELMCoxEN, ELMCoxBoost ve ELMmBoost yöntemlerini RSFL, RSFM, RSFC, Cox-L_1 ve Cox-L_2 yöntemleri ile karşılaştırmıştır. Yaptıkları analiz sonucunda elde ettikleri C-indeks değerlerine göre, çalışmalarında kullandıkları veri setlerinin dört tanesinde ELMBJEN ve ELMCoxEN; bir tanesinde ELMCoxBoost ve ELMmBoost yöntemlerinin en iyi performans gösteren yöntemler olduğunu; bir tanesinde ise tüm yöntemlerin performanslarının birbirine çok yakın olduğunu rapor etmişlerdir. Buna ek olarak, yöntemlerin tüm veri setlerindeki C-indeks değerlerinin ranklarına göre yaptıkları karşılaştırmada ise SELM yöntemleri arasından ELMBJEN ve ELMCoxEN yöntemlerinin öne çıkan yöntemler olduğunu ifade etmişlerdir. Yüksek boyutlu sağkalım verilerinin analiz edilmesi için birbirine çok yakın sonuçlar veren yöntemler arasından hangisinin en uygun yöntem olduğunun tespit edilmesi süreci gerçek veri setlerinin analiz edilmesine karşı simülasyon ile daha kolay yönetilebilir bir durumdur. Bu çalışmada, değişen sansür oranına göre tüm performans değerlendirme kriterlerinden elde edilen sonuçlara aşamalı kümeleme analizi uygulandığında; ELMCoxBoost ve Cox-L_2 yöntemlerinin performanslarının birbirine çok yakın olduğu, dolayısıyla bu yöntemlerin birbirlerinin yerine kullanılabileceği gösterilmiştir.

SELM yöntemleri ve bu yöntemlere ilişkin parametreler esnek olduğundan, sağkalım verilerinin SELM yöntemleri ile analiz edilmesinde sansür oranının yanı sıra, kullanılan çekirdek fonksiyonunun türü ve parametre değerleri de sonuçları etkileyebilmektedir. Bu sebeple; yöntemleri doğrudan karşılaştırmadan önce, modellere ilişkin en uygun parametre belirlenmelidir. (Wang ve Li, 2019), geliştirdikleri ELMCoxBAR yöntemini farklı bağımsız değişken sayısı, hazard fonksiyonu şekil parametresi ve çekirdek fonksiyonuna göre türetilmiş veri setlerinin yanı sıra sansür oranı %32,09 ve %78,57 arasında değişen dokuz gerçek veri setinde uygulamışlardır. Elde ettikleri bulgulara göre, ELMCoxBAR modelinin RBF ile eğitilmesinin zaman alıcı olduğunu, veri yapısına uygun olmayan çekirdek fonksiyonu kullanıldığında dahi ELMCoxBAR modelinin sağlam performans gösterdiğini ve c parametre

değerinden çok etkilenmediğini belirterek modelin eğitimi aşamasında RBF yerine doğrusal çekirdek fonksiyonunun kullanılmasını önermişlerdir. Bu çalışmanın bulguları ile denemeler sonucu elde ettiğimiz bulgulara dayanarak, çalışmamızda ELMCoxBAR modelinin eğitilmesi aşamasında doğrusal çekirdek fonksiyonu ve c=0,5 parametresi kullanılmış; ELMCoxBAR modelinin performansının, artan sansür oranından olumsuz olarak etkilendiği belirlenmiştir.

SELM modelleri, yüksek boyutlu sağkalım verilerinin analizi için alternatif olarak önerilmelerinin yanı sıra, düşük boyutlu sağkalım verilerinin analizinde de kullanılabilmektedir. (Dhillon ve Singh, 2020) ise, meme kanseri hastalarına ilişkin patolojik görüntüleri içeren düşük boyutlu gerçek genomik sağkalım veri setini ELMBJ, ELMBJEN, ELMCox, ELMCoxEN, ELMCoxBoost ve ELMmBoost yöntemleriyle analiz etmiştir. Yaptıkları analiz sonunda, yöntemlerin meme kanseri hastalarının 5-yıllık sağkalım durumuna ilişkin tahmin performanslarını karşılaştırmıştır. Çalışmalarının sonuçlarına göre, tüm yöntemlerin birbirine yakın performans gösterdiğini; ancak, ELMmBoost yönteminin diğer yöntemlerden daha öne çıktığını ifade etmişlerdir. (Yang ve diğerleri., 2021), düşük boyutlu sağkalım verilerinin analizinde ELMCox, RSF ve Cox-L_1 yöntemlerini kullandıkları simülasyon çalışmalarında, sansür oranı arttıkça ELMCox modelinin daha kötü performans sergilediğini göstermişlerdir. Ayrıca; sansür oranı %25 iken ELMCox ve RSF modellerinin performanslarının neredeyse aynı olduğuna ve Cox-L_1 modelinin diğer iki yönteme kıyasla biraz daha kötü performans gösterdiğine dikkat çekmişlerdir.

