Ayşegül YABACI

T. C.
ULUDAĞ ÜNİVERSİTESİ
SAĞLIK BİLİMLERİ ENSTİTÜSÜ
BİYOİSTATİSTİK
ANABİLİMDALI

SAĞKALIM VERİLERİNDE KULLANILAN AĞAÇ TABANLI YÖNTEMLERİN KARŞILAŞTIRILMASI

Ayşegül YABACI

(YÜKSEK LİSANS TEZİ)

BURSA-2017

BİYOİSTATİSTİK ANABİLİM DALI YÜKSEK LİSANS TEZİ

2017

T.C.
ULUDAĞ ÜNİVERSİTESİ
SAĞLIK BİLİMLERİ ENSTİTÜSÜ
BİYOİSTATİSTİK ANABİLİM DALI

SAĞKALIM VERİLERİNDE KULLANILAN AĞAÇ TABANLI YÖNTEMLERİN KARŞILAŞTIRILMASI

Ayşegül YABACI

(YÜKSEK LİSANS TEZİ)

DANIŞMAN:
Doç. Dr. Deniz SIĞIRLI

BURSA-2017

T.C.
ULUDAĞ ÜNİVERSİTESİ
SAĞLIK BİLİMLERİ ENSTİTÜSÜ

ETİK BEYANI

Yüksek Lisans tezi olarak sunduğum;
"Sağkalım Verilerinde Kullanılan Ağaç Tabanlı Yöntemlerin Karşılaştırılması" adlı çalışmanın, proje safhasından sonuçlanmasına kadar geçen bütün süreçlerde bilimsel etik kurallarına uygun bir şekilde hazırlandığını ve yararlandığım eserlerin kaynaklar bölümünde gösterilenlerden oluştuğunu belirtir ve beyan ederim.

Ayşegül YABACI

KABUL ONAY

SAĞLIK BİLİMLERİ ENSTİTÜSÜ MÜDÜRLÜĞÜ'NE

Uludağ Üniversitesi Tıp Fakültesi Biyoistatistik Anabilim Dalı Yüksek Lisans öğrencisi Ayşegül YABACI tarafından hazırlanan "Sağkalım Verilerinde Kullanılan Ağaç Tabanlı Yöntemlerin Karşılaştırılması" konulu Yüksek Lisans tezi 19/01/2017 günü, ...13.....-..15... saatleri arasında yapılan tez savunma sınavında jüri tarafından oy birliği/oy çokluğu ile kabul edilmiştir.

	Adı-Soyadı	**İmza**
Tez Danışmanı	**Doç.Dr.Deniz SIĞIRLI**	
Üye	**Prof. Dr. Berna YAZICI**	
Üye	**Doç.Dr.Güven ÖZKAYA**	

Bu tez Enstitü Yönetim Kurulu'nun ..... tarih ve ..... sayılı toplantısında alınan ..... numaralı kararı ile kabul edilmiştir.

Prof. Dr. Gülşah ÇEÇENER
Enstitü Müdürü

TEZ KONTROL ve BEYAN FORMU

19/01/2017

Adı Soyadı: Ayşegül YABACI

Anabilim Dalı: Biyoistatistik Anabilim Dalı

Tez Konusu: Sağkalım Verilerinde Kullanılan Ağaç Tabanlı Yöntemlerin Karşılaştırılması

ÖZELLİKLER	UYGUNDUR	UYGUN DEĞİLDİR	AÇIKLAMA
Tezin Boyutları	☑	☐	
Dış Kapak Sayfası	☑	☐	
İç Kapak Sayfası	☑	☐	
Kabul Onay Sayfası	☑	☐	
Sayfa Düzeni	☑	☐	
İçindekiler Sayfası	☑	☐	
Yazı Karakteri	☑	☐	
Satır Aralıkları	☑	☐	
Başlıklar	☑	☐	
Sayfa Numaraları	☑	☐	
Eklerin Yerleştirilmesi	☑	☐	
Tabloların Yerleştirilmesi	☑	☐	
Kaynaklar	☑	☐	

DANIŞMAN ONAYI

Unvanı Adı Soyadı: Doç. Dr. Deniz SIĞIRLI

İmza:

İÇİNDEKİLER

TABLOLAR DİZİNİ

TÜRKÇE ÖZET

Karar ağaçları, sınıflama ve regresyon probleminin çözümünde çok aşamalı ve ardışık bir yaklaşım ile karmaşık yapıdaki verileri aşamalı bir hale dönüştürerek basit bir karar verme işlemini gerçekleştirmektedir. Sağkalım ağaçları ve ormanları ise parametrik ve yarı parametrik modellerin popüler parametrik olmayan bir alternatifidir. Bu yöntemler diğer yöntemlere göre oldukça esnek olup daha önceden belirlenmeden etkileşimlerin otomatik olarak ortaya konulmasını sağlarlar.

Koşullu çıkarsama ağaçları (KÇA) yöntemi, iyi tanımlanmış koşullu çıkarsama prosedürleri içinde ağaç tabanlı regresyon modellerinin parametrik olmayan bir sınıfıdır. Koşullu çıkarsama ağaçları yöntemi sınıflayıcı, sıralayıcı, sayısal, sansürlü ve bunlara ek olarak çoklu yanıt değişkenleri ve rasgele ölçekle ölçeklendirilmiş ortak değişkenleri içeren tüm regresyon problemlerinde uygulanabilir.

Koşullu çıkarsama ormanları (KÇO), çok sayıda KÇA'nın birleştirilmesiyle gerçekleştirilen bir sağkalım ormanı yöntemidir. KÇO yöntemi, sansürlenme varlığında topluluk öğrenmesi için birleştirilmiş ve esnek bir yapı önermektedir. Bu yöntem sağdan sansürlü veriler için hastaların sağkalım zamanının tahmininde kullanılır.

Rasgele sağkalım ormanları (RSO) yöntemi, rasgele ormanlar yönteminin bir uzantısıdır. Bu yöntemde rasgelelik iki şekilde tanımlanmaktadır. İlk olarak ağacın büyümesi için verinin rasgele olarak bootstrap örnekleminden çekilmesi, ikinci olarak ise ağacın her bir düğümünde ayırma için ortak değişkenlerin alt kümelere rasgele olarak seçilmesidir. RSO yöntemi, düşük genelleme hatasını sürdürürken zengin sınıf ayrımları sağlamaktadır.

Bu çalışmada KÇA, KÇO ve RSO yöntemleri açıklanmış ve simülasyon çalışması ile sağkalım ormanları yöntemleri olan KÇO ve RSO'nun performansları karşılaştırılmıştır. Simülasyon çalışmasından elde edilen sonuçlara göre RSO yönteminin KÇO' ya göre daha iyi performans gösterdiği belirlenmiştir.

Anahtar Kelimeler: Ağaç-tabanlı yöntemler, Koşullu Çıkarsama Ağaçları, Koşullu Çıkarsama Ormanları, Rasgele Sağkalım Ormanları.

İNGİLİZCE ÖZET

Comparison of Tree-Based Methods Used in Survival Data

Decision trees, carry out a simple decision process by transforming data which are in a complex structure to a gradual form, using multi stage and sequantial approach in classification and regression problems. Survival trees and forests are popular non parametric alternatives of parametric and semi-parametric survival models. These methods are more flexible than the other methods and provide putting forward the interactions automatically which have not been determined before.

Conditional inference trees (Ctree) is a non-parametric class of regression trees embedding tree-structured regression models into a well defined theory of conditional inference procedures. It is applicable to all kinds of regression problems, including nominal, ordinal, numeric, censored as well as multivariate response variables and arbitrary measurement scales of the covariates.

Conditional inference forests (Cforest) is a survival forest method which is conducted by combining a large number of Ctrees. Cforest propose an unified and flexible framework for ensemble learning in the presence of censoring. The methodology is utilized for predicting the survival time of patients for right cencored data.

Random survival forests (RSF) methodology extends Breiman's random forests (RF) method. In RF, randomization is introduced in two forms. First, a randomly drawn bootstrap sample of the data is used to grow a tree. Second, at each node of the tree, a randomly selected subset of covariates are chosen as candidate variables for splitting. In addition, RSF enables to approximate rich classes of functions while maintaining low generalization error.

In the present study, Ctree, Cforest and RSF methods have been explanined in detail and the performances of the survival forest methods namely Cforest and RSF have been compared with the simulation study. According to results the simulation part of the study, it is determined that the RSF method performs better than the other two tree-based method.

Keywords: Tree-Based Methods, Conditional Inferences Trees, Conditional Inferences Forests, Random Survival Forests.

1. GİRİŞ

Ağaç tabanlı yöntemler, bir problemi oluşturan veri setlerinin yapısına göre bir ağaç yapısı şeklinde sınıflandırma ve regresyon modelleri oluşturmaktadır. Söz konusu ağaç yapılarının oluşturulmasında kullanılan karar kurallarının anlaşılabilir olması yöntemin kullanımını yaygın hale getirmiştir. Karar ağaçları, sınıflama ve regresyon probleminin çözümünde çok aşamalı ve ardışık bir yaklaşım ile karmaşık yapıdaki verileri aşamalı bir hale dönüştürerek basit bir karar verme işlemini gerçekleştirmektedir (Safavian ve ark., 1991).

Sınıflama ve regresyon ağaçları (SRA) bağımsız değişkene ait hiçbir ön koşul öne sürmeden kesikli ya da sürekli bağımlı değişkenin sınıf üyeliğini tahmin etmeye yarayan parametrik olmayan bir analiz yöntemidir. Bu modeller kategorik ya da sürekli bir ya da birden fazla bağımsız değişkenin kombinasyonlarını kullanarak tekrarlamalı ikili homojen bölünmelerle bağımlı değişkendeki değişimi ortaya çıkarmayı ve bağımlı değişkenin değerlerini tahmin etmeyi sağlamaktadırlar. Bağımsız değişkenlerin bağımlı değişken üzerindeki etkisinin ve aralarındaki etkileşimin görsel olarak ortaya konulması karar ağacı modelleri ile mümkündür. Ağaç modellerinin işleyiş yapısı, bağımsız değişkene ait temel basit sorulardan alınan cevapların oluşturduğu yolları (ağaç dalları) takip etmektedir. Bu yollar ise (ağaç dalları) bağımlı değişkeni hangi bağımsız değişken ya da değişkenlerin etkilediğini gösterir. Genel olarak eğer bağımlı değişken kategorik ise yöntemin adı sınıflandırma ağacı, sürekli ise regresyon ağacı olarak isimlendirilmektedir. (Breiman ve ark., 1984).

Sınıflama ve regresyonda kullanılan ağaç tabanlı yöntemler ilk olarak sosyal bilimler alanında uygulanmış olup (Morgan ve Sonquist, 1963) sonrasında istatistik alanında yapılan çalışmalarla ağaç tabanlı yöntemlerin uygulamalı ve teorik yönleri ele alınarak ağaç yapıları için yeni algoritmalar önerilmiştir (Brieman,1984). Aynı dönemlerde karar ağaçları makine öğrenme ve mühendislik alanlarında kullanılmaya başlanmıştır (Breiman ve ark., 1984; Morgan ve Sonquist,1963).

Ağaç topluluklarına ilişkin yöntemlerle ilgili olarak bagging, boosting ve rastgele orman olarak adlandırılan çeşitli yöntemler önerilmiştir. Bagging tahmin edicilerin çoklu versiyonlarının üretilmesi ve kümelenmiş tahmincilerin elde edilmesinde bu çoklu versiyonların kullanılması için ‘bagging trees’ yöntemi önerilmiştir (Breiman, 1984). Tahmin edicilerin çoklu versiyonları, öğrenme setinin boostrap tekrarlarının yapılmasıyla

elde edilmiştir ve kümeleme boostrap tahminlerinin ortalamasının alınmasıyla gerçekleştirilmiştir. Sonrasında rastgele seçilmiş alt uzayda oluşturulan çoklu ağaçların birleştirilmesi için rastgele karar ormanları yöntemi önerilmiştir (Ho, 1995). Ağaç yapraklarındaki sonsal olasılık tahminlerinin ortalamalarının birleştirilmesi için bir karar ormanı yöntemi olan rasgele alt uzay yöntemi önerilmiş olup (Ho, 1998) bu yöntem bagging ve boosting yöntemleriyle karşılaştırılmıştır.

Sağkalım analizi tıpta çeşitli hastalıklara ilişkin risk, prognostik etkenler, tedavi başarısı vb. verilerin incelenmesinde büyük önem taşımakta ve tıbbın birçok alanında uygulanmaktadır. Sağkalım analizinin amaçları; farklı zamanlarda sağkalım olasılığı tahminlerinin elde edilmesi, sağkalım süresi dağılımının tahmin edilmesi, farklı hasta gruplarının yaşam süresi dağılımlarının karşılaştırılması olarak sıralanabilmektedir. Sağkalım analizindeki en önemli gelişmeler, yaşam fonksiyonlarının tahmini için geliştirilen Kaplan-Meier yöntemi (Kaplan ve Meier,1958), iki yaşam dağılımını karşılaştırmaya olanak sağlayan log-rank testi (Mantel ve Haenszel,1959) ve sağkalım analizinde sıkça kullanılan yöntemlerden biri olan yaşam süresi üzerinde etkili olan faktörlerin belirlenmesinde kullanılan Cox regresyon (Cox,1972) modelidir.

Sağkalım analizinde kullanılan Cox regresyon analizinin varsayımları, bağımsız değişkenlerin hazard fonksiyonu üzerindeki etkilerinin loglineer olması ve bağımsız değişkenlerin loglineer fonksiyonu ile hazard fonksiyonu arasındaki ilişkinin çarpımsal olmasıdır. Bu iki varsayıma ek olarak gözlemlerin birbirinden bağımsız olmaları ve hazard oranının zamana göre değişmemesi, yani sabit olması gerekmektedir (Cox, 1972). Cox regresyon gibi yarı parametrik yöntemlerin yanı sıra hızlandırılmış başarısızlık zamanı (Accurated Failure Time - AFT) yöntemi gibi parametrik yöntemler de önerilmiştir (Wei, 1992). Sağkalım ağaçları ve ormanları ise parametrik ve yarı parametrik modellerin popüler parametrik olmayan bir alternatifidir. Bu yöntemler diğer yöntemlere göre oldukça esnek olup, belirli ayırma kriterleri kullanılarak bağımsız değişkenler ile bağımlı değişken arasındaki etkileşimlerin ortaya konulmasını sağlarlar. Ayrıca, tek bir ağaç birimleri bağımsız değişkenler dikkate alınarak sağkalım özelliklerine göre doğal olarak sınıflandırılabilirken, ağaçların kombinasyonu ile oluşturulan ağaç kümeleri ile oldukça güçlü tahmin edici araçlar elde edilebilmektedir.