Literatürde yer alan az sayıdaki çalışmanın sonuçlarına paralel olarak, bu çalışmada da sansür oranı daha az olduğunda SELM, DTBA ve Cox-L_2 yöntemlerinin daha iyi performans sergiledikleri belirlenmiştir. (Wang ve Zhou, 2018)'nun gerçek veri setleriyle yaptıkları çalışmalarında olduğu gibi, bu tez çalışmasında da ELMCoxBoost yöntemi en iyi performans gösteren yöntem olarak öne çıkmıştır; ancak, (Dhillon ve Singh, 2020)'in çalışmasında olduğu gibi yöntemlerin performansları arasında çok büyük farklılıklar olmadığı görülmüştür. Modellerin teorik yapısı incelendiğinde; yinelemesiz öğrenme yapısı ile tek öğrenme parametresinin gizli ve çıktı katmanları arasındaki ağırlıklar olması nedeniyle SELM modellerinin geleneksel algoritmalara göre çok daha hızlı öğrendiği bilinmektedir. Ayrıca, diğer modeller tarafından kullanılan gradyan iniş tabanlı optimizasyon yöntemleri modellerin genelleme yeteneğini azalttığı için, SELM modellerinin yinelemesiz öğrenme yapısı rastgele parametrelerle optimal bir çözüm sağlamaktadır. Bunun yanı sıra, gizli düğüm parametrelerinin ayarlanması gerekmediği gibi karmaşık bir parametre ayarlama işlemi de gerektirmediği için SELM modelleri diğer modellere göre daha tercih edilebilirdir.

6. SONUÇ VE ÖNERİLER

Sansür oranı %10, %30, %50 ve %70 olacak şekilde n=200 birim ve p=1000 gen ifade değeri için 1000 döngü ile gerçekleştirilen simülasyon çalışmamızda yüksek boyutlu sağkalım verileri türetildi. Türetilen sağkalım verilerinde Cox-L_2, DTBA, ELMCox, ELMCoxBAR, ELMCoxBoost, ELMCoxEN, ve ELMmBoost yöntemleri ile sağkalım süresi ve kısa dönem sağkalım durumu tahmini yapıldı. Yöntemlerin sağkalım süresi tahminine ilişkin performansları C-indeks ve IBS performans ölçütleri ile; kısa dönem sağkalım durumu tahminine ilişkin performansları ise duyarlılık, özgüllük, doğruluk oranı ile NTO, PTO, AUPR, AUC, F_1 skoru, kappa katsayısı ve MKK performans ölçütleri ile karşılaştırıldı. Birbirine yakın performans gösteren yöntemler aşamalı kümeleme analizi ile belirlendi.

Modellerin sağkalım süresi tahmin performanslarının sansür oranına göre değişiklik gösterdiği belirlendi. Bu değişimin sansür oranındaki değişim ile ters yönlü olduğu, sansür oranı arttıkça modellerin sağkalım süresi tahmin performanslarının düştüğü gözlendi. C-indeks ve IBS sonuçları ile uygulanan aşamalı kümeleme analizine göre sansür oranı %10 ve %30 olan yüksek boyutlu sağkalım verilerini ELMCoxBoost, Cox-L_2, ELMCox ve ELMCoxBAR modellerinin benzer performansla analiz ettiği gözlendi. %50 sansürlü gözlem oranına sahip veri setlerinin analiz bulgularına göre ise ELMCox, Cox-L_2 ile ELMCoxBoost modellerinin, bunun yanı sıra; ELMCoxEN ile ELMCoxBAR modellerinin birbirlerine yakın sonuçlar verdiği belirlendi. Sansür oranı %70 olduğunda ise ELMCox, ELMCoxEN ve ELMCoxBAR; ELMCoxBoost, Cox-L_2 ve DTBA yöntemlerinin birbirine yakın performans gösterdiği gözlendi.

Modellerin kısa dönem sağkalım durumu tahmin performanslarına ilişkin sonuçlar incelendiğinde ise performansların birbirine yakın olduğu tespit edildi. Duyarlılık, özgüllük, doğruluk oranı ile NTO, PTO, F_1 skoru, AUPR, AUC, kappa katsayısı ve MKK sonuçları ile uygulanan aşamalı kümeleme analizine göre sansür oranı %10 olan yüksek boyutlu sağkalım verilerini ELMCox, ELMCoxBAR, ELMCoxBoost ile Cox-L_2; DTBA ile ELMCoxEN modellerinin benzer performansla analiz ettiği gözlendi. %30 sansür oranına sahip sağkalım verilerinin analizinde ise ELMCoxBoost ile Cox-L_2; ELMCox, ELMCoxBAR ile ELMCoxEN modellerinin yakın sonuçlar verdiği belirlendi. Veri setindeki gözlemlerin yarısı sansürlü olduğunda ise ELMCoxBoost ile Cox-L_2; ELMCoxEN, ELMCoxBAR, ELMCox ile DTBA yöntemlerinin birbirine benzer performans gösterdiği gözlendi. Sansür oranı %70 olduğunda

ise, ELMCoxBoost ile Cox-L_2; DTBA, ELMCoxBAR ile ELMCoxEN yöntemlerinin birbirine yakın performans gösterdikleri bulundu.

Sonuç olarak; çalışmamızda yüksek boyutlu sağkalım verilerinin analizinde kullanılan sağkalım analizi yöntemleri, sansür oranındaki artıştan olumsuz olarak etkilenmektedir. Modellerin kısmi logaritmik olabilirlik fonksiyon yapısı nedeniyle, sansür oranındaki artış, tahminlerin hatasını artırarak modelin tahmin performansını düşürmektedir. Modeller birbirine yakın performans gösterse de tüm sansür senaryolarındaki ortak çıkarım ELMCoxBoost ve Cox-L_2 yöntemlerinin birbirleri yerine kullanılabilir olduğudur. Cezalı modellerde ceza parametresinin veri yapısına en uygun değerinin belirlenmesi gerektiğinden uygulanması daha pratik olan ELMCoxBoost yöntemi Cox-L_2 yöntemi yerine tercih edilebilir. Veri setinde değişken sayısının fazla olması, parametre optimizasyon süreci ve simülasyon döngü sayısından dolayı çok zaman alan analiz süreci, yüksek boyutlu sağkalım verileri için en uygun yöntemin belirlenmesinde karşılaşılan önemli bir problemdir. Buna karşılık, farklı veri yapılarının analizinden elde edilecek sonuçlar değişkenlik gösterebileceğinden, farklı veri setlerinde aynı amaçla kullanılan bu yöntemlerden en uygun olanı simülasyon çalışmalarıyla denetlenerek belirlenmelidir.

Bu çalışma, değişen sansür oranlarına göre türetilmiş yüksek boyutlu sağkalım verilerinde bu yöntemlerin performanslarını ve birbirlerine benzerliklerini belirlemek açısından literatüre bilimsel katkıda bulunmaktadır. DTBA gibi boyut indirgeme yöntemlerinin ve cezalı modellerin yerine yüksek boyutlu sağkalım verilerini özellikle doğrudan analiz edebilen SELM modellerinin kullanılabilirliği bu çalışmayla ortaya konmaktadır. Çalışmamızı özgün kılan başka bir durum ise; DTBA, Cox-L_2 ve SELM yöntemleri ile hem sağkalım süresi hem de kısa dönem sağkalım durumu tahmini yapılan literatürde başka bir simülasyon çalışmasının olmamasıdır.

KAYNAKLAR

Aktürk Hayat, E., Türe, M., & Şenol, Ş. (2016). An Alternative Dimension Reduction Approach to Supervised Principal Components Analysis in High Dimensional Survival Data. *Turkiye Klinikleri Journal of Biostatistics, 8*(1), 21-29.

Altman, D. G., & Bland, J. M. (1994). Diagnostic tests. 1: Sensitivity and specificity. *BMJ: British Medical Journal, 308*(6943), 1552.

Altman, N., & Krzywinski, M. (2017). Ensemble methods: bagging and random forests. *Nature Methods, 14*(10), 933-935.

Bair, E., Hastie, T., Paul, D., & Tibshirani, R. (2006). Prediction by supervised principal components. *Journal of the American Statistical Association, 101*(473), 119-137.

Bair, E., & Tibshirani, R. (2004). Semi-supervised methods to predict patient survival from gene expression data. *PLoS Biol, 2*(4), e108.

Binder, H., Allignol, A., Schumacher, M., & Beyersmann, J. (2009). Boosting for high-dimensional time-to-event data with competing risks. *Bioinformatics, 25*(7), 890-896.

Binder, H., Benner, A., Bullinger, L., & Schumacher, M. (2013). Tailoring sparse multivariable regression techniques for prognostic single-nucleotide polymorphism signatures. *Statistics in medicine, 32*(10), 1778-1791.