Bu tez çalışmasının amacı sağkalım verilerinde kullanılan ağaç tabanlı yöntemlerden rasgele sağkalım ormanları (Random Survival Forests - RSO) ve koşullu çıkarsama ormanları (Conditional Inference Forests - KÇO) yöntemlerinin orantısal hazard

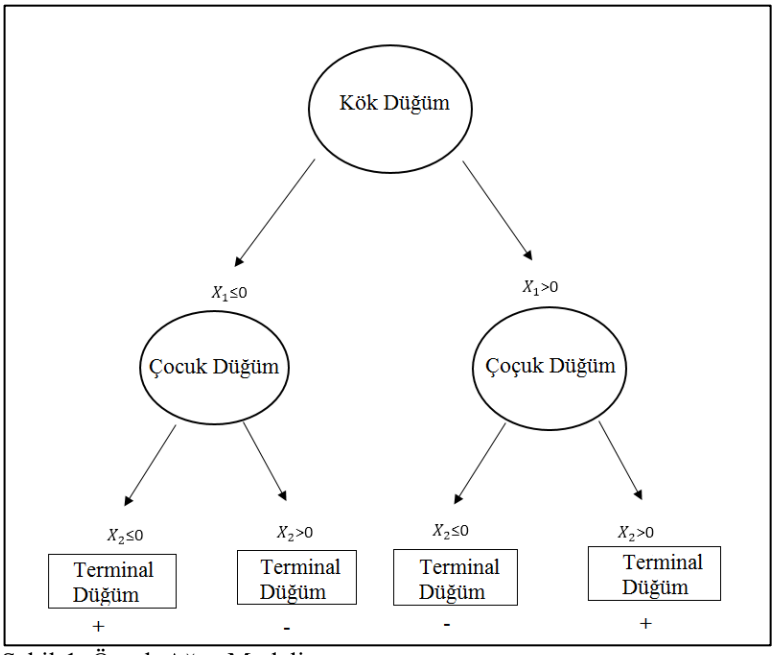
varsayımının sağlandığı ve sağlanmadığı durumlarda, sansürlü sağkalım fonksiyonunun farklı tahmin edicilerinin kullanılması durumunda ve farklı örneklem büyüklüklerinde performanslarının değerlendirilmesidir.

2.GENEL BİLGİLER

2.1.Ağaç Tabanlı Yöntemler

Ağaç tabanlı yöntemler ortak değişken uzayını ayrık bölgelere tekrarlı olarak ayırırlar ve bunlara karşılık gelen verileri gruplara (düğümlere) bölerler. Bölünecek her bir düğüm için düğümden elde edilen iki çocuk düğüm arasındaki dağılımın ayrımı için bir ayırma ölçütü hesaplanır. Her bir ortak değişken için mümkün tüm düğümler değerlendirilir ve çocuk düğümleri en iyi ayıran ayırma noktası ve buna karşılık gelen değişken seçilir. Bu prosedür düğüm sayısını artıracak şekilde her bir düğüm sadece birkaç birim içerene kadar tekrarlı olarak uygulanır. Sonuç olarak elde edilen model ikili bir ağaç olarak gösterilebilir. Son olarak ağaç elde edildikten sonra ağacı budamak ve boyutunu yeniden düzenlemek için çeşitli kurallar bulunmaktadır (Akpınar,2000).

Bir ağaç modelinde bağımsız değişkenler X_1 ve X_2 olarak ele alınırsa, X_1 ve X_2; $[-1; +1]$ tanım aralığında değişen düzgün dağılımdan rasgele olarak seçilen n_1 ve n_2 büyüklükteki örneklemlerin değişken değerlerine çarpma kuralı uygulandığında aşağıdaki sonuçlar elde edilir. Bu sonuçlara göre; $X_1 \,.\, X_2 > 0$ ise sonuç pozitif, $X_1 \,.\, X_2 \leq 0$ ise sonuç negatiftir. Bu örnekte bağımlı değişken pozitif ve negatif olmak üzere iki seviyeye sahip olmaktadır (Akpınar,2000).



Şekil-1. Örnek Ağaç Modeli

Ağaç modellerinde karar verme noktalarına düğüm denmektedir. Şekil-1'deki ağaç modelinde başlangıç düğümü, gözlem değerlerinin tümünü ihtiva eden ve en karmaşık

düğüm olan kök düğümü ya da aile düğümüdür. Kök düğümü iki alt düğüme (çocuk düğümüne) bölünür. Çocuk düğümlerinde henüz karar verme gerçekleşmemiştir. Kök düğümünden her çocuk düğümüne bölünme gerçekleştiği için çocuk düğümü aile düğümüne göre daha homojendir. Daha sonra çocuk düğümleri ayırma kriterleri dikkate alınarak karar noktalarına yani terminal düğümlere ulaşılır. Terminal düğümlerde ele alınan özelliklerin sınıf üyelikleri tanımlanır. Terminal düğümler ağaçtaki en homojen düğümler olduğu için daha sonra bölünme gerçekleşmemektedir.

Ağaç modellerinde, başlangıç düğümünden başlayarak ikili tekrarlı ayırmalarla daha homojen alt gruplara ulaşılıp karar noktalarında bağımlı değişkenin durumu tanımlanmaktadır. Bu şekilde regresyon ağaçlarındaki düğüm noktalarında yer alan gözlemler sahip oldukları bağımsız değişkenin değerlerine göre iki çocuk düğümden uygun olana atanırlar.

2.2.Ağaç Tabanlı Yöntemlerde En Sık Kullanılan Algoritmalar

2.2.1. AID (Automatic Interaction Detection) Algoritması

Sosyal ve ekonomik olayları daha güvenilir bir şekilde gösterebilmek için standart istatistiksel tekniklerin dışında yeni analiz tekniklerinin geliştirilmesi amacı ile Morgan ve Sonquist tarafından önerilen otomatik etkileşim belirleme (Automatic Interaction Detection-AID) algoritması, karar ağacı temelli ilk algoritmadır. Bu teknik en kuvvetli ve en iyi tahmini gerçekleştirebilmek için bağımlı ve bağımsız değişkenler arasındaki tüm ilişkilerin incelenmesine dayanır. Teknikte en kuvvetli ilişkiye sahip bağımsız değişken bulunduğunda veri kümesi bu bağımsız değişken değerlerine göre ikiye ayrılmakta ve süreç mümkün bölünmeler tamamlanıncaya kadar devam etmektedir. İlk temelleri AID yöntemi ile atılan karar ağacı modelleri çeşitli algoritmalar ile sürdürülmüştür (Akpınar, 2000).

2.2.2.CART (Classification and Regression Trees) Algoritması

AID adlı karar ağacı algoritmasının devamı niteliğinde olup orantısal, eşit aralıklı, sıralı ve isimsel ölçekle ölçülen değişkenler olmak üzere tüm değişken türleri için kullanılabilen sınıflandırma ve regresyon ağaçları (Classification and Regression Trees-CART) algoritması sınıflandırma ve regresyon problemlerinde bir çözüm olarak kullanılabilir.

CART, bütün bağımsız değişkenleri kullanıp verileri alt gruplara ayırarak oluşturulan bir ağaçtır. Sınıflama ve regresyon ağaçlarının en başında herhangi bir parçalanma içermeyen ve bağımlı değişkenin yer aldığı kök düğümü bulunur. İlk olarak bu kök düğümü iki parçaya ayrılır. Bu iki parçaya çocuk düğümleri adı verilir. Regresyon ağacının oluşturulmasında temel prensip, bağımlı değişkenin maksimum homojenliği sağlayacak şekilde yinelemeli olarak iki çocuk düğüme ayrılmasıdır. Ayrılma ve bölünme sonucu oluşan düğümler, alt küme olarak da adlandırılır. Ağacın oluşturulma sürecinde bağımlı değişkenlerde yinelemeli olarak oluşturulan herhangi bir çocuk düğümde homojenlik mümkün olduğunca sağlanmışsa bu düğümlerde artık parçalanma süreci sona erer ve bu düğüm terminal olarak adlandırılır. Bu süreçte çeşitli safsızlık ölçütleri (Gini, Twoing, Ordered Twoing ve Least 15 Squared Deviation) kullanılarak modele alınan bütün açıklayıcı değişkenler test edilir ve en iyi açıklayıcı değişken sırayla seçilir. Sonuçta yeni oluşacak olan düğümde en yüksek homojenliği sağlayacak şekilde açıklayıcı değişkenin kesim değeri (eğer açıklayıcı değişken kategorik ise kategorisi) belirlenir (Morgan ve Sonquist,1963; Brieman,1984).

2.2.3.CHAID (Chi-squared Automatic Interaction Detector) Algoritması

CART'ın dışında en çok kullanılan karar ağacı algoritmalarından biri olan ki-kare otomatik etkileşim belirleyici (Chi-squared Automatic Interaction Detector-CHAID) algoritması en iyi bölmeyi hesaplamak için istatistiksel olarak anlamlı bir farklılığın olmadığı, hedef değişkene uyan çiftlerde tahmin değişkeninin olası kategori çiftini birleştirmesiyle oluşturulmuştur. Bu algoritma istatistiksel anlamlılık konseptine sahip olan ve durdurma kriterinin hipotez testlerine dayandığı ilk algoritmadır. Bu algoritma istatistiksel bir testin anlamlılığını kriter olarak kullanarak bir potansiyel ön kestirici değişkenin tüm değerlerini değerlendirir. Başka bir ifade ile bağımlı değişkene göre homojen olarak değerlendirilen tüm değerleri birleştirir ve diğer tüm değerleri heterojen (benzer olmayan) olarak değerlendirir. Ardından karar ağacındaki ilk dalın formuna göre en iyi ön kestirici değişkenin seçilmesiyle her bir düğümün, seçilen değişkenin homojen değerlerinin bir grubunu oluşturmasını sağlar. Bu süreç ağaç tamamıyla büyüyene kadar sürer. Kullanılan istatistiksel test, bağımlı değişkenin ölçüm düzeyine bağlıdır. Özellikle büyük veri setlerine uygun olan, diğer algoritmalara göre daha basit olan bir algoritma kullanılarak ikili olmayan (bir düğümden ikiden fazla dal çıkabilen) ağaçlar oluşturur. Hem sürekli hem de kategorik değişkenler için regresyon tipi ya da sınıflandırma tipi problemlerde uygulanabilmektedir.

Algoritma sınıflandırma tipi problemlerde her adımda en iyi bölmeyi elde etmek için ki-kare testini, regresyon tipi problemlerde F- testini kullanmaktadır (Kass, 1980).

2.2.4.CRUISE (Classification Rule with Unbiased Interaction Selection and Estimation) Algoritması

Sınıflandırma kuralı ve yansız etkileşim seçim ve tahmin (Classification Rule with Unbiased Interaction Selection and Estimation-CRUISE) yönteminde, her bir düğümde veriye doğrusal diskriminant analizi yapılarak ayırma gerçekleştirilir. CRUISE, CART ve hızlı yansız etkin istatistiksel ağaç (Quick Unbiased Efficient Statistical Tree-QUEST) metotlarında ortak amaç, her bir terminal düğümdeki örneklemin olabildiğince saf olduğu bir ağaç elde etmektir. CRUISE yönteminin hızlı çalışan bir algoritma olması ve az sayıda dala sahip karmaşık olmayan ağaçlar oluşturması gibi avantajları bulunmaktadır (Kim, 2011).

2.2.5.QUEST (Quick Unbiased Efficient Statistical Tree) Algoritması

İsimsel ortak değişkenler, kukla değişkenler kullanılarak sıralı değişkenlermiş gibi ele alındığında değişken seçiminde yanlılık ortaya çıkmaktadır. Bu problemi ortadan kaldırmak için önerilmiş olan QUEST algoritmasında sürekli ortak değişkenler için varyans analizinden elde edilen p değerleri kullanılırken, isimsel ortak değişkenler için ise ki-kare testinden elde edilen p değerleri kullanılmaktadır (Loh ve Shih,1997).

2.2.6.GUIDE (Generalized Unbiased Interaction Detection and Estimation) Algoritması

Sürekli bağımlı değişkenler için yansız değişken seçimi sağlanması amacıyla geliştirilmiş olan bir başka algoritma olan genelleştirilmiş yansız etkileşim belirleme ve tahmin (Generalized Unbiased Interaction Detection and Estimation-GUIDE) algoritmasında ise modelden elde edilen artıkların işareti ile her bir ortak değişken arasındaki ilişki ki-kare testinden elde edilen p değeri ile ölçülmektedir. Bu algoritmada sürekli ortak değişkenler değişken seçiminden önce dört düzeye kategorize edilmekte fakat daha sonra modeller her bir düğüm içinde kategorize edilmemiş ortak değişkenlere uydurulmaktadır (Loh, 2002).

2.3.Sağkalım Analizi

Belirli bir hastalığa yakalanmış olan bir bireyin hastalık tanısı konulduktan sonra veya tedaviye başladıktan sonra daha ne kadar süre hayatta kalabileceğini ve hastalığın ne kadar süre içerisinde tekrar nüksedebileceğini tahmin etmek ya da sağkalım süresini etkileyen faktörleri incelemek amacıyla geliştirilmiş yöntemler bütünü sağkalım analizi olarak adlandırılmaktadır. Sağkalım analizi, araştırıcı tarafından tanımlanan herhangi bir olgunun ortaya çıkmasına kadar geçen sürenin incelenmesinde kullanılan çözümleme yöntemleri topluluğudur. Bu analizler belirli bir süre içinde n sayıdaki bireyden elde edilen hayatta kalma sürelerinin dağılımının yapısını inceleyerek, hayatta kalma süresini etkilediği düşünülen değişkenleri içeren modeller kurup bu yolla parametre tahmini yapmayı amaçlayan tekniklerdir. Sağkalım analizinde sıkça geçen başarısızlık terimi, araştırılan durumun denekte görülmesi durumu olarak açıklanabilir. Başarısızlık, canlılar için genelde ölüm veya hastalık anlamına gelir (Aalen, 1982).

Sağkalım analizinin temel amaçları ise aşağıdaki gibi ifade edilebilmektedir (Aalen, 1982):

- Tedaviden sonra hastaların beklenen sağkalım sürelerinin tahmini, genel sağkalım ve /veya hastalıksız sağkalım eğrilerinin elde edilmesi,
- Sağkalım karakteristiklerinin tahmin edilmesi ve yorumlanması,
- Hastalıkların doğal seyrinin incelenmesi,
- Bağımsız değişkenler ile sağkalım süresi arasındaki ilişkinin incelenmesi.

2.3.1.Sağkalım Süresi

Sağkalım analizinde, çalışma periyodunun başlangıcından (örneğin tanı tarihi, tedaviye başlama tarihi, tıbbi girişim tarihi vb.); ölüm, tedavi başarısızlığı, hastalığın nüksü, bir hastalığın ortaya çıkması ve bazen de tedaviye yanıt, iyileşme gibi olumlu olan belirli bir olaya kadar geçen süre sağkalım süresi olarak tanımlanır (Lee ve Wang, 2003; Zhao, 2008).

Sağkalım analizinde diğer süre kavramı da izlem süresidir. İzlem süresi, ilgilenilen olayın gerçekleşip gerçekleşmediğini gözlemlemek için bir çalışmadaki bireylerin ne kadar takip edileceğini belirten süredir ve çalışma başlangıcında belirlenir (Altman,1990).

Sağkalım süresi, tedavi ve hastalık ile ilgili olmasının yanında; tıpta bir cihazın (işitme cihazı, platin vb) ömrü, ya da diş hekimliğinde bir dolgunun, bir implantın ömrü de olabilir (Van Belle ve Fisher, 1993; Kim ve Dailey,2008). Sağkalım süresi; ilgilenilen olaya bağlı olarak yıl, ay, hafta, gün gibi zaman birimleri olabilmektedir (Kleinbaum ve Klein,2005).