Binder, H., & Schumacher, M. (2009). Incorporating pathway information into boosting estimation of high-dimensional risk prediction models. *BMC bioinformatics, 10*(1), 1-11.

Breiman, L. (2001). Random forests. *Machine learning, 45*(1), 5-32.

Breiman, L. (2004). Consistency for a simple model of random forests.

Bühlmann, P., & Hothorn, T. (2007). Boosting algorithms: Regularization, prediction and model fitting. *Statistical science, 22*(4), 477-505.

Chen, X., Wang, L., Smith, J. D., & Zhang, B. (2008). Supervised principal component analysis for gene set enrichment of microarray data with continuous or survival outcomes. *Bioinformatics, 24*(21), 2474-2481.

Chicco, D., & Jurman, G. (2020). The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. *BMC genomics, 21*(1), 1-13.

Chicco, D., Tötsch, N., & Jurman, G. (2021). The Matthews correlation coefficient (MCC) is more reliable than balanced accuracy, bookmaker informedness, and markedness in two-class confusion matrix evaluation. *BioData mining, 14*(1), 1-22.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and psychological measurement, 20*(1), 37-46.

Davis, J., & Goadrich, M. (2006). *The relationship between Precision-Recall and ROC curves.* Paper presented at the Proceedings of the 23rd international conference on Machine learning.

De Bin, R. (2016). Boosting in Cox regression: a comparison between the likelihood-based and the model-based approaches with focus on the R-packages CoxBoost and mboost. *Computational Statistics, 31*(2), 513-531.

Delen, D., Walker, G., & Kadam, A. (2005). Predicting breast cancer survivability: a comparison of three data mining methods. *Artificial intelligence in medicine, 34*(2), 113-127.

Dhillon, A., & Singh, A. (2020). eBreCaP: extreme learning-based model for breast cancer survival prediction. *IET Systems Biology, 14*(3), 160-169.

Fawcett, T. (2006). An introduction to ROC analysis. *Pattern recognition letters, 27*(8), 861-874.

Fill, J. A., & Fishkind, D. E. (2000). The Moore--Penrose Generalized Inverse for Sums of Matrices. *SIAM Journal on Matrix Analysis and Applications, 21*(2), 629-635.

Fletcher, R. (2013). *Practical methods of optimization*: John Wiley & Sons.

Gitto, S., Magistri, P., Marzi, L., Mannelli, N., De Maria, N., Mega, A., . . . Villa, E. (2022). Predictors of solid extra-hepatic non-skin cancer in liver transplant recipients and analysis of survival: a long-term follow-up study. *Annals of Hepatology*, 100683.

Graf, E., Schmoor, C., Sauerbrei, W., & Schumacher, M. (1999). Assessment and comparison of prognostic classification schemes for survival data. *Statistics in medicine, 18*(17-18), 2529-2545.

Hanley, J. A., & McNeil, B. J. (1982). The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology, 143*(1), 29-36.

Harrell, F. E., Califf, R. M., Pryor, D. B., Lee, K. L., & Rosati, R. A. (1982). Evaluating the yield of medical tests. *Jama, 247*(18), 2543-2546.

Harrell Jr, F. E., Lee, K. L., & Mark, D. B. (1996). Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. *Statistics in medicine, 15*(4), 361-387.

Hoerl, A. E., & Kennard, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics, 12*(1), 55-67.

Hothorn, T., Bühlmann, P., Kneib, T., Schmid, M., & Hofner, B. (2010). Model-based boosting 2.0. *The Journal of Machine Learning Research, 11*, 2109-2113.

Huang, G.-B. (2014). An insight into extreme learning machines: random neurons, random features and kernels. *Cognitive Computation, 6*(3), 376-390.

Huang, G.-B., Chen, L., & Siew, C. K. (2006). Universal approximation using incremental constructive feedforward networks with random hidden nodes. *IEEE Trans. Neural Networks, 17*(4), 879-892.

Huang, G.-B., Zhou, H., Ding, X., & Zhang, R. (2011). Extreme learning machine for regression and multiclass classification. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics), 42*(2), 513-529.

Huang, G.-B., Zhu, Q.-Y., & Siew, C.-K. (2004). *Extreme learning machine: a new learning scheme of feedforward neural networks.* Paper presented at the 2004 IEEE international joint conference on neural networks (IEEE Cat. No. 04CH37541).

Huang, G.-B., Zhu, Q.-Y., & Siew, C.-K. (2006). Extreme learning machine: theory and applications. *Neurocomputing, 70*(1-3), 489-501.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning* (Vol. 112): Springer.