2.3.2.Sağkalım Analizinde Veri Türleri

İleriye dönük izleme çalışmalarında en sık karşılaşılan problemlerin başında, deneklerin araştırma periyodu süresince tam olarak izlenememesi ve uzun süren izlemelerde hastalara uygulanan tedavi yöntemlerindeki olası değişiklikleri gösterebiliriz. Bir çalışmada, ilgilenilen olay bir bireyin sağkalım süresi olduğunda, her bir bireyin çalışmanın başlangıcından sonuna kadar gözlem altında bulundurulması çeşitli nedenlerden dolayı olanaksızdır ve tam olarak izlenemeyen hastalar, izlendikleri süre kadarıyla çözümlemeye alınır. Bu durumda veri sansürlüdür denir. Genel olarak 3 tip sansürlü veri bulunur. Bunlar; sağdan sansürlü veriler, soldan sansürlü veriler ve aralık sansürlü verilerdir.

i. Sağdan Sansürleme

Başarısızlık olarak adlandırılan (ölüm, bozulma, çürüme vb.) olay, çalışma için belirlenen bir sonlanım noktasına ya da çalışma sonuçlanana kadar gerçekleşmezse, bireyin yaşam süresinin uzunluğu çalışmanın sonlanma zamanının sağ tarafına geçer. Böyle bir durumda, bu bireyin yaşam süresi kesin olarak bilinmeyecek ve birey gözleme alınmayacaktır. Yani, bireyin sağkalım süresi sansürlenecektir. Bu tip sansürlemeye "sağdan sansürleme" denir.

ii. Soldan Sansürleme

İlgilenilen olayın belli bir zamandan önce gerçekleştiğinin bilindiği, ancak kesin olarak zamanının bilinmediği durumda soldan sansürlü veriler ortaya çıkar (Stevenson,2009).

iii. Aralık sansürlü veriler

Çalışma periyodu içerisinde, ilgilenilen olayın iki zaman arasında gerçekleştiği bilindiğinde fakat kesin olarak hangi zamanda meydana geldiği bilinmediğinde ortaya çıkan veri tipidir (Stevenson, 2009; Klein ve Moeschberger, 2005).

2.3.3.Sağkalım Analizinde Kullanılan Bazı Önemli Fonksiyonlar

Sağkalım sürelerini tanımlamak için temel olarak 3 tip fonksiyon kullanılmaktadır. Bunlar; olasılık yoğunluk fonksiyonu, sağkalım fonksiyonu ve hazard fonksiyonudur. Bu durumda aşağıdaki tanımlar verilebilir.

i. Olasılık Yoğunluk Fonksiyonu

T sağkalım süresini gösteren bir sürekli rasgele değişken olmak üzere *T* rasgele değişkeninin olasılık yoğunluk fonksiyonu başarısızlık zamanının tam olarak *t* anında olması olasılığıdır. Olasılık yoğunluk fonksiyonu aşağıdaki gibi tanımlanır:

$$f(t) = \lim_{\Delta t \to 0} \frac{P(t \leq T < t + \Delta t)}{\Delta t} \qquad (1)$$

ii. Sağkalım Fonksiyonu

Sağkalım fonksiyonu $S(t)$, *T* rasgele değişkeninin belirlenmiş bir sağkalım süresi *t*'den daha büyük olması olasılığını vermektedir. Bu tanıma göre;

$$S(t) = P(T > t) = \int_{s=t}^{\infty} f(s)ds \quad 0 < t < \infty \qquad (2)$$

Sağkalım fonksiyonunda *t*, kuramsal olarak 0 ile ∞ arasında bir değer aldığından sağkalım fonksiyonu düzgün bir eğri şeklinde olur.

Tüm sağkalım fonksiyonları şu üç özelliği taşımaktadır (Kleinbaum, Klein);

1. Artmayan fonksiyonlardır. *t* arttıkça sağkalım fonksiyonu monoton azalır.
2. *t*=0 anında $S(t) = S(0) = 1$ eşitliği mevcuttur. Bu hiçbir deneğin *t*=0 anında tanımlanan olayı yaşamamış olduğunu ve sağkalım olasılığının 1'e eşit olduğunu gösterir.
3. $t = \infty$ anında $S(t) = S(\infty) = 0$ eşitliği mevcuttur. Bu çalışma periyodu teoriksel açıdan limitsiz bir şekilde arttığında artık hiçbir deneğin hayatta kalmayacağını ve sağkalım olasılığının 0 olacağını gösterir.

Yukarıda tanımlanan üç özellik sağkalım fonksiyonlarının teorik özellikleridir. Gerçekte sağkalım fonksiyonları basamak şeklinde elde edilir. Ayrıca sağkalım süreleri uygulamada hiçbir zaman sonsuz olamayacağından ve çalışma süresi boyunca tanımlanan olayı bazı bireyler yaşamayacağından $S(t)$ sıfır olmak zorunda değildir.

iii. Hazard Fonksiyonu

Hazard fonksiyonu, *t* anındaki başarısızlığın koşullu yoğunluk fonksiyonu olarak tanımlanır. Hazard fonksiyonu, *t* zamanına kadar ilgilenilen olayın gerçekleşmediği bilindiğinde, $[t, t+\Delta t]$ aralığında gerçekleşmesi olasılığıdır ve ilgilenilen olayın anlık olasılığını veren bir fonksiyondur (Aalen, 1982). Hazard fonksiyonu eşitlik-3'deki gibi tanımlanır :

$$h(t) = \lim_{\Delta t \to 0} \frac{P(t \leq T \leq t + \Delta t \mid T > t)}{\Delta t} \qquad (3)$$

ıv. Kümülatif Hazard Fonksiyonu

Kümülatif hazard fonksiyonu *t* zamanına kadar hesaplanan başarısızlık hızlarının kümülatif fonksiyonudur, kümülatif hazard fonksiyonu $H(t)$ ile gösterilir (Aalen, 1982). Eşitlik-4'deki gibidir;

$$H(t) = \int_{s=0}^{t} h(s)ds \qquad (4)$$

Hazard fonksiyonuyla sağkalım fonksiyonu arasındaki ilişki eşitlik-5'deki gibidir;

$$H(t) = -\log S(t) \quad \text{veya} \quad \exp\big(-H(t)\big) = S(t) \qquad (5)$$

2.4.Sağkalım Verilerinde Kullanılan Ağaç Tabanlı Yöntemler

Sağkalım verilerinde ağaç tabanlı yöntemler, tamamlanmış veriler içeren geleneksel regresyon ve sınıflandırma ağacı yöntemlerinin bir uzantısıdır. Sağkalım verilerinin analizi için geliştirilmiş diğer parametrik ve yarı parametrik modellerin parametrik olmayan bir alternatifi olarak geliştirilmişlerdir (Gordon ve Olshen,1985). Tek bir ağaç kullanılarak

oluşturulan modellerin tahmin performansını artırmak için geliştirilmiş olan rasgele ormanlar yönteminin sansürlü verilere uyarlanması ile sağkalım verileri için çeşitli ağaç topluluğu yöntemleri önerilmiştir (Brieman, 2001).

2.4.1.Koşullu Çıkarma Ağaçları (Conditional Inference Trees- KÇA) Yöntemi

Tekrarlı bölme algoritmalarında çeşitli safsızlık ölçütleri yaklaşımları kullanılmaktadır. Bu yaklaşımın aşırı uyum ve ortak değişkenler arasından seçim yanlılığı olmak üzere iki temel dezavantajı bulunmaktadır. Aşırı uyum problemiyle ilgili olarak algoritmanın istatistiksel anlamlılığının test edilmesi ve safsızlık ölçütündeki gelişmenin anlamlı olup olmadığının ortaya konulması gerekmektedir (Mingers, 1987).

Koşullu çıkarsama ağaçları (KÇA) permütasyon testlerini kullanarak algoritmanın dağılımsal özelliklerini dikkate alan bir istatistiksel yaklaşım öne sürmektedir. KÇA, bağımlı değişken ile ortak değişkenler arasındaki ilişkiyi ölçen, istatistiklerin koşullu dağılımı dikkate alınarak, farklı ölçeklerde ölçülmüş ortak değişkenler arasından yansız bir seçim yapılmasını sağlamaktadır. Ayrıca, ortak değişkenlerden herhangi biri ve bağımlı değişken arasında anlamlı bir ilişki olup olmadığını belirlemek ve bölünmenin durdurulmasının gerekli olup olmadığını belirlemek amacı ile çoklu test prosedürleri uygulanmaktadır (Hothorn ve ark.,2006b).

2.4.1.1.Notasyon ve Algoritma

$\tilde{T}$ gerçek ölüm zamanı ve C sansürlenme zamanı olmak üzere, $T = min(\tilde{T}, C)$ bağımlı değişken ve $\Delta = I(\tilde{T} \leq C)$ durum değişkenidir. $\mathcal{X} = \mathcal{X}_1 \times ... \times \mathcal{X}_p$ örneklem uzayından gelen p boyutlu ortak değişkenler vektörü $\boldsymbol{X} = (X_1, ..., X_p)$ olarak verilsin. Ortak değişkenlerin herhangi bir ölçekte ölçülmüş olduğu durum ele alınmıştır. $\boldsymbol{X}$ ortak değişkenleri verildiğinde T bağımlı değişkenin koşullu dağılımı $\mathcal{F}_{\boldsymbol{T}|\boldsymbol{X}}$ şeklinde gösterilmek üzere, $\mathcal{F}_{\boldsymbol{T}|\boldsymbol{X}}{}'$ nın eşitlik-6'daki gibi ortak değişkenlerin bir fonksiyonuna bağlı olduğu varsayılsın.

$$\mathcal{F}_{T|\boldsymbol{X}} = \mathcal{F}(T \mid f(X_1, ... X_p)) \qquad (6)$$

$\mathcal{L}$, bazı X_{ji} (j=1,..,p ; i=1,..,n) ortak değişken değerlerinin eksik olduğu, bağımsız ve aynı dağılıma sahip gözlem değerlerinin n birimlik rasgele örneklemi olan ‘’öğrenme örneklemi’’ olmak üzere eşitlik-7’deki gibi verilsin.

$$\mathcal{L} = \{(\boldsymbol{T}_i, \Delta_i, \boldsymbol{X}_i) \;\; i = 1, \dots, n\} \tag{7}$$

Oluşturulan ağaçtaki her bir düğüm için bir birim ağırlıkları vektörü bulunmaktadır. Birim ağırlıkları vektörü $\boldsymbol{w} = [w_1, \dots w_n]$ şeklinde gösterilsin. Bir birim ağırlıkları vektörünün elemanlarının herbiri ilgili değişkenin gözlem değerleri bu düğümde yer alıyorsa ağırlık vektöründe karşılık gelen değeri 1, yer almıyorsa 0’dır. Örneğin; 5 adet gözlem değerinden oluşan bir değişkenin 1. ve 3. gözlemleri bir düğüme, 2. 4. ve 5. gözlemleri ise diğer düğüme ayrılsın. Bu durumda ilk düğüme karşılık gelen ağırlık vektörü $w = [1\ 0\ 1\ 0\ 0]$ olacaktır. Sıralı ikili ayırma gerçekleştiren algoritma adımları aşağıdaki gibidir (Hothorn ve ark., 2006b):

Adım-1: $\boldsymbol{w}$ birim ağırlıkları için p tane ortak değişkenlerden herhangi biri ile bağımlı değişken arasında bağımsızlık olduğu yönündeki genel yokluk hipotezi test edilir. Eğer bu hipotez reddedilemiyorsa durulur. Diğer durumda T ile en güçlü ilişkiye sahip j^*’ncı ortak değişken olan $\boldsymbol{X}_{j^*}$ seçilir.

Adım-2: $\boldsymbol{X}_{j^*}$’ı, A^* ve $\boldsymbol{X}_{j^*}/A^*$ biçiminde iki ayrık kümeye ayıracak $A^* \subset \boldsymbol{X}_{j^*}$ kümesi seçilir. $\boldsymbol{w}_{sol}$ ve $\boldsymbol{w}_{sağ}$ birim ağırlıkları, tüm i=1,…,n’ler için iki alt grubu $\boldsymbol{w}_{sol,i} = w_i\, \mathrm{I}(\boldsymbol{X}_{j^*i} \in A^*)$ ve $\boldsymbol{w}_{sağ,i} = w_i\, \mathrm{I}(\boldsymbol{X}_{j^*i} \notin A^*)$ fonksiyonları ile belirlenir. Burada I(.) indikatör fonksiyondur.

Adım-3: Adım-1 ve Adım-2 $\boldsymbol{w}_{sol}$ ve $\boldsymbol{w}_{sağ}$ birim ağırlıkları modifiye edilerek tekrar edilir. Adım-1’de $\boldsymbol{w}$ birim ağırlıkları ile tanımlanan her bir düğümde genel bağımsızlık yokluk hipotezi aşağıdaki gibidir.

$$H_0 = \bigcap_{j=1}^{p} H_0^j$$

Buradaki ; p tane kısmi hipotez ise aşağıdaki gibi tanımlanmaktadır:

$H_0^j : \mathcal{F}_{T|\boldsymbol{X}_j} = \mathcal{F}_T$; j = 1, … , p

Belirlediğimiz bir α anlamlılık düzeyinde H_0 hipotezi reddedilemediğinde bölme durmaktadır. Genel H_0 hipotezi reddedildiğinde T ile her bir X_j , $j = 1, \dots, p$ ortak değişkeni arasındaki ilişki kısmi hipotezler olan H_0^j hipotezleri ile test edilmektedir. Bu hipotez için test istatistikleri veya p değerleri kullanılarak en çok ilişkili olan ortak değişken seçilmektedir. w_i ağırlıkları 0 ya da 1 değerini alabilmektedir. $w_i = 1$ birim ağırlıklarına karşılık gelen elemanların tüm permütasyonlarının simetrik grubu $S(\mathcal{L}, w)$ ile gösterilsin. Bu durumda T ile X_j , ($j = 1, \dots, p$) arasındaki ilişki aşağıda verilen doğrusal test istatistiği ile ölçülür (Hothorn ve ark., 2006b).

$$\boldsymbol{T}_j(\mathcal{L}, w) = vec(\sum_{i=1}^{n} w_i g_j(X_{ji}) h((T_i, (T_1, \dots . T_n))')) \in \mathbb{R}^{p_j q} \quad (8)$$

Burada $g_j : \mathcal{X}_j \to \mathbb{R}^{p_j}$ X_j ortak değişkeninin rasgele olmayan dönüşümüdür. Sürekli ortak değişkenler için $g_{ji}(x) = x$ birim transformasyonu da uygulanabilir. Rank ya da doğrusal olmayan transformasyonlar da mümkündür. Etki fonksiyonu $h : \mathcal{T} \times \mathcal{T}^n \to \mathbb{R}^q$, simetrik permütasyondaki yanıt değişkenlerine $(T_1, \dots . T_n)$' e bağlı olup eşitlik-9'daki gibi elde edilir. Sağkalım verilerinde ise h, log-rank skoru olarak seçilebilir (Segal,1988). Eşitlik-8'deki vec operatörü pxq boyutlu bir matrisi pq boyutlu sütun vektörüne dönüştürür.

$$h(T_i, (T_1, \dots . T_n)) = \sum_{k=1}^{n} w_k \, I(T_k \leq T_i) \quad i = 1, .. n \quad (9)$$

$\boldsymbol{T}_j(\mathcal{L}, w)$'nin dağılımı H_0^j hipotezi altında X_j ve T 'nin ortak dağılımına bağlıdır, fakat bilinmemektedir. Bu nedenle permütasyon testlerinden yararlanılır. Bağımlı değişkenin $\sigma \in S(\mathcal{L}, w)$ şeklindeki permütasyonları verildiğinde H_0 hipotezi altında $\boldsymbol{T}_j(\mathcal{L}, w)$' nin koşullu beklenen değeri $\mu_j \in \mathbb{R}^{p_j q}$ ve kovaryansı $\sum_j \in \mathbb{R}^{p_j q \times p_j q}$ Strasser ve Weber tarafından aşağıdaki eşitlik-10'da olduğu gibi verilmiştir (Hothorn ve ark., 2006b)

$$\mu_j = \mathbb{E}\left(\boldsymbol{T}_j(\mathcal{L}, w) \middle| S(\mathcal{L}, w)\right) = vec((\sum_{i=1}^{n} w_i g_j(X_{ji})) \, \mathbb{E}\,(h \mid S(\mathcal{L}, w))') \quad (10)$$

$$\sum_j = \mathbb{V}(\boldsymbol{T}_j(\mathcal{L}, w) \mid S(\mathcal{L}, w)\,) \quad (11)$$

$$= \frac{w_.}{w_. - 1}\, \mathbb{V}(h \mid S(\mathcal{L}, w)) \otimes \left(\sum_i w_i g_j(X_{ji}) \otimes w_i g_j(X_{ji})'\right)$$
$$- \frac{1}{w_. - 1}(h \mid S(\mathcal{L}, w)) \otimes \left(\sum_i w_i g_j(X_{ji})\right) \otimes \left(\sum_i w_i g_j(X_{ji})\right)'$$