Kasun, L. L. C., Yang, Y., Huang, G.-B., & Zhang, Z. (2016). Dimension reduction with extreme learning machine. *IEEE transactions on Image Processing, 25*(8), 3906-3918.

Kawaguchi, E. S., Suchard, M. A., Liu, Z., & Li, G. (2017). Scalable Sparse Cox's Regression for Large-Scale Survival Data via Broken Adaptive Ridge. *arXiv preprint arXiv:1712.00561*.

Keilwagen, J., Grosse, I., & Grau, J. (2014). Area under precision-recall curves for weighted and unweighted data. *PloS one, 9*(3), e92209.

Kronek, L.-P., & Reddy, A. (2008). Logical analysis of survival data: prognostic survival models by detecting high-degree interactions in right-censored data. *Bioinformatics, 24*(16), i248-i253.

Krzanowski, W. J., & Hand, D. J. (2009). *ROC curves for continuous data*: Chapman and Hall/CRC.

Lai, Y., Hayashida, M., & Akutsu, T. (2013). Survival analysis by penalized regression and matrix factorization. *The Scientific World Journal, 2013*.

Li, J., Zhou, Z., Dong, J., Fu, Y., Li, Y., Luan, Z., & Peng, X. (2021). Predicting breast cancer 5-year survival using machine learning: a systematic review. *PloS one, 16*(4), e0250370.

Lou, S.-J., Hou, M.-F., Chang, H.-T., Lee, H.-H., Chiu, C.-C., Yeh, S.-C. J., & Shi, H.-Y. (2022). Breast Cancer Surgery 10-Year Survival Prediction by Machine Learning: A Large Prospective Cohort Study. *Biology, 11*(1), 47.

Lu, J., Huang, J., & Lu, F. (2019). Distributed kernel extreme learning machines for aircraft engine failure diagnostics. *Applied Sciences, 9*(8), 1707.

McNeil, B. J., & Adelstein, S. J. (1976). Determining the value of diagnostic and screening tests. *Journal of Nuclear Medicine, 17*(6), 439-448.

Perperoglou, A. (2014). Cox models with dynamic ridge penalties on time-varying effects of the covariates. *Statistics in medicine, 33*(1), 170-180.

Powers, D. M. (2020). Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation. *arXiv preprint arXiv:2010.16061*.

Ridgeway, G. (2020). Generalized Boosted Models: A guide to the gbm package. *Update, 1*, 1.

Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). Learning representations by back-propagating errors. *nature, 323*(6088), 533-536.

Schmidt, M., Böhm, D., Von Törne, C., Steiner, E., Puhl, A., Pilch, H., . . . Gehrmann, M. (2008). The humoral immune system has a key prognostic impact in node-negative breast cancer. *Cancer research, 68*(13), 5405-5413.

Sun, D., Li, A., Tang, B., & Wang, M. (2018). Integrating genomic data and pathological images to effectively predict breast cancer clinical outcome. *Computer methods and programs in biomedicine, 161*, 45-53.

Türe, M., & Kurt Ömürlü, İ. (2018). Development of a New Supervised Principal Component Analysis Based on Artificial Neural Networks in Gene Expression Data. *Osmangazi Tıp Dergisi, 40*(1), 20-27.

Van Houwelingen, H. C., Bruinsma, T., Hart, A. A., van't Veer, L. J., & Wessels, L. F. (2006). Cross-validated Cox regression on microarray gene expression data. *Statistics in medicine, 25*(18), 3201-3216.

Verweij, P. J., & Van Houwelingen, H. C. (1994). Penalized likelihood in Cox regression. *Statistics in medicine, 13*(23-24), 2427-2436.

Wang, H., & Li, G. (2019). Extreme learning machine Cox model for high-dimensional survival analysis. *Statistics in medicine, 38*(12), 2139-2156.

Wang, H., Wang, J., & Zhou, L. (2018). A survival ensemble of extreme learning machine. *Applied Intelligence, 48*(7), 1846-1858.

Wang, H., & Zhou, L. (2018). SurvELM: an R package for high dimensional survival analysis with extreme learning machine. *Knowledge-Based Systems, 160*, 28-33.

Warrens, M. J. (2015). Five ways to look at Cohen's kappa. *Journal of Psychology & Psychotherapy, 5*(4), 1.

Witten, D. M., & Tibshirani, R. (2010). Survival analysis with high-dimensional covariates. *Statistical methods in medical research, 19*(1), 29-51.

Yang, H., Tian, J., Meng, B., Wang, K., Zheng, C., Liu, Y., . . . Zhang, Y. (2021). Application of Extreme Learning Machine in the Survival Analysis of Chronic Heart Failure Patients With High Percentage of Censored Survival Time. *Frontiers in cardiovascular medicine, 8*.

Yu, Y., Xu, S., Zhao, E., Dong, Y., Chen, J., Rao, B., . . . Qiu, F. (2022). Identification of a 10-pseudogenes signature as a novel prognosis biomarker for ovarian cancer. *Biocell, 46*(4), 999.

T.C.

AYDIN ADNAN MENDERES ÜNİVERSİTESİ

SAĞLIK BİLİMLERİ ENSTİTÜSÜ

BİLİMSEL ETİK BEYANI

"Yüksek Boyutlu Sağkalım Verilerinin Denetimli Temel Bileşenler, Cezalı Cox Regresyon ve Aşırı Öğrenme Makineleri Yöntemleri ile Karşılaştırmalı Analizi" başlıklı Doktora tezimdeki bütün bilgileri etik davranış ve akademik kurallar çerçevesinde elde ettiğimi, tez yazım kurallarına uygun olarak hazırlanan bu çalışmada, bana ait olmayan her türlü ifade ve bilginin kaynağına eksiz atıf yaptığımı bildiririm. İfade ettiklerimin aksi ortaya çıktığında ise her türlü yasal sonucu kabul ettiğimi beyan ederim.

Fulden CANTAŞ TÜRKİŞ

08/06/2022

ÖZGEÇMİŞ

Soyadı, Adı : CANTAŞ TÜRKİŞ, Fulden

Uyruk : T.C.

Doğum yeri ve tarihi : Karşıyaka, İzmir / 19.04.1989

E-posta : fulden.cantas@adu.edu.tr

Yabancı dil : İngilizce

EĞİTİM

Derece	Kurum	Mezuniyet tarihi
Lisans	Manisa Celal Bayar Üniversitesi	2011
Yüksek Lisans	Aydın Adnan Menderes Üniversitesi	2016
Doktora	Aydın Adnan Menderes Üniversitesi	2022

İŞ DENEYİMİ

Yıl	Yer/Kurum	Ünvan
2011-2014	İzmir Yüksek Teknoloji Enstitüsü	Araştırma Görevlisi
2014-2022	Aydın Adnan Menderes Üniversitesi	Araştırma Görevlisi

AKADEMİK YAYINLAR

1. MAKALELER

1. Ünsal, A. İ. A., Junıor, F. J. L., **Cantaş, F.,** Kurt Ömürlü, İ., Ünal, A., & Demirci, B. (2021). Awareness, Attitudes and Behaviours of Medicine, Dentistry, Nursing and Midwifery Students on Eye Health: A Cross-sectional Study. Meandros Medical And Dental Journal, 22(1), 93-104.
2. Çeri, N., G., Şahmelikoğlu, A., G., **Cantaş, F.**, Sakallı, G. (2021). The anatomic study of extracranial structures related to tuberculum pharyngeum. European Journal of Anatomy, 25(2), 241-246
3. Bekmez, S., Cakmak, H., Kocaturk, T., **Cantas, F.,** & Dundar, S. (2021). Biomechanical properties of the cornea following intravitreal ranibizumab injection. Graefe's Archive for Clinical and Experimental Ophthalmology, 259(3), 691-696.
4. Türe, M., Öztürk, H., Kurt Ömürlü, İ., **Cantaş, F.** (2020). Comparison of Logistic Regression, K-Nearest Neighbor, Support Vector Machine, Multilayer Perceptron, and Bagging Classification and Regression Trees Using Wavelet-Based Features in Detection of Epileptic Signals. International Journal of Imaging and Robotics, 20 (3), 14-24
5. Omurlu, I. K., **Cantas, F.**, Ture, M., & Ozturk, H. (2020). An empirical study on performances of multilayer perceptron, logistic regression, ANFIS, KNN and bagging CART. Journal of Statistics and Management Systems, 23(4), 827-841.
6. Örs, B. S., **Cantas, F.**, Gungor, E. O., & Simsek, D. (2019). Assessment and comparison of visual skills among athletes. Spor ve Performans Araştırmaları Dergisi, 10(3), 231-241.
7. Kısa Makale, Yavaşoğlu İrfan, Gencer Nadide, **Cantaş Fulden**, Döger Füruzan, Bolaman Ali Zahit (2019). Can Histochemical c-mpl Positivity in Bone Marrow be a Predictor for Splenectomy in Immune Thrombocytopenia?. International Blood Research Reviews, 9(4).
8. **Cantaş, F.**, Kurt Ömürlü İ., & Türe, M. (2018). Comparison of hierarchical and non-hierarchical fuzzy models with simulation and an application on hypertension data set. Meandros Medical and Dental Journal, 19(2), 138.