Burada $w_. = \sum_{i=1}^{n} w_i$ birim ağırlıklarının toplamı, $\otimes$ ise kroneker çarpımı olup etki fonksiyonunun koşullu beklenen değeri

$$\mathbb{E}\,(h \mid S(\mathcal{L}, w)) = w_.^{-1} \sum_i w_i h(T_i, (T_1, \ldots . T_n)) \in \mathbb{R}^q \tag{12}$$

ve $q \times q$ boyutlu kovaryans matrisi eşitlik-13'de olduğu gibidir.

$$\mathbb{V}(h \mid S(\mathcal{L}, w) = w_.^{-1} \sum_i w_i (h(T_i, (T_1, \ldots . T_n)) - \mathbb{E}\,(h \mid S(\mathcal{L}, w)))(h(T_i, (T_1, \ldots . T_n))$$
$$- \mathbb{E}\,(h \mid S(\mathcal{L}, w)))' \tag{13}$$

Bu koşullu beklenen değer ve kovaryans kullanılarak eşitlik-8'de verilen $\boldsymbol{T}_j(\mathcal{L}, w) \in \mathbb{R}^{pq}$ doğrusal istatistiği $p \in \{p_1, \ldots, p_p\}$ değerleri için standartlaştırılabilir. Orjinal çok değişkenli doğrusal istatistik olan $t \in \mathbb{R}^{pq}$ 'yu tek değişkenli test istatistiğine dönüştüren eşitlik-14'de verilmiştir (Hothorn ve ark., 2006b).

$$c_{max}(\boldsymbol{t}, \mu, \Sigma) = \max_{k=1,\ldots pq} \left| \frac{(\boldsymbol{t} - \mu)_k}{\sqrt{(\Sigma)_{kk}}} \right| \tag{14}$$

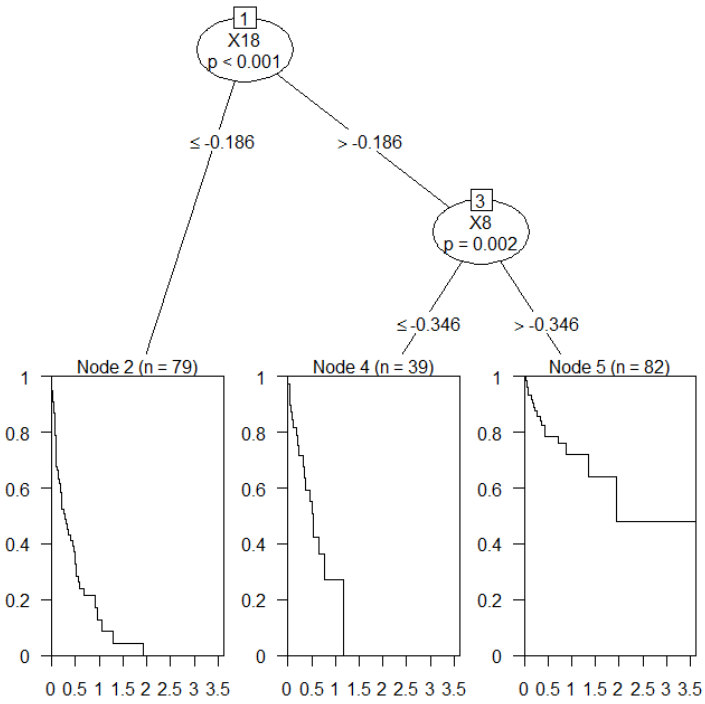
Başka bir alternatif ise $c_{quad}(\boldsymbol{t}, \mu, \Sigma) = (\boldsymbol{t} - \mu)\Sigma^{+}\,(\boldsymbol{t} - \mu)'$ 'dir. Ancak bu Σ 'nın Moore-Penrose tersi olan Σ^{+}'yı içerdiği için hesaplama bakımından daha karmaşıktır. Ayrıca bütün ortak değişkenler aynı ölçekte olmadığında yani $j = 2, \ldots, p$ için $p_1 = p_j$ olmadığında $c = (t_j, \mu_j, \Sigma_j),\ j = 1, \ldots, p$ test istatistikleri yansız bir şekilde karşılaştırılamaz. Yansız bir değişken seçimi gerçekleştirmek için p değeri ölçeğine geçilmelidir. Çünkü $c(\boldsymbol{T}_j(\mathcal{L}, w), \mu_j, \Sigma_j)$ test istatistiğinin koşullu dağılımı için p değerleri farklı ölçeklerde ölçülmüş olan ortak değişkenler arasında direk olarak karşılaştırılabilirler.

Adım-1'de minimum p değerine sahip olan ortak değişkenler seçilir. Bu ortak değişken $j^* = argmin_{j=1,..m}\ P_j$ olmak üzere X_j^* ile gösterilir. Burada P_j eşitlik-15'de olduğu gibi hesaplanır ve H_0^j hipotezi için p değeridir.

$$P_j = \mathbb{P}_{H_0^j}(c(\boldsymbol{T}_j(\mathcal{L}, w), \mu_j, \Sigma_j) \geq c(t_j, \mu_j, \Sigma_j) \mid S(\mathcal{L}, w)) \qquad (15)$$

Ancak bu yaklaşım kayıp değerlere sahip öğrenme örneklemleri için çok uygun değildir. Genel H_0 hipotezini test etmek için yaygın olarak kabul edilen yaklaşımlar $P_1, \ldots, P_p$ değerlerine dayanan çoklu test prosedürleridir. Çoklu test prosedürlerine Bonferroni düzeltmeli p değeri ve $min\ p$ değeri yeniden örnekleme yaklaşımı örnek verilebilir. Eğer düzeltilmiş p değerlerinin minimumu α değerinden küçükse H_0 hipotezi rededilir aksi durumda algoritma durdurulur. Bu durumda elde edilecek ağacın büyüklüğünü belirleyecek tek parametre α olacaktır (Hothorn ve ark., 2006b).

Şekil-2'de KÇA yöntemi ile elde edilmiş örnek bir ağaç verilmiştir.



Şekil-2. Örnek KÇA Ağacı

2.4.1.2. Ayırma Kriteri

Adım-1'de seçilen ortak değişken $X_j^{*\prime}\iota$, Adım-2'de ikiye bölmek için permütasyon testi kullanılır. Bölmenin uyumu, $\boldsymbol{T}_j(\mathcal{L}, w)$ test istatistiğin özel durumu olan X_j^* 'ın mümkün tüm A alt gruplarının test istatistiği eşitlik-16'da olduğu gibi hesaplanır.

$$\boldsymbol{T}_{j^*}^{A}(\mathcal{L}, w) = vec\left(\sum_{i=1}^{n} w_i\, I\big(X_{j*i} \in A\big) h(T_i, (T_1, \ldots . T_n)'\right) \in \mathbb{R}^q \qquad (16)$$

Bu doğrusal istatistik, $\{T_i | w_i > 0\ ve\ X_{ji} \in A; i = 1, .., n\}$ ve $\{T_i | w_i > 0\ ve\ X_{ji} \notin A; i = 1, .. n\}$ örneklemleri arasındaki uyumsuzluğu ölçen iki örneklem test istatistiği verir. Koşullu beklenen değer $\mu_{j^*}^{A}$ ve kovaryans $\Sigma_{j^*}^{A}$ sırasıyla eşitlik-10 ve eşitlik-11'deki gibi hesaplanmaktadır. Bu beklenen değer ve kovaryans kullanılarak $\boldsymbol{T}_{j^*}^{A}(\mathcal{L}, w)$'nin standartlaştırılmış test istatistiği $c(t_{j^*}^{A}, \mu_{j^*}^{A}, \Sigma_{j^*}^{A})$ elde edilir. Bu test istatistiklerinden maksimum olana karşılık gelen ayrım A^* ile gösterilir. Mümkün tüm A altkümeleri üzerinden maksimize edilen test istatistiği eşitlik-17'deki gibidir (Hothorn ve ark., 2006b).

$$A^* = argmax_A\, c(t_{j^*}^{A}, \mu_{j^*}^{A}, \Sigma_{j^*}^{A}) \qquad (17)$$

Daha sonra algoritmanın Adım-2'de belirtildiği gibi w_{sol} ve $w_{sağ}$ birim ağırlıkları, tüm i=1,…,n'ler için iki alt grubu $\boldsymbol{w}_{sol,i} = w_i\, I(\boldsymbol{X}_{j^*i} \in A^*)$ ve $\boldsymbol{w}_{sağ,i} = w_i\, I(\boldsymbol{X}_{j^*i} \notin A^*)$ fonksiyonları ile belirlenir ve ağırlıklar modifiye edilerek algoritmada Adım-1 ve Adım-2 tekrar edilir.

Çok yüksek α değeri ile ağacın veriye aşırı uyum göstermesini engellemek için son adım olarak ağaç çeşitli yöntemlerle budanmalıdır. Örneğin, terminal bölmeler α anlamlılık düzeyinde anlamlı oluncaya kadar bütün terminal düğümler elimine edilebilir. Burada α başta belirlenen α'dan oldukça düşük bir değer olarak alınır. Bu yaklaşım ilk olarak Segal (1998) tarafından verilmiştir. Alternatif bir yaklaşım ise Molinaro, Dudoit ve van der Laan (2004) tarafından önerilen ağırlıklandırma şemasıdır. Bu şemada $w(x)$ birim ağırlıklarının tahmini için ağırlıklı Kaplan-Meier fonksiyonu kullanılır (Hothorn ve ark., 2006b).

2.4.2. Koşullu Çıkarsama Ormanları (Conditional Inference Forest -KÇO) Yöntemi

X ortak değişkeni verildiğinde T bağımlı değişkenin koşullu dağılım fonksiyonunun $f: \mathcal{X} \to \mathbb{R}$ fonksiyonu ile X'e bağlı olduğu varsayılsın. Bu durumda $\mathcal{F}_{T|X} = \mathcal{F}_{T|f(X)}$ olmaktadır. Koşullu sansürleme sağkalım fonksiyonu $G(T \mid \mathbf{X}) \approx \mathbb{P}(C > t \mid \mathbf{X} = \mathrm{x})$ biçiminde verilsin. Ψ, f regresyon fonksiyonu için tüm aday tahminciler olan $\psi: \mathcal{X} \to \mathbb{R}$'lerin fonksiyon uzayı olsun. f regresyon fonksiyonu tahmini, tam veri kayıp fonksiyonu olan $L_{tam}\big(T, \psi(\mathbf{X})\big)$ ile tanımlanan risk fonksiyonunun beklenen değeri minimize edilerek bulunur. Ancak sansürlü gözlem olduğunda tüm verilere ulaşılamayacağından tam veri fonksiyonu hesaplanamaz. Dolayısıyla tam veri fonksiyonu yerine, gözlenen veri fonksiyonu $L = (T, \psi(\mathbf{X}) \mid \eta)$ kullanılır. Bu durumda gözlenen veri fonksiyonunun beklenen değeri eşitlik-18'de verildiği gibi elde edilir. Burada tam veri fonksiyonunun beklenen değerinin $\psi \in \Psi$ aday tahmincilerine göre minimize edilmesi amaçlanmaktadır (Hothorn ve ark., 2006a).

$$\mathbb{E}_{T,\boldsymbol{X}} L_{tam}\big(T, \psi(\mathbf{X})\big) = \int L(T, \psi(\mathbf{X}) \mid \eta)\, d\mathcal{F}_{T,\Delta,\mathbf{X}} = \mathbb{E}_{T,\Delta,\mathbf{X}}\, L(T, \psi(\mathbf{X}) \mid \eta) \tag{18}$$

Eşitlik-18'de η nuisance parametresidir ve koşullu sansürlenme sağkalım fonksiyonu olarak alınabilir. Gözlenen kayıp veri fonksiyonu, $G(T \mid \boldsymbol{X})^{-1}$ kullanılarak eşitlik-19'daki gibi tanımlanabilir (Hothorn ve ark., 2006a).

$$L(T, \psi(\boldsymbol{X}) \mid G) = L\big(T, \psi(\boldsymbol{X})\big) \frac{\Delta}{G(T \mid \boldsymbol{X})} \tag{19}$$

Görüldüğü gibi esasında tam veri kayıp fonksiyonu, $\boldsymbol{X}$ ortak değişkeni verildiğinde T zamanından sonra sansürlenme olasılığının tersi ile ağırlıklandırılır**.** Bu durumda gözlenen veri kayıp fonksiyonunun beklenen değeri eşitlik-20'deki gibi elde edilir.

$$\widehat{\mathbb{E}}_{T,\Delta,\boldsymbol{X}}\, L(T, \psi(\boldsymbol{X}) \mid G) = n^{-1} \sum_{i=1}^{n} L\big(T_i, \psi(\boldsymbol{X_i}) \mid \hat{G}\big) = n^{-1} \sum_{i=1}^{n} L\big(T_i, \psi(\boldsymbol{X_i}) \mid \hat{G}\big) \frac{\Delta_i}{\hat{G}(T_i \mid \boldsymbol{X_i})} \tag{20}$$

Regresyon fonksiyonu tahmincisi $\hat{f}$ bu eşitliğin aday tahminciler $\psi \in \Psi$'ye göre minimize edilmesi ile elde edilir. Burada G koşullu sansürlü sağkalım fonksiyonu bilinmemektedir ve yerine onun tahmincisi $\hat{G}$ kullanılmaktadır. $\hat{G}$ tahmincisi olarak

parametrik olmayan tahmin edici, Cox tahmin edicisi veya toplamsal Aalen regresyon tahmin edicisi kullanılabilir. Burada $w_i = \Delta_i \hat{G}(T_i \mid \boldsymbol{X_i})^{-1}$ olmak üzere, $\mathbf{w} = (w_1, w_2, .., w_n)$ IPC (sansürlü ağırlıkların ters olasılığı - inverse probability of censoring weights) olarak adlandırılır. Hothorn ve arkadaşları eşitlik-20'de verilen gözlenen veri kayıp fonksiyonunun beklenen değerini minimize eden ψ değerlerini bulmak için koşullu sağkalım ormanları (conditional ınference forest- cforest) algoritmasını önermişlerdir(Hothorn ve ark., 2006a).

2.4.2.1. Algoritma

Gözlenen öğrenme örneklemi $\mathcal{L} = \{(T_i, \Delta_i, \boldsymbol{X_i}); i = 1, \dots, n\}$ ve $w_i = \Delta_i \hat{G}(T_i \mid \boldsymbol{X_i})^{-1}$'den $\mathbf{w}$ ağırlık vektörü hesaplanır. Eğer öğrenme örneklemi sansürlü gözlem değeri içeriyorsa bu gözlemler için $\Delta_i = 0$ olduğundan $w_i = 0$ olur. Algoritmanın adımları ise aşağıdaki gibidir (Hothorn ve ark., 2006a):

Adım 1 : $m = 1$ ve $M > 1$ olarak ayarlanır.

Adım 2 : n ve $(\sum_{i=1}^{n} w_i)^{-1}\boldsymbol{w}$ parametreli multinomial dağılımdan birim sayılarının bir rasgele vektörü $\mathbf{v}_m = (v_{m1}, \dots, v_{mn})$ çekilir.

Adım 3 : Bir regresyon ağacı ile $\mathcal{X}$ örneklem uzayı $K(m)$ tane hücrelere ayrılarak $\pi_m = (R_{m1}, \dots, R_{mK(m)})$ parçaları oluşturulur. Bu ağaç birim sayıları $\mathbf{v}_m$ ile öğrenme örneklemi $\mathcal{L}$ kullanılarak oluşturulur. $\mathcal{L}$ öğrenme örnekleminin permütasyonlarında i'nci gözlem v_{mi} kez yer alır.

Adım 4 : m birer artırılarak $m = M$ olana kadar adım 2 ve adım 3 tekrarlanır.

Adım-3'de, Adım-2'de belirlenen öğrenme örneklemi kullanılarak koşullu çıkarsama ağaçları (ctree) algoritması ile bir sağkalım ağacı elde edilir. Algoritma sonunda M tane KÇA elde edilmiş olacaktır.

$\mathcal{T}_m$ oluşturulan m'nci sağkalım ağacı , $\mathcal{T}_m(\boldsymbol{x})$ $\boldsymbol{x}$ ortak değişken değeri ile m'nci ağaçtaki terminal düğümü göstersin. Her bir $\boldsymbol{x}$ değeri tek bir terminal düğümde yer alacaktır.

$\tilde{N}_i(s) = I(T_i \leq s, \Delta_i = 1)$ ve $\tilde{Z}_i(s) = I(T_i > s)$ olmak üzere;

$$\tilde{N}_m^*(s, x) = \sum_{i=1}^{n} v_{im} I(X_i \in \mathcal{T}_m(x)) \tilde{N}_i(s) \tag{21}$$

$$\tilde{Z}_m^*(s, x) = \sum_{i=1}^{n} v_{im} I\big(X_i \in \mathcal{T}_m(x)\big) \tilde{Z}_i(s) \tag{22}$$

elde edilir. Burada $\tilde{N}_m^*(s,x)$ ile $\tilde{Z}_m^*(s,x)$ sırasıyla $\boldsymbol{x}$ ortak değişken değerine karşılık gelen terminal düğümdeki s zamanına kadar olan sansürsüz olayların sayısını ve s zamanında risk altındaki birim sayısını verir. Bu durumda $\boldsymbol{x}$ ortak değişken değeri verildiğinde *t* zamanı için topluluk sağkalım fonksiyonu eşitlik-23'de olduğu gibidir.

$$\hat{S}^{cforest}(\,t \mid x\,) = \prod_{s \le t}\left(1 - \frac{\sum_{m=1}^{M}\tilde{N}_m^*(s,x)}{\sum_{m=1}^{M}\tilde{Z}_m^*(s,x)}\right) \tag{23}$$

Bu eşitlik asimptotik olarak eşitlik-24'e eşittir.

$$\exp\left(-\int_{s=0}^{t}\frac{\sum_{m=1}^{M}\tilde{N}_m^*(s,x)}{\sum_{m=1}^{M}\tilde{Z}_m^*(s,x)}\right) \tag{24}$$

$\boldsymbol{x}$ ortak değişkenli bir birimin sağkalım süresinin tahmin edilmesi istendiğinde tahmin ağırlıklarının hesaplanması gerekmektedir. $\boldsymbol{x}$ ortak değişken değerine sahip *i*'nci birim için tahmin ağırlığı $a_i(\boldsymbol{x})$ olarak gösterilir ve eşitlik-25'deki elde edilir.

$$a_i(\boldsymbol{x}) = \sum_{m=1}^{M} v_{mi} \sum_{k=1}^{K(m)} I(\boldsymbol{X}_i \in R_{mk}\ ve\ \boldsymbol{x} \in R_{mk}); \quad i = 1, \dots, n \tag{25}$$

$a_i(\boldsymbol{x})$ tahmin ağırlığı, *x* değerinin öğrenme örneklemindeki *i*'nci birim ile kaç defa aynı hücreye düştüğü hesaplanarak *x* ile X_i $(i = 1, \dots, n)$ arasındaki benzerliği ölçer (Brieman, 1996 ; Hothorn ve ark.,2004).

2.4.3. Rasgele Sağkalım Ormanlar (Random Survival Forest-RSO) Yöntemi

Ishwaran ve arkadaşları (2008a) tarafından geliştirilen rasgele sağkalım ormanlar yöntemi bir topluluk öğrenme yöntemi olup birden fazla sağkalım ağacının sonucunu birleştirerek bir risk tahmin modeli oluşturur. Rasgele ormanlar yöntemi, rasgele örneklemeyi ve topluluk yöntemlerindeki geliştirilmiş özellikleri içermesi nedeniyle daha iyi genellemeler sunar ve geçerli tahminlerde bulunur. Mümkün olduğunca birbirinden farklı ağaçlar oluşturarak da düşük korelasyon yapısında bir topluluk elde edilir.

2.4.3.1 Algoritma

Kullanılan genel algoritma aşağıdaki gibidir:

Adım-1: Orijinal veriden M tane bootstrap örneklemi çekilir. Her bir bootstrap örneklemi orijinal verinin ortalama %37'sini dışarıda bırakmalıdır. Dışarıda bırakılan veri out-of-bag data (OOB) olarak adlandırılmaktadır.

Adım-2: Her bir bootstrap örneklemi için bir sağkalım ağacı oluşturulur. Ağacın her düğümünde, rasgele olarak $\sqrt{p}$ aday değişkeni seçilir. Düğüm, çocuk düğümler arasındaki sağkalım farkını maksimize eden aday değişkenler kullanılarak ayrılır.

Adım-3: Her bir terminal düğümde en az 1 tane ilgilenilen olay, gözlenen birim kalana kadar bölme işlemine devam edilir.

Adım-4: Her bir ağaç için kümülatif hazard fonksiyonu (KHF) hesaplanır. Topluluk kümülatif hazard fonksiyonunu elde etmek için ortalama alınır.

Adım-5: OOB verisini kullanarak topluluk kümülatif hazard fonksiyonu için tahmin hatası hesaplanır (Ishwaran ve ark., 2008a).

2.4.3.2. Ayırma Kriteri

Rasgele sağkalım ormanları yönteminde ayırma kriteri önemli bir unsurdur. Algoritmada ayırma kriteri olarak iki yöntem kullanılabilmektedir. Bunlardan ilki log-rank ayrımı, ikincisi ise log-rank skor ayrımıdır (Segal, 1988; Ciampi ve ark., 1986; Hothorn ve Lausen, 2003).

i. Logrank Ayırma Kriteri

T_i ; $i = 1, .., n$, i'nci birimin sağkalım zamanı olmak üzere X_j ortak değişkeni için bir düğümdeki ayrım, c kesim noktasına göre $X_j \leq c$ ve $X_j > c$ şeklinde gösterilsin. $z = 1, \dots, N$ için $s_1 < s_2 < \cdots < s_z$ bir düğümdeki ayrık ölüm zamanları olsun. $\tilde{N}^*_{md}(s_z, x)$, m'nci ağaç için d=1,2'nci çocuk düğümlerinde s_z zamanında ölen kişilerin sayısını göstersin. $\tilde{N}^*_m(s_z, x) = \tilde{N}^*_{m1}(s_z, x) + \tilde{N}^*_{m2}(s_z, x)$ biçimindedir. $\tilde{Z}^*_{md}(s_z, x)$, m'nci ağaç için d=1,2'nci çocuk düğümlerinde s_z zamanında risk altındaki birim sayısını göstermek üzere, $\tilde{Z}^*_m(s_z, x) = \tilde{Z}^*_{m1}(s_z, x) + \tilde{Z}^*_{m2}(s_z, x)$ olur ve $\tilde{Z}^*_{m1}(s_z, x) = \#\{T_i \geq s_z\, , x_i \leq c\}$, $\tilde{Z}^*_{m2}(s_z, x) = \#\{T_i \geq s_z\, , x_i > c\}$ olur. Burada x_i, i'nci birim için X_j ortak değişkeninin aldığı değerdir. n_d, d'nci çocuk düğümdeki gözlenen toplam birim sayısı olsun. Böylece $n_1 = \#\{i: x_i \leq c\}$ ve $n_2 = \#\{i: x_i > c\}$ olmak üzere $n = n_1 + n_2$ 'dir. X_j ortak değişkeninin c kesim değeri için log-rank test istatistiği eşitlik-26'daki gibidir.

$$LogRank(X_j, c)$$

$$= \frac{\sum_{z=1}^{N}\left(\tilde{N}_{m1}^{*}(s_z, X_j) - \tilde{Z}_{m1}^{*}(s_z, X_j)\frac{\tilde{N}_{m}^{*}(s_z, X_j)}{\tilde{Z}_{m}^{*}(s_z, X_j)}\right)}{\sqrt{\sum_{z=1}^{N}\frac{\tilde{Z}_{m1}^{*}(s_z, X_j)}{\tilde{Z}_{m}^{*}(s_z, X_j)}\left((1-\frac{\tilde{Z}_{m1}^{*}(s_z, X_j)}{\tilde{Z}_{m}^{*}(s_z, X_j)})(\frac{\tilde{Z}_{m}^{*}(s_z, X_j) - \tilde{N}_{m}^{*}(s_z, X_j)}{\tilde{Z}_{m}^{*}(s_z, X_j) - 1}\right)\tilde{N}_{m}^{*}(s_z, X_j)\right)}} \quad (26)$$

$\left|LogRank(X_j, c)\right|$, düğüm ayrımı için bir ölçü vermektedir. $\left|LogRank(X_j, c)\right|$ ölçüm değeri ne kadar büyük olursa iki terminal düğüm arasındaki fark o kadar büyük olur ve en iyi ayrımı verir. Ortak değişkenler arasında ve kesim değerleri arasında bütün X_j ortak değişkenleri ve c kesim noktaları için $\left|LogRank(X_j^*, c^*)\right| \geq \left|LogRank(X_j, c)\right|$ değerini veren X_j^* ortak değişkeni ve c^* kesim değeri bulunarak en iyi ayrım belirlenir (Segal, 1988, Hothorn ve Lausen, 2003).

***ii*. Logrankskor Ayırma Kriteri**

Diğer bir ayırma kuralı, Hothorn ve Lausen (2003) tarafından önerilen log-rank skor ayrım kuralıdır. Bu kuralı tanımlamak için X_j ortak değişkeninin aldığı değerlerin $x_1 \leq x_2 \leq \cdots \leq x_n$ olarak sıralandığı varsayılsın. Her bir T_i sağkalım zamanı için ranklar eşitlik-27'deki gibi elde edilir.

$$\alpha_i = \Delta_i - \sum_{k=1}^{\Gamma_i}\frac{\Delta_k}{n - \Gamma_k + 1} \quad (27)$$

Burada $\Gamma_k = \#\{s: T_s \leq T_k\}$ eşittir. Bu durumda log-rank skor istatistiği eşitlik-28'de olduğu gibi elde edilir.

$$LogRankskor(X_j, c) = \frac{\sum_{x_i \leq c}\alpha_i - n_1\overline{\alpha}}{\sqrt{n_1(1-\frac{n_1}{n})s_\alpha^2}} \quad (28)$$

Eşitlikte $\overline{\alpha}$ ve s_α^2 sırasıyla rankların örneklem ortalaması ve örneklem varyansını göstermektedir. $LogRankskor(X_j, c)$, düğüm ayrımı için log-rank skor ölçüsü vermektedir. Bu değeri maksimum yapan X_j ortak değişkeni ve c kesim değeri seçilir.

2.4.3.3. Topluluk Kümülatif Hazard Fonksiyonunun (KHF) Elde Edilmesi

Ishwaran ve arkadaşları (2008a) tarafından önerilen rasgele sağkalım ağaçlarında ağaç topluluğu, ağaç tabanlı Nelson-Aalen tahmin edicilerinin birleştirilmesi ile oluşur. m'inci

bootstrap örneklemi için oluşturulan ağacın her bir terminal düğümündeki Nelson-Aalen tahmin edicisi ile koşullu kümülatif hazard fonksiyonu eşitlik-29'daki gibi elde edilir.

$$\widehat{H}_m(\,t \mid x\,) = \int_0^t \frac{\widetilde{N}_m^*(s,x)}{\widetilde{Z}_m^*(s,x)} \tag{29}$$

Bir topluluk KHF'sini hesaplamak için *M* tane sağkalım ağacı üzerinde ortalama alınır ve topluluk KHF'si eşitlik-30'daki gibi elde edilir.

$$H_{\mathrm{e}}(t|\mathrm{x}) = \frac{1}{M}\sum_{i=1}^{M} \widehat{H}_m(\,t \mid x\,) \tag{30}$$

Unutulmamalıdır ki ormandaki her ağaç bağımsız bootstrap örneklemini kullanarak büyür. Ayrıca bir düğümdeki tüm birimler aynı KHF'ye sahiptir. Dolayısıyla *i*'nci birim için KHF, x_i'nin terminal düğümü için Nelson-Aalen tahmin edicisidir.

$\widehat{H}_m(\,t \mid x\,) = \widehat{H}_m(\,t \mid x_i\,)$ *i*=1,…,n için x_i, *i*'nci birimin *j* boyutlu ortak değişkenler vektörüdür. Eğer *m*'nci bootstrap örneklemi için *i*, OOB birimi ise $I_{i,m} = 1$ aksi durumda $I_{i,m} = 0$ olsun. Bu durumda OOB için topluluk KHF'si eşitlik-31'de verildiği gibidir.

$$H_{\mathrm{e}}^*(t|\mathrm{x}) = \frac{\sum_{m=1}^{M} I_{i,m}\,\widehat{H}_m(\,t \mid x\,)}{\sum_{m=1}^{M} I_{i,m}} \tag{31}$$

Rasgele sağkalım ağaçları için topluluk sağkalım fonksiyonu ise eşitlik-32'deki gibidir.

$$\hat{S}^{rsf}(\,t \mid \boldsymbol{x}\,) = exp\left(-\frac{1}{M}\sum_{m=1}^{M} \widehat{H}_m(\,t \mid x\,)\right) \tag{32}$$

2.4.3.4. Rasgele Sağkalım Ormanları Yönteminin Özellikleri

2.4.3.4.1. Genelleme Hatası

Veri setinden bir bootstrap örneklemi seçildiğinde bazı gözlemler ağaç oluşturma aşamasında yer almaz. OOB olarak adlandırdığımız bu gözlemler ile genelleme hatasına yönelik bir iç tahmin yapılır. OOB hata oranını elde etmek için her ağaç OOB veri seti için bir sınıf değeri tahmin eder ve bu tahminler kaydedilir. Herhangi bir noktada her bir

gözlem için OOB olduğu ağaçlardaki hata oranı tahminlerinin ortalaması alınarak genelleme hatası hesaplanabilir. Genel bir hata oranı ise tüm gözlemlerin ortalaması alınarak hesaplanabilir (Ishwaran ve ark., 2008a).

2.4.3.4.2. Parametrelerin Ayarlanması

Rasgele sağkalım ormanları yönteminde karar ormanı oluşturulurken belirlenmesi gereken iki parametre vardır; bunlardan ilki her düğümde rasgele seçilecek olan değişken sayısı ikincisi oluşturulacak ağaç sayısıdır. Yöntem bu parametrelerin seçiminde hassas bir yapı göstermez. Breiman, bu parametrelerin seçimi için bazı önerilerde bulunmuştur. Pek çok sınıflandırma problemi için her düğümde rasgele seçilecek olan değişken sayısı $\sqrt{p}$ olarak alınır. Burada p, veri setindeki bağımsız değişken sayısını göstermektedir.

Rasgele sağkalım ormanları yönteminde ormana daha fazla sayıda ağaç eklemek aşırı uyumun oluşmasına neden olmamaktadır. Oluşturulan ağaçların sayısı için ilgilenilen önemli nokta ağaçların yeterli büyüklükte olmasıdır. Bu sayı, OOB hata oranı kullanılarak kontrol edilir. OOB hata oranı belli bir ağaç sayısından sonra sabit bir değere yakınsamaktadır (Ishwaran ve ark., 2008a).

2.5. *G* Koşullu Sansürlü Sağkalım Fonksiyonunun Tahmininde Kullanılan Tahminciler

i. Parametrik Olmayan Tahmin Edici

$G(T \mid X) \approx P(C > t \mid X = x)$ sansürlenme zamanlarının koşullu sağkalım fonksiyonu ve $K(t)$ herhangi bir kernel fonksiyonu olmak üzere Graf ve ark. (1999) tarafından kullanılan parametrik olmayan tahmin edicisi eşitlik-33'deki gibi verilmektedir (Gerds ve Schumacher, 2007).

$$\hat{G}_{NonPar} = \left\{ G: sup_t \frac{|G(T \mid X) - G(T|X')|}{|X - X'|^{\alpha}} \leq K(t) > 0 \right\} \tag{33}$$

ii. Cox Tahmin Edicisi

α regresyon katsayısı ve $H_0(t)$ başlangıç kümülatif hazard fonksiyonu ile Cox regresyon tahmin edicisi eşitlik-34'deki gibi verilmektedir (Gerds ve Schumacher, 2007).

$$\hat{G}_{Cox} = \left\{ G_{\alpha, H_0(t)}: G(T \mid X) = exp\{-\exp(\alpha' X) H_0(t)\}; \alpha \in \mathbb{R}^d \right\} \tag{34}$$

***iii*. Aalen Tahmin Edicisi**

$\alpha(t)$ zamana bağlı regresyon katsayısı ile toplamsal Aalen regresyon tahmin edicisi eşitlik-35'deki gibi verilmektedir (Gerds ve Schumacher, 2007).

$$\hat{G}_{Aalen} = \left\{ G_{\alpha} : G(T \mid X) = exp\left\{ - \int_{s=0}^{t} X'\alpha(s).ds \right\} \right\} \quad (35)$$

2.6. Model Performansının Değerlendirilmesinde Kullanılan Ölçütler

2.6.1. Brier Skoru

Sağkalım analizinde model performansını değerlendirmek için çeşitli ölçütler bulunmaktadır. Bunlardan bir tanesi zamana bağlı beklenen Brier skorudur. Brier skoru modelin ürettiği olasılık değeri ile durum değişkeninin gerçek değeri arasındaki farkın kareli ortalamasının beklenen değeri olarak tanımlanmaktadır. Skorun düşük olması istenilen bir durum olup skor ne kadar sıfıra yakınsa yapılan sınıf tahminleri o kadar güvenilirdir. t zamanı için i'nci birimin durumu $\Delta_i = I(\tilde{T}_i \leq t)$, ve X ortak değişkeni verildiğinde i'nci birim için t zamanında tahmin edilen sağkalım olasılığı $\hat{S}(t \mid X_i)$ olarak gösterilsin. Bu durumda Brier skoru eşitlik-36'daki gibidir.

$$BS(t, \hat{S}) = E\left[I(\tilde{T}_i > t) - \hat{S}(t \mid X_i)^2 \right] \quad (36)$$

Burada beklenen değer, öğrenme setinde bulunmayan i'nci birimin verisine bağlı olarak hesaplanmaktadır. Brier skoru için kritik değerlerden ilki %33'tür. Bu U[0,1] dağılımından çekilen rasgele sayı ile tahmin edilen riske karşılık gelir. İkincisi %25 değeridir ve her birim için %50 risk tahminine karşılık gelmektedir. Bir diğer kriter ise tüm bağımsız değişkenlerin çıkarıldığı modelden elde edilen Brier skor değeridir (Ishwaran ve ark., 2008a).

Artık kareler ise eşitlik-37'de verilen sansürlenme ağırlıklarının ters olasılıkları kullanılarak ağırlıklandırılır.

$$\widehat{W}_i(t) = \frac{I(\tilde{T}_i \leq t)\Delta_i}{\hat{G}(\tilde{T}_i - \mid X_i)} + \frac{I(\tilde{T}_i > t)}{\hat{G}(t \mid X_i)} \quad (37)$$

Burada $\hat{G}(t \mid x) \approx P(C_i > t \mid X_i = x)$ i. birim için sansürlenme zamanlarının koşullu sağkalım fonksiyonunun tahminidir. Eğer bir bağımsız veri seti D_n mevcutsa beklenen Brier skoru eşitlik-38'deki gibidir.

$$\widehat{BS}(t,\hat{S}) = \frac{1}{n}\sum_{i\in D_n} \widehat{W}_i(t)\{I(\tilde{T}_i > t) - \hat{S}(t \mid X_i)\}^2 \quad (38)$$

Burada n, D_n'deki birim sayısıdır (i=1,….,n) ve $\hat{S}$ öğrenme verisinden hesaplanmaktadır. Eşitlik-38'deki ağırlıklar parametrik olmayan tahmin edici, Cox regresyon modeli veya toplamsal Aalen regresyon modeli ile optimal olarak tahmin edilebilirler (Mogensen ve ark.,2012).

i. Brier Skorunun Hesaplanmasında Kullanılan Çapraz Geçerlilik Yöntemleri

Modelin doğruluğunun test edildiği çapraz geçerlilik yönteminde veri başlangıçta test ve eğitim verisi olarak tesadüfi olarak ayrılır. Eğitim verisi modelin kurulumu (eğitimi) aşamasında kullanılır. Test verisi model kurulumunda kullanılmaz modelin doğruluğu bu yeni veri seti üzerinde test edilir (Kurtz, 1948; Moiser, 1951).

Basit çapraz geçerlilikte; veri setinin %5-%33'lük bir kısmı test verisi olarak ayrılmakta ve modelin öğrenme aşamasında bu kısım kullanılmamaktadır. Geriye kalan kısım üzerinde ise model kurulmakta ve gerçek değerler ile tahmin değerleri karşılaştırılarak modelin doğruluğu hesaplanmaktadır (Krus ve Fuller,1982).

Çift çapraz geçerlilik ve çoklu çapraz geçerlilik, veri setinin sınırlı sayıda olması durumunda kullanılmaktadır. Bu yöntemler, modelleme aşamasında bütün veri setinin kullanımına imkân sağlamaktadır. Veri seti, bir bölümü eğitim verisi diğer bölümü test verisi olmak üzere tesadüfî olarak iki eşit parçaya ayrılır ve bu şekilde bir doğruluk hesaplaması yapılır. Ardından test verisi ile eğitim verisi yer değiştirilerek yeniden bir doğruluk hesaplaması yapılır ve doğruluk değerlerinin ortalaması alınarak modelin doğruluk oranı hesaplanır. Bu algoritma çift çapraz geçerlilik olarak tanımlanmaktadır.

Çoklu çapraz geçerlilik yöntemi ise, her verinin bir kez eğitim bir kez de test verisi olarak kullanabileceği geçerlilik yöntemlerinin genellenmiş halidir. Bu yöntemde, veri n adet eşit gruba ayrılır. Bir grup test olarak geriye kalan (n-1) grup ise eğitim verisi olarak kullanılmaktadır ve bu durum n kez tekrar edilir. Yine doğruluk değerlerinin ortalaması alınarak modelin doğruluk oranı belirlenir (Efron ve Tibshirani, 1997).

Ağaç tabanlı yöntemlerin tahmin performansının ortaya konulması için çeşitli çapraz geçerlilik yöntemleri önerilmiştir. Bunlar aşağıdaki gibi sıralanabilir (Ishwaran ve ark., 2008a).

- **Görünen Tahmini (Apparent Estimate – AppErr)**

D_n veri setindeki n ($i=1,....,n$) birimden hesaplanan tahmin hatasının "görünen tahmini" eşitlik-39'daki gibidir.

$$AppErr\left(t,\hat{S}\right)=\frac{1}{n}\sum_{i\in D_n}\widehat{W}_i(t)\{I\left(\tilde{T}_i>t\right)-\hat{S}(t\mid X_i)\}^2 \qquad (39)$$

- **Bootstrap Çapraz Geçerlilik Tahmini (Bootstrap Cross-Validation Estimate-BootCvErr)**

Bootstrap çapraz geçerlilik yaklaşımı, D_n veri setini çok sayıda D_b öğrenme örneklemlerine ve bunlara karşılık gelen $D_n\backslash D_b$ test örneklemlerine ayırır. Burada B bootstrap örneklem sayısı olmak üzere b=1,..,B biçimindedir. Bootstrap örneklemleri orijinal veri setinden iadeli ya da iadesiz olarak seçilebilirler. Bootstrap öğrenme örneklemi kullanılarak model elde edilir. Karşılık gelen test örneklemleri ile artıklar elde edilir. Son olarak tahmin hatasının "bootstrap çapraz geçerlilik tahmini" bütün test setleri üzerinden ortalama alınarak eşitlik-40'daki hesaplanır.

$$BootCvErr\left(t,\hat{S}\right)=\frac{1}{B}\sum_{b=1}^{B}\frac{1}{Z_b}\sum_{i\in D_n\backslash D_b}\widehat{W}_i(t)\{I\left(\tilde{T}_i>t\right)-\hat{S}_b(t\mid X_i)\}^2 \qquad (40)$$

Bootstrap yöntemi yerine koyulmadan yapıldığında Z_b, n'den daha küçük olan kullanıcı tarafından belirlenen sabit bir sayıdır ve bootstrap örneklemlerinin büyüklüğünü vermektedir. Bootstrap yöntemi yerine konularak yapıldığında ise Z_b, D_b bootstrap örnekleminden çekilmeyen birimlerin sayısıdır. Yerine konularak elde edilen "bootstrap çapraz geçerlilik tahmin edicisi" eşitlik-41'de verilen toplamsal ifadenin sırasının değiştirilmesi ile elde edilir. Burada K_i, i'nci birim çıkartıldığında elde edilen bootstrap örneklemlerinin sayısıdır.

$$\frac{1}{n}\sum_{i=1}^{n}\frac{1}{K_i}\sum_{b:i\in D_n\backslash D_b}\widehat{W}_i(t)\{I\left(\tilde{T}_i>t\right)-\hat{S}_b(t\mid X_i)\}^2 \qquad (41)$$

- **k Katlı Çapraz Geçerlilik Tahmini (k Fold Cross Validation Estimate-CrossvalErr)**

"k katlı çapraz geçerlilik"te bootstrap çapraz geçerlilikten farklı olarak öğrenme örneklemlerinin sayısı k olarak alınır. D_n veri seti k alt kümeye $(D_j, j = 1, .., k)$ ayrılır. j'nci alt kümenin yer almadığı $D_n \backslash D_j$ veri seti ile $\hat{S}_j$'ler elde edilir. j'nci veri seti olan D_j ise test veri seti olarak kullanılır.

Tahmin hatasının çapraz geçerlilik tahmini eşitlik-42'deki gibidir.

$$crossvalErr(t, \hat{S}) = \frac{1}{k}\sum_{j=1}^{k}\sum_{i \in D_j} \widehat{W}_i(t)\{I(\tilde{T}_i > t) - \hat{S}_j(t \mid X_i)\}^2 \qquad (42)$$

- **Birini Dışarıda Bırak Tahmini (Leave-One-Out Cross Validation Estimate-loocvErr)**

Birini dışarıda bırak (leave-one-out) çapraz geçerlilik yaklaşımı k katlı çapraz geçerlilik yaklaşımına benzerdir. $D_n^{-i} = D_n \backslash \{(T_i, X_i)\}$ eğitim seti S_i'ler elde edilir ve $\{(T_i, X_i)\}$ üzerinde geçerliliği kontrol edilir. Tahmin hatasının birini dışarıda bırak (leave-one-out) tahmini eşitlik-43'deki gibidir.

$$loocvErr(t, \hat{S}) = \frac{1}{k}\sum_{i \in D_n} \widehat{W}_i(t)\{I(\tilde{T}_i > t) - \hat{S}_i(t \mid X_i)\}^2 \qquad (43)$$

- **Efron'un 0.632 Tahmini (Efron's 0.632 estimate-Boot632Err)**

Tahmin hatasının "bootstrap.632 tahmini", görünen tahmin hatası ve bootstrap çapraz geçerlilik tahmin hatasının ağırlıklı doğrusal kombinasyonudur ve eşitlik-44'de olduğu gibi elde edilir.

$$Boot632Err(t, \hat{S}) = (1 - 0{,}632) \cdot AppErr(t, \hat{S}) + 0{,}632 \cdot BootCvErr(t, \hat{S}) \qquad (44)$$

0,632 sabiti örneklem büyüklüğünden bağımsızdır ve i. birimin bootstrap örnekleminden iadeli olarak seçilme olasılığına karşılık gelir. Eşitlik-45'deki gibi ifade edilir.

$$P(\{(T_i, X_i)\} \in D_b) = 1 - \left(1 - {}^{1}/_{n}\right)^{n} \approx (1 - e^{-1}) \approx 0{,}632 \qquad (45)$$

- **Efron ve Tibshirani'nin 0.632+ tahmini(Efron&Tibshirani's 0.632+ Estimate-Boot632plusErr)**

Tahmin hatasının "bootstrap.632+ tahmini" görünen tahmin hatası, bootstrap çapraz geçerlilik tahmin hatası ve eşitlik-48'de verilen bilgisizlik tahmini (no information estimate) ile eşitlik-46'daki gibi elde edilir.

$$Boot632Err(t,\hat{S}) = \left(1 - \frac{0{,}632}{(1 - 0{,}368 \cdot \omega)}\right) \cdot AppErr(t,\hat{S}) + \frac{0{,}632}{(1 - 0{,}368 \cdot \omega)} \cdot BootCvErr(t,\hat{S}) \quad (46)$$

$$\omega = \frac{\min\left(BootCvErr(t,\hat{S}), NoInfErr(t,\hat{S})\right) - AppErr(t,\hat{S})}{NoInfErr(t,\hat{S}) - AppErr(t,\hat{S})} \quad (47)$$

$BootCvErr(t,\hat{S}) < AppErr(t,\hat{S})$ olduğu özel durumda $\omega = 0{,}632$ olarak alınır.

- **Bilgisizlik Tahmini (No Information Estimate- NoInfErr)**

"Bilgisizlik tahmini" bootstrap.632+ tahmini için gerekli olup eşitlik-48'deki gibi verilmiştir (Efron ve Tibshirani, 1997; Gerds ve Schumacher, 2007).

$$NoInfErr(t,\hat{S}) = \frac{1}{n^2}\sum_{j\in D_n}\sum_{i\in D_n}\widehat{W}_i(t)\{I(\tilde{T}_j > t) - \hat{S}_i(t \mid X_i)\}^2 \quad (48)$$

2.6.2. İntegrali Alınmış Brier Skoru (Integrated Brier Score- IBS)

Tahmin hataları IBS ile aşağıdaki gibi özetlenebilir:

$$IBS(TH,\tau) = \frac{1}{\tau}\int_{t=0}^{\tau} TH(t,\hat{S})dt \quad (49)$$

Burada *TH* yukarıda belirtilmiş herhangi bir yöntemle elde edilmiş tahmin hatasıdır.τ ise maksimum gözlem zamanıdır ($\tau > 0$) (Ishwaran ve ark., 2008a).

2.6.3. Harrel'in Uyum İndeksi (Concordance İndex - C indeks)

Uyum indeksi (C-indeks) rasgele seçilmiş bir birimde ilgilenilen olayı (ölüm) ilk kez yaşayan birimin daha kötü tahmin edilen sonuç değerine sahip olması olasılığını tahmin eder. C-indeks değeri 1'e ne kadar yakınsa model performansının o kadar iyi olduğu söylenebilir. C-indeksinin önemli bir özelliği sağkalım performansını ölçen diğer

indekslerden farklı olarak değerlendirmenin sabit tek bir zaman noktasına dayanmamasıdır (Hothorn ve ark., 2006a).

C-indeksi aşağıdaki adımlar takip edilerek hesaplanır.

Adım-1. Veri seti üzerindeki mümkün tüm birim çiftleri oluşturulur.

Adım-2. Bu çiftlerden daha kısa sağkalım zamanına karşılık gelen birim sansürlü ise çift ihmal edilir. Eğer çiftlerden her ikisi yaşıyor ve $T_i = T_j$ ise i, j çiftleri ihmal edilir. "İzin verilebilen", ihmal edilmeyen çiftlerin toplam sayısı olarak ifade edilebilir.

Adım-3. $T_i \neq T_j$ olduğunda her bir izin verilebilen çift için eğer daha kısa sağkalım zamanına sahip olan daha kötü tahmin sonuçlarına sahipse 1, eğer tahmin sonuçları eşit ise 0,5 değerini alır. Her bir izin verilebilir çift için $T_i = T_j$ olduğunda fakat her ikisinin birden ölü olmadığı durumda ölü olan daha kötü tahmin sonucuna sahipse 1, aksi halde 0,5 değerini alır. "Uyumlu", izin verilebilen tüm çiftlerin aldıkları değerlerin toplamını gösterir.

Adım-4. C-indeksi, $C = \frac{Uyumlu}{\text{İzin } Verilebilen}$ olarak tanımlanır.

3. GEREÇ ve YÖNTEM

Bu tez çalışmasında, sağkalım verilerinde kullanılan güncel ağaç tabanlı yöntemlerden RSO ile KÇO yöntemlerinin performanslarının karşılaştırılması amacıyla farklı senaryolar altında simülasyon çalışması yapılmıştır. Ayrıca RSO yönteminin kendi içinde farklı ayırma kriterlerinin kullanılması durumundaki performansı ve koşullu sansürlenme sağkalım fonksiyonunun tahmini için Aalen, Cox ve parametrik olmayan tahmincilerin performansı değerlendirilmiştir (Gerds ve Schumacher, 2007). Bunun için iki farklı senaryo ile veri türetimi yapılmıştır. İlk senaryoda orantısal hazard varsayımının gerçekleştiği durum dikkate alınırken, ikinci senaryoda orantısal hazard varsayımının sağlanmadığı durum dikkate alınmıştır (Ishwaran ve ark.,2010; Zhu ve Kosorok, 2012). Her iki senaryoda da her ayrımda rasgele seçilen bağımsız değişken sayısı kriteri, değişken sayısı p'nin karekökü olarak alınmıştır. Örneklem büyüklüğü 100,200 ve 300 olarak belirlenmiştir. Oluşturulan ağaç sayısı M=100, bootstrap tekrar sayısı (yerine koymadan) B=100 olarak alınmış olup, test seti (out of bag data) toplam örneklem büyüklüğünün %37'sini, eğitim seti (in bag data) ise %63'ünü kullanmaktadır. Her terminal düğümde yer alacak birim sayısı 6 ile sınırlandırılmıştır. Simülasyon çalışması 1000 tekrar ile yürütülmüştür.

Senaryo 1:

Bağımsız değişken sayısı p=25 olarak alınmıştır. $X = (X_1, \dots, X_{25})$, $\Sigma_{ij} = \rho^{|i-j|}$ (ρ=0,9) ve köşegen elemanlar 1 olmak üzere $\Sigma_{p\times p}$ varyans-kovaryans matrisli, ortalamalar vektörü $[0]_{p\times 1}$ olan çok değişkenli normal dağılımdan türetilmiştir. Sağkalım süreleri b_0=0,1 olmak üzere $\mu = b_0 \times \sum_{i=11}^{20} X_i$ ortalamalı üstel dağılımdan bağımsız olarak türetilmiştir. Sansürlenme zamanları $\mu/2$ ortalamalı üstel dağılımdan bağımsız olarak türetilmiştir. Durum değişkeni $\Delta = I(\tilde{T} \leq C)$ şeklinde elde edilmiştir. Bu senaryo için sansürlenme oranı yaklaşık %30 olarak elde edilmiştir.

Senaryo 2:

Bağımsız değişken sayısı p=25 olarak alınmıştır. $X = (X_1, \dots, X_{25})$, $\Sigma_{ij} = \rho^{|i-j|}$ (ρ=0,75) olmak üzere $\Sigma_{p\times p}$ varyans-kovaryans matrisli ortalamalar vektörü $[0]_{p\times 1}$ olan çok değişkenli normal dağılımdan türetilmiştir. Sağkalım süreleri ortalaması $\mu = 0{,}1 \times \left|\sum_{i=1}^{5} X_i\right| + 0{,}1 \times \left|\sum_{i=21}^{25} X_i\right|$ ortalamalı log normal dağılımdan bağımsız olarak türetilmiştir. Sansürlenme zamanları $\mu + 0{,}5$ ortalamalı log normal dağılımdan

türetilmiştir. Örneklem büyüklüğü 100, 200 ve 300 olarak alınmıştır. Durum değişkeni $\Delta = I(\tilde{T} \leq C)$ şeklinde elde edilmiştir. Bu senaryo için sansürlenme oranı yaklaşık olarak %30 olarak elde edilmiştir. Oluşturulan model performansları IBS ve C indeksi ile değerlendirilmiştir.

Simülasyon çalışmasında R 3.4.1 programında pec, party, randomForestSRC paketleri kullanılmıştır (Hothorn ve ark., 2005; Mogensen ve ark., 2012a; Ishwaran ve ark., 2008b).

4. BULGULAR

Sağkalım verilerinin analizinde kullanılan ağaç tabanlı orman yöntemleri olan RSO ve KÇO yöntemlerinin performanslarının karşılaştırılması için oransal hazard varsayımının sağlandığı durum olan senaryo 1 ve orantısal hazard varsayımının sağlanmadığı senaryo 2 dikkate alınarak yapılan simülasyon çalışması ile elde edilen bulgular tablolar (4.1- 4.12) aracılığıyla sunulmuştur. Sunulan tablolarda, çalışmada belirlenmiş olan üç farklı örneklem büyüklüğü (100, 200, 300) ile IPC ağırlıklarının hesaplanmasında kullanılan üç tahminci olan Aalen, Cox ve parametrik olmayan tahminciler için iki farklı ayırma kriterine sahip olan RSO yöntemi ve KÇO yönteminin C-indeksve IBS ölçütlerine ilişkin ortalama ve standart hata değerleri sunulmuştur.

Tablo 4-1. Örneklem büyüklüğünün n=100 olduğu ve oransal hazard varsayımının sağlandığı durumda farklı sağkalım zamanları için RSO ve KÇO yönteminin C-indeks ölçütüne göre ortalama ve standart hata değerleri

C-indeks (senaryo 1, n=100)	**Sağkalım zamanı**					
	0,5		**2**		**3,5**	
	$\bar{x}$	$s_{\bar{x}}$	$\bar{x}$	$s_{\bar{x}}$	$\bar{x}$	$s_{\bar{x}}$
RSO (logrank-parametrik olmayan)	0,9111	0,0007	0,8780	0,0010	0,8639	0,0016
RSO (logrank-Cox)	0,9133	0,0006	0,8808	0,0009	0,8652	0,0009
RSO (logrank-Aalen)	0,9202	0,0004	0,8887	0,0008	0,8689	0,0008
RSO (logrankskor-parametrik olmayan)	0,8849	0,0009	0,8477	0,0010	0,8165	0,0012
RSO (logrankskor-Cox)	0,8868	0,0008	0,8481	0,0009	0,8268	0,0011
RSO (logrankskor-Aalen)	0,8927	0,0007	0,8575	0,0008	0,8309	0,0010
KÇO (parametrik olmayan)	0,8333	0,0010	0,8205	0,0020	0,7931	0,0024
KÇO (Cox)	0,8319	0,0009	0,8249	0,0010	0,8126	0,0011
KÇO (Aalen)	0,8319	0,0009	0,8249	0,0010	0,8126	0,0011

Tablo 4-2. Örneklem büyüklüğünün n=200 olduğu ve oransal hazard varsayımının sağlandığı durumda farklı sağkalım zamanları için RSO ve KÇO yönteminin C-indeks ölçütüne göre ortalama ve standart hata değerleri

C-indeks (senaryo 1, n=200)	**Sağkalım zamanı**					
	0,5		**2**		**3,5**	
	$\bar{x}$	$s_{\bar{x}}$	$\bar{x}$	$s_{\bar{x}}$	$\bar{x}$	$s_{\bar{x}}$
RSO (logrank-parametrik olmayan)	0,9675	0,0008	0,8786	0,0009	0,8639	0,0016
RSO (logrank-Cox)	0,9678	0,0007	0,8908	0,0008	0,8652	0,0009
RSO (logrank-Aalen)	0,9680	0,0005	0,8987	0,0007	0,8689	0,0008
RSO (logrankskor-parametrik olmayan)	0,8850	0,0009	0,8475	0,0010	0,8170	0,0017
RSO (logrankskor-Cox)	0,8870	0,0008	0,8485	0,0009	0,8281	0,0011
RSO (logrankskor-Aalen)	0,8935	0,0007	0,8590	0,0008	0,8309	0,0010
KÇO (parametrik olmayan)	0,8689	0,0010	0,8249	0,0020	0,7931	0,0024
KÇO (Cox)	0,8689	0,0010	0,8596	0,0010	0,8126	0,0011
KÇO (Aalen)	0,8769	0,0009	0,8596	0,0010	0,8126	0,0011

Tablo 4-3. Örneklem büyüklüğünün n=300 olduğu ve oransal hazard varsayımının sağlandığı durumda farklı sağkalım zamanları için RSO ve KÇO yönteminin C-indeks ölçütüne göre ortalama ve standart hata değerleri

C-indeks (senaryo 1, n=300)	**Sağkalım zamanı**					
	0,5		**2**		**3,5**	
	$\bar{x}$	$s_{\bar{x}}$	$\bar{x}$	$s_{\bar{x}}$	$\bar{x}$	$s_{\bar{x}}$
RSO (logrank-parametrik olmayan)	0,9838	0,0009	0,8886	0,0010	0,8739	0,0012
RSO (logrank-Cox)	0,9857	0,0006	0,8908	0,0009	0,8752	0,0006
RSO (logrank-Aalen)	0,9869	0,0004	0,8990	0,0007	0,8789	0,0003
RSO (logrankskor-parametrik olmayan)	0,8950	0,0009	0,8475	0,0012	0,8276	0,0014
RSO (logrankskor-Cox)	0,8965	0,0006	0,8485	0,0010	0,8381	0,0011
RSO (logrankskor-Aalen)	0,8970	0,0005	0,8590	0,0008	0,8509	0,0010
KÇO (parametrik olmayan)	0,8879	0,0010	0,8249	0,0020	0,7931	0,0024
KÇO (Cox)	0,8889	0,0010	0,8596	0,0010	0,8125	0,0013
KÇO (Aalen)	0,8969	0,0009	0,8596	0,0010	0,8126	0,0011

Tablo 4-4. Örneklem büyüklüğünün n=100 olduğu ve orantısal hazard varsayımının sağlanmadığı durumda farklı sağkalım zamanları için RSO ve KÇO yönteminin C-indeks ölçütüne göre ortalama ve standart hata değerleri

C-indeks (senaryo 2, n=100)	**Sağkalım zamanı**					
	0,5		**2**		**3,5**	
	$\overline{x}$	$s_{\overline{x}}$	$\overline{x}$	$s_{\overline{x}}$	$\overline{x}$	$s_{\overline{x}}$
RSO (logrank-parametrik olmayan)	0,9802	0,0002	0,9434	0,0004	0,9076	0,0006
RSO (logrank-Cox)	0,9801	0,0002	0,9435	0,0004	0,9079	0,0006
RSO (logrank-Aalen)	0,9834	0,0001	0,9552	0,0002	0,9282	0,0002
RSO (logrankskor-parametrik olmayan)	0,9801	0,0002	0,8576	0,0015	0,8265	0,0009
RSO (logrankskor-Cox)	0,9799	0,0002	0,8581	0,0009	0,8368	0,0006
RSO (logrankskor-Aalen)	0,9825	0,0002	0,8775	0,0008	0,8409	0,0002
KÇO (parametrik olmayan)	0,9323	0,0012	0,8602	0,0010	0,8342	0,0011
KÇO (Cox)	0,9324	0,0012	0,8603	0,0010	0,8361	0,0009
KÇO (Aalen)	0,9336	0,0010	0,8605	0,0009	0,8398	0,0008

Tablo 4-5. Örneklem büyüklüğünün n=200 olduğu ve orantısal hazard varsayımının sağlanmadığı durumda farklı sağkalım zamanları için RSO ve KÇO yönteminin C-indeks ölçütüne göre ortalama ve standart hata değerleri

C-indeks (senaryo 2, n=200)	**Sağkalım zamanı**					
	0,5		**2**		**3,5**	
	$\bar{x}$	$s_{\bar{x}}$	$\bar{x}$	$s_{\bar{x}}$	$\bar{x}$	$s_{\bar{x}}$
RSO (logrank-parametrik olmayan)	0,9765	0,0009	0,9385	0,0005	0,9056	0,0010
RSO (logrank-Cox)	0,9789	0,0003	0,9436	0,0003	0,9083	0,0004
RSO (logrank-Aalen)	0,9795	0,0001	0,9462	0,0002	0,9152	0,0003
RSO (logrankskor-parametrik olmayan)	0,8950	0,0017	0,8675	0,0015	0,8275	0,0018
RSO (logrankskor-Cox)	0,8970	0,0005	0,8785	0,0007	0,8381	0,0010
RSO (logrankskor-Aalen)	0,8995	0,0003	0,8990	0,0005	0,8409	0,0007
KÇO (parametrik olmayan)	0,9015	0,0015	0,8194	0,0016	0,8002	0,0012
KÇO (Cox)	0,9028	0,0010	0,8291	0,0009	0,8013	0,0008
KÇO (Aalen)	0,9041	0,0002	0,8291	0,0009	0,8035	0,0005

Tablo 4-6. Örneklem büyüklüğünün n=300 olduğu ve orantısal hazard varsayımının sağlanmadığı durumda farklı sağkalım zamanları için RSO ve KÇO yönteminin C-indeks ölçütüne göre ortalama ve standart hata değerleri

C-indeks (senaryo 2, n=300)	**Sağkalım zamanı**					
	0,5		**2**		**3,5**	
	$\overline{x}$	$s_{\overline{x}}$	$\overline{x}$	$s_{\overline{x}}$	$\overline{x}$	$s_{\overline{x}}$
RSO (logrank-parametrik olmayan)	0,9948	0,0008	0,9100	0,0012	0,8839	0,0011
RSO (logrank-Cox)	0,9957	0,0005	0,9108	0,0010	0,8852	0,0006
RSO (logrank-Aalen)	0,9969	0,0003	0,9120	0,0004	0,8889	0,0002
RSO (logrankskor-parametrik olmayan)	0,8970	0,0009	0,8475	0,0010	0,8576	0,0012
RSO (logrankskor-Cox)	0,8985	0,0006	0,8585	0,0008	0,8581	0,0011
RSO (logrankskor-Aalen)	0,8990	0,0005	0,8690	0,0007	0,8609	0,0010
KÇO (parametrik olmayan)	0,8979	0,0013	0,8749	0,0020	0,7998	0,0022
KÇO (Cox)	0,8989	0,0010	0,8896	0,0010	0,8125	0,0009
KÇO (Aalen)	0,8989	0,0009	0,8896	0,0010	0,8126	0,0005

Tablo 4-7. Örneklem büyüklüğünün n=100 olduğu ve oransal hazard varsayımının sağlandığı durumda farklı sağkalım zamanları için RSO ve KÇO yönteminin IBS ölçütüne göre ortalama ve standart hata değerleri

IBS (senaryo 1, n=100)	**AppErr**		**BootCvErr**		**NoInfErr**		**Boot632plusErr**	
	$\overline{x}$	$s_{\overline{x}}$	$\overline{x}$	$s_{\overline{x}}$	$\overline{x}$	$s_{\overline{x}}$	$\overline{x}$	$s_{\overline{x}}$
RSO (logrank-parametrik olmayan)	0,0298	0,0056	0,1660	0,0360	0,2850	0,0060	0,1389	0,0279
RSO (logrank-Cox)	0,0292	0,0042	0,1646	0,0253	0,2814	0,0050	0,1384	0,0258
RSO (logrank-Aalen)	0,0286	0,0021	0,1630	0,0156	0,2787	0,0038	0,1372	0,0168
RSO (logrankskor-parametrik olmayan)	0,0300	0,0070	0,1745	0,0380	0,3050	0,0120	0,1439	0,0289
RSO (logrankskor-Cox)	0,0295	0,0068	0,1736	0,0293	0,3014	0,0090	0,1424	0,0280
RSO (logrankskor-Aalen)	0,0287	0,0035	0,1720	0,0166	0,2987	0,0058	0,1412	0,0267
KÇO (parametrik olmayan)	0,1010	0,0200	0,1534	0,0210	0,2576	0,0104	0,1386	0,0225
KÇO (Cox)	0,1007	0,0191	0,1512	0,0198	0,2399	0,0098	0,1384	0,0210
KÇO (Aalen)	0,0974	0,0120	0,1489	0,0127	0,2342	0,0090	0,1363	0,0133

AppErr: Görünen tahmin, BootCvErr: Boostrap Çapraz Geçerlilik Tahmini, NoInfErr: Bilgisizlik Tahmin hatası, Boot632plusErr: 0.632+ tahmini

Tablo 4-8. Örneklem büyüklüğünün n=200 olduğu ve oransal hazard varsayımının sağlandığı durumda farklı sağkalım zamanları için RSO ve KÇO yönteminin IBS ölçütüne göre ortalama ve standart hata değerleri

IBS (senaryo 1, n=200)	**AppErr**		**BootCvErr**		**NoInfErr**		**Boot632plusErr**	
	$\overline{x}$	$s_{\overline{x}}$	$\overline{x}$	$s_{\overline{x}}$	$\overline{x}$	$s_{\overline{x}}$	$\overline{x}$	$s_{\overline{x}}$
RSO (logrank-parametrik olmayan)	0,0288	0,0036	0,1640	0,0335	0,2845	0,0050	0,1379	0,0259
RSO (logrank-Cox)	0,0282	0,0022	0,1626	0,0233	0,2794	0,0045	0,1373	0,0238
RSO (logrank-Aalen)	0,0276	0,0019	0,1615	0,0145	0,2767	0,0028	0,1362	0,0148
RSO (logrankskor-parametrik olmayan)	0,0290	0,0070	0,1645	0,0367	0,3030	0,0110	0,1418	0,0260
RSO (logrankskor-Cox)	0,0285	0,0068	0,1636	0,0291	0,2914	0,0085	0,1412	0,0270
RSO (logrankskor-Aalen)	0,0277	0,0035	0,1620	0,0164	0,2867	0,0050	0,1409	0,0167
KÇO (parametrik olmayan)	0,1008	0,0197	0,1514	0,0187	0,2456	0,0094	0,1376	0,0215
KÇO (Cox)	0,0987	0,0172	0,1508	0,0166	0,2297	0,0066	0,1368	0,0200
KÇO (Aalen)	0,0961	0,0110	0,1469	0,0115	0,2210	0,0050	0,1338	0,0113

AppErr: Görünen tahmin, BootCvErr: Boostrap Çapraz Geçerlilik Tahmini, NoInfErr: Bilgisizlik Tahmin hatası, Boot632plusErr: 0.632+ tahmini

Tablo 4-9. Örneklem büyüklüğünün n=300 olduğu ve oransal hazard varsayımının sağlandığı durumda farklı sağkalım zamanları için RSO ve KÇO yönteminin IBS ölçütüne göre ortalama ve standart hata değerleri

IBS (senaryo 1, n=300)	**AppErr**		**BootCvErr**		**NoInfErr**		**Boot632plusErr**	
	$\overline{x}$	$s_{\overline{x}}$	$\overline{x}$	$s_{\overline{x}}$	$\overline{x}$	$s_{\overline{x}}$	$\overline{x}$	$s_{\overline{x}}$
RSO (logrank-parametrik olmayan)	0,0275	0,0032	0,1628	0,0325	0,2275	0,0045	0,1365	0,0247
RSO (logrank-Cox)	0,0271	0,0020	0,1616	0,0228	0,2283	0,0037	0,1355	0,0222
RSO (logrank-Aalen)	0,0265	0,0012	0,1609	0,0136	0,2247	0,0018	0,1320	0,0136
RSO (logrankskor-parametrik olmayan)	0,0285	0,0063	0,1635	0,0357	0,3020	0,0100	0,1417	0,0249
RSO (logrankskor-Cox)	0,0283	0,0064	0,1626	0,0271	0,2904	0,0075	0,1410	0,0260
RSO (logrankskor-Aalen)	0,0276	0,0027	0,1617	0,0154	0,2357	0,0040	0,1401	0,0147
KÇO (parametrik olmayan)	0,1006	0,0177	0,1513	0,0167	0,2454	0,0094	0,1366	0,0215
KÇO (Cox)	0,0977	0,0162	0,1504	0,0146	0,2294	0,0066	0,1358	0,0200
KÇO (Aalen)	0,0951	0,0106	0,1459	0,0105	0,2308	0,0050	0,1328	0,0113

AppErr: Görünen tahmin, BootCvErr: Boostrap Çapraz Geçerlilik Tahmini, NoInfErr: Bilgisizlik Tahmin hatası, Boot632plusErr: 0.632+ tahmini

Tablo 4-10. Örneklem büyüklüğünün n=100 olduğu ve orantısal hazard varsayımının sağlanmadığı durumda farklı sağkalım zamanları için RSO ve KÇO yönteminin IBS ölçütüne göre ortalama ve standart hata değerleri

IBS (senaryo 2, n=100)	**AppErr**		**BootCvErr**		**NoInfErr**		**Boot632plusErr**	
	$\overline{x}$	$s_{\overline{x}}$	$\overline{x}$	$s_{\overline{x}}$	$\overline{x}$	$s_{\overline{x}}$	$\overline{x}$	$s_{\overline{x}}$
RSO (logrank-parametrik olmayan)	0,0296	0,0056	0,1658	0,0350	0,2840	0,0054	0,1379	0,0269
RSO (logrank-Cox)	0,0290	0,0042	0,1642	0,0243	0,2804	0,0044	0,1374	0,0248
RSO (logrank-Aalen)	0,0285	0,0021	0,1627	0,0146	0,2777	0,0032	0,1362	0,0158
RSO (logrankskor-parametrik olmayan)	0,0297	0,0070	0,1743	0,0370	0,3040	0,0116	0,1429	0,0279
RSO (logrankskor-Cox)	0,0292	0,0068	0,1732	0,0283	0,3004	0,0087	0,1414	0,0270
RSO (logrankskor-Aalen)	0,0295	0,0035	0,1718	0,0156	0,2977	0,0054	0,1402	0,0257
KÇO (parametrik olmayan)	0,1007	0,0200	0,1531	0,0200	0,2566	0,0102	0,1376	0,0215
KÇO (Cox)	0,1005	0,0191	0,1508	0,0188	0,2389	0,0097	0,1374	0,0200
KÇO (Aalen)	0,0964	0,0120	0,1479	0,0117	0,2332	0,0087	0,1353	0,0123

AppErr: Görünen tahmin, BootCvErr: Boostrap Çapraz Geçerlilik Tahmini, NoInfErr: Bilgisizlik Tahmin hatası, Boot632plusErr: 0.632+ tahmini

Tablo 4-11. Örneklem büyüklüğünün n=200 olduğu ve orantısal hazard varsayımının sağlanmadığı durumda farklı sağkalım zamanları için RSO ve KÇO yönteminin IBS ölçütüne göre ortalama ve standart hata değerleri

IBS (senaryo 2, n=200)	**AppErr**		**BootCvErr**		**NoInfErr**		**Boot632plusErr**	
	$\overline{x}$	$s_{\overline{x}}$	$\overline{x}$	$s_{\overline{x}}$	$\overline{x}$	$s_{\overline{x}}$	$\overline{x}$	$s_{\overline{x}}$
RSO (logrank-parametrik olmayan)	0,0278	0,0026	0,1630	0,0325	0,2835	0,0040	0,1369	0,0249
RSO (logrank-Cox)	0,0272	0,0012	0,1616	0,0223	0,2784	0,0035	0,1363	0,0228
RSO (logrank-Aalen)	0,0266	0,0009	0,1605	0,0135	0,2757	0,0018	0,1352	0,0138
RSO (logrankskor-parametrik olmayan)	0,0280	0,0060	0,1635	0,0357	0,3020	0,0100	0,1408	0,0249
RSO (logrankskor-Cox)	0,0275	0,0058	0,1626	0,0281	0,2904	0,0075	0,1405	0,0260
RSO (logrankskor-Aalen)	0,0267	0,0025	0,1610	0,0154	0,2857	0,0040	0,1402	0,0157
KÇO (parametrik olmayan)	0,1007	0,0187	0,1504	0,0177	0,2446	0,0084	0,1366	0,0205
KÇO (Cox)	0,0977	0,0162	0,1506	0,0156	0,2287	0,0056	0,1358	0,0195
KÇO (Aalen)	0,0951	0,0100	0,1459	0,0105	0,2300	0,0043	0,1328	0,0108

AppErr: Görünen tahmin, BootCvErr: Boostrap Çapraz Geçerlilik Tahmini, NoInfErr: Bilgisizlik Tahmin hatası, Boot632plusErr: 0.632+ tahmini

Tablo 4-12. Örneklem büyüklüğünün n=300 olduğu ve orantısal hazard varsayımının sağlanmadığı durumda farklı sağkalım zamanları için RSO ve KÇO yönteminin IBS ölçütüne göre ortalama ve standart hata değerleri

IBS (senaryo 2, n=300)	**AppErr**		**BootCvErr**		**NoInfErr**		**Boot632plusErr**	
	$\overline{x}$	$s_{\overline{x}}$	$\overline{x}$	$s_{\overline{x}}$	$\overline{x}$	$s_{\overline{x}}$	$\overline{x}$	$s_{\overline{x}}$
RSO (logrank-parametrik olmayan)	0,0265	0,0022	0,1618	0,0315	0,2817	0,0035	0,1358	0,0237
RSO (logrank-Cox)	0,0261	0,0010	0,1606	0,0218	0,2773	0,0027	0,1352	0,0222
RSO (logrank-Aalen)	0,0255	0,0002	0,1608	0,0126	0,2737	0,0008	0,1347	0,0126
RSO (logrankskor-parametrik olmayan)	0,0275	0,0053	0,1625	0,0347	0,3010	0,0098	0,1407	0,0239
RSO (logrankskor-Cox)	0,0273	0,0054	0,1616	0,0261	0,2902	0,0065	0,1400	0,0250
RSO (logrankskor-Aalen)	0,0266	0,0017	0,1607	0,0144	0,2847	0,0030	0,1399	0,0137
KÇO (parametrik olmayan)	0,1005	0,0167	0,1503	0,0157	0,2444	0,0084	0,1356	0,0205
KÇO (Cox)	0,0967	0,0152	0,1501	0,0136	0,2305	0,0056	0,1348	0,0197
KÇO (Aalen)	0,0941	0,0104	0,1449	0,0102	0,2284	0,0040	0,1318	0,0103

AppErr: Görünen tahmin, BootCvErr: Boostrap Çapraz Geçerlilik Tahmini, NoInfErr: Bilgisizlik Tahmin hatası, Boot632plusErr: 0.632+ tahmini

5. TARTIŞMA ve SONUÇ

İlk ağaç tabanlı yöntemler tek bir ağaç yapısı oluşturmaktaydı ve elde edilen tahmin kurallarının yorumlanması kolaydı. Ancak elde edilen modellerin basitliği zayıf doğruluk tahminleri ile sonlanmaktaydı (Zhu ve Kosorok, 2012). Ağaç tabanlı yöntemlerde tahmin doğruluğunu arttırmak için ağaç topluluklarının kullanılması fikri ilk olarak Brieman (1996) ve Dietterich (2000a) tarafından ortaya konulmuş ve daha sonra Brieman (2001) tarafından önerilen rasgele ormanlar adı verilen yöntem ağaç toplulukları için genel bir çerçeve sağlayarak en popüler yöntem olmuştur.

Sağkalım verilerinde ağaç tabanlı yöntemlerin kullanılması ise ilk olarak ayırma kriteri olarak Kaplan-Meier eğrileri arasındaki uzaklık ölçütünü kullanan Gordon ve Olshen (1984) tarafından önerilmiştir. Daha sonra Ciampi ve arkadaşları (1986) olabilirlik-oranı istatistiğine dayalı ayırma kuralları oluşturmuştur ve sonrasında Segal (1988) parametrik olmayan düğümler içindeki homojenlik yerine düğümler arasındaki bölmeye dayanan logrank test istatistiğinin Harrington-Fleming sınıflamasını kullanmıştır. Sansürlü veriler için ağaç tabanlı yöntemlerin ağaç toplulukları yöntemlerinde kullanılmasına yönelik olarak düğümler için üstel log-olabilirliğe dayanan bir yöntem ve martingal artıklarının hata kareleri kaybı ile CART algoritmasında doğrudan kullanılan bir yöntem önerilmiştir (Davis ve Anderson, 1989; Therneau ve ark., 1990). Hothorn ve ark. (2004) bagging sağkalım ağaçlarını önermiş ve tek sağkalım ağacı kullanan modellerle karşılaştırmışlardır. Hothorn ve ark. (2006a) sansürlü ağırlıkların ters olasılığını kullanarak sağdan sansürlü veriler için sağkalım süresinin ağırlıklı tahminini veren KÇO yöntemini önermişlerdir.

Bu tez çalışmasında, sağdan sansürlü veriler için önerilen ampirik risk fonksiyonunu minimize etmeyi amaçlayan ve birbirinden farklı ağaçlar oluşturarak düşük korelasyon yapısında bir topluluk elde eden KÇO yöntemi (Hothorn ve ark., 2006a) ve sağdan sansürlü veriler için önerilen, Brieman'nın (2001) rasgele ormanlar yönteminin bir uzantısı olan RSO yöntemi (Ishwaran ve ark., 2008a) ele alınmış olup, yöntemlerin C-indeks ve IBS ölçütlerine göre karşılaştırılması amaçlanmıştır.

Çalışmada C-indeks ölçütüne göre; bütün durumlarda RSO