2. PROJELER

1. Türe, M., Kurt Ömürlü, İ., Öztürk, H., **Cantaş, F.** (2019). Zaman serisi analizinde CART, CTREE, CART BAGGING, CTREE BAGGING, RF, SVR ve ARIMA yöntemlerinin performanslarının simülasyonla karşılaştırılması.

2. Bek, V.S., Çaman M. B., Özdemir, İ., Daloğlu, E., **Cantaş, F**., Kutlu, G. (Devam ediyor). Vagal sinir stimülasyonu uygulanan hastalarda stimülasyonun kapatılmasının serebral hemodinami üzerine etkileri.

3. Bek, V.S., Çaman M. B., Mercan, T., Daloğlu, E., Özdemir, İ., **Cantaş, F.,** Kutlu, G. (Devam ediyor). Vagal sinir stimülasyonunun elektroensefalografi kaydı üzerine etkileri.

4. Bek, V.S., Özdemir, İ., Çaman M. B., Daloğlu, E., **Cantaş, F.,** Kutlu, G. (Devam ediyor). Vagal sinir stimülasyonunun serebral kan akımı üzerine etkileri.

5. Bek, V.S., Daloğlu, E., Çaman M. B., Özdemir, İ., **Cantaş, F.,** Kutlu, G. (Devam ediyor). Vagal sinir stimülasyonu uygulanan hastalarda uykunun polisomnografi ile değerlendirilmesi.

6. Bek, V.S., Daloğlu, E., Özdemir, İ., Çaman M. B., **Cantaş, F.,** Kutlu, G. (Devam ediyor). Vagal sinir stimülasyonunun somatosensoriyel uyarılmış potansiyel (SSEP) üzerine etkileri.

3. BİLDİRİLER

A) Uluslararası Kongrelerde Sunulan Bildiriler

1. **Cantaş Fulden**, Kurt Ömürlü İmran, Türe Mevlüt (2021). An Empirical Study on High Dimensional, Censored Survival Data Analysis with Supervised Principal Components, Extreme Learning Machine-Based and Penalized Cox Models. II. International Applied Statistics Conference (Özet Bildiri/Sözlü Sunum)

2. **Cantaş Fulden**, Kurt Ömürlü İmran, Türe Mevlüt, Öztürk Hakan (2021). A Survival Simulation Study on Usage of Train-Test Splitting, Cross Validation and Bagging Methods on Training Weibull, Exponential, Log-normal, Log-logistic, Quantile and Cox Regression Models. II. International Applied Statistics Conference (Özet Bildiri/Sözlü Sunum)

3. **Cantaş Fulden**, Kurt Ömürlü İmran, Türe Mevlüt (2021). Survival Prediction with Supervised Principal Components Analysis, Penalized Cox Models and Extreme Learning Machine Based Penalized Cox Models on Simulated High Dimensional Data. XXII. Ulusal ve V. Uluslararası Biyoistatistik Kongresi (Özet Bildiri/Sözlü Sunum)

4. Oturakçı İbrahim, Deper İpek, Örs Berfin Serdil, **Cantaş Fulden**, Onarıcı Güngör Elvin, Şimşek Deniz (2018). Sporcularda görsel motor kontrolün değerlendirilmesi. 9th International Biomechanics Congress (Tam Metin Bildiri/Sözlü Sunum)

5. Ayaz Ece, Yıldız Gülnur, Şahbaz Selin, Örs Berfin Serdil, **Cantaş Fulden**, Onarıcı Güngör Elvin, Şimşek Deniz (2018). Visual control of basketball free throw shooting: A pilot study. 9th International Biomechanics Congress (Tam Metin Bildiri/Sözlü Sunum)

6. Kalender Hülya, Uzuner Kubilay, Bayram İsmail, **Cantaş Fulden**, Şimşek Deniz (2018). Comparison of Foot Posture, Ankle Joint Forces And Plantar Pressure Values of the Athletes From Different Sports. 9th International Biomechanics Congress (Özet Bildiri/Sözlü Sunum)

7. Ayaz Ece, Şahbaz Selin, Örs Berfin Serdil, **Cantaş Fulden**, Onarıcı Güngör Elvin, Şimşek Deniz (2018). Basketbolda serbest atışın görsel kontrolü:Pilot çalışma. 9th International Biomechanics Congress (Tam Metin Bildiri/Sözlü Sunum)

8. Oturakçı İbrahim, Deper İpek, Örs Berfin Serdil, **Cantaş Fulden**, Onarıcı Güngör Elvin, Şimşek Deniz (2018). The role of vision on perceptual-motor skill performance. 9th International Biomechanics Congress (Tam Metin Bildiri/Sözlü Sunum)

9. Oturakçı İbrahim, Deper İpek, Örs Berfin Serdil, **Cantaş Fulden**, Onarıcı Güngör Elvin, Şimşek Deniz (2018). Assessment of Visual-Motor Control in Athletes. 9th International Biomechanics Congress (Tam Metin Bildiri/Sözlü Sunum)

10. Örs Berfin Serdil, Şimşek Deniz, **Cantaş Fulden** (2018). Farklı kategorilerde yarışan ritmik cimnastikçilerin patlayıcı kuvvetlerinin ve ön split esnekliklerinin belirlenmesi. International Congress on Recreation and Sport Management (Özet Bildiri/Sözlü Sunum)

11. Kurt Ömürlü İmran, **Cantaş Fulden**, Türe Mevlüt, Öztürk Hakan (2018). An empirical study on classification performances of artificial neural networks, logistic regression, ANFIS, k-nearest neighbor algorithm and bagging CART. IRSYSC 2018 (Özet Bildiri/Poster)

12. Öztürk Hakan, Türe Mevlüt, Kurt Ömürlü İmran, **Cantaş Fulden** (2018). Comparison of wavelet subbands for epileptic seizure detection using EEG signals. IRSYSC 2018 (Özet Bildiri/Sözlü Sunum)

13. Türe Mevlüt, Kurt Ömürlü İmran, Öztürk Hakan, **Cantaş Fulden** (2018). Comparison of CART, CART with bagging and random forests on seasonal time series forecasting using simulation. IRSYCS 2018 (Özet Bildiri/Sözlü Sunum)

14. Alkan Aslı Gül, Aydın Elif, **Cantaş Fulden**, Kurt Ömürlü İmran, Şendur Ömer Faruk (2018). Diz Effüzyonunun Denge Parametreleri Üzerine Etkisinin İncelenmesi. Uluslararası katılımlı Türk Romatoloji Kongresi (Özet Bildiri/Sözlü Sunum)

15. Alkan Aslı Gül, Aydın Elif, **Cantaş Fulden**, Kurt Ömürlü İmran, Şendur Ömer Faruk (2018). Diz Effüzyonunun Düşme Riski Üzerine Etkisi. Uluslararası katılımlı Türk Romatoloji kongresi (Özet Bildiri/Poster)

16. Yavaşoğlu İrfan, Gencer Nadide, **Cantaş Fulden**, Döger Füruzan, Bolaman Ali Zahit (2017). Can Histochemical C-Mpl Positivıiy in Bone Marrow Be A Predictor For Splenectomy

in Immune Thrombocytopenia?. 22nd Congress of European Hematology Association (Özet Bildiri/Poster)

17. Kurt Ömürlü İmran, **Cantaş Fulden**, Türe Mevlüt, Öztürk Hakan (2017). Comparison of Classification Performances of Artificial Neural Networks, Adaptive Neuro-Fuzzy Inference System and Logistic Regression. 3rd International Researchers, Statisticians and Young Statisticians (Özet Bildiri/Sözlü Sunum)

18. Türe Mevlüt, Öztürk Hakan, Kurt Ömürlü İmran, Kıylıoğlu Nefati, **Cantaş Fulden** (2017). The Comparison of Wavelet-based Features and Classification Methods in EEG Signals. 3rd Internatıonal Researchers,Statıstıcıans and Young Statıstıcıans Congress (Özet Bildiri/Sözlü Sunum)

19. **Cantaş Fulden**, Kurt Ömürlü İmran, Türe Mevlüt (2016). Comparison of Classification Performances of Hierarchical and Non-Hierarchical Fuzzy Models. XVIII. Ulusal ve I. Uluslararası Biyoistatistik Kongresi (Özet Bildiri/Sözlü Sunum)

B) Ulusal Kongrelerde Sunulan Bildiriler

1. Çeri Nazlı Gülriz, Polat Ayşe Gizem, **Cantaş Fulden** (2017). Anatomical characteristics of tuberculum pharyngeum and its gender-related change. 18. Ulusal Anatomi Kongresi, 11(2), 141-142. (Özet Bildiri/Poster)

2. Çakmak Harun, Bekmez Sinan, Kocatürk Tolga, Dündar Sema, **Cantaş Fulden** (2015). İntravitreal enjeksiyon yapılan hastalarda kornealbiyomekanik özellikler. Türk Oftalmoloji Derneği 49. Ulusal Kongresi (Özet Bildiri/Poster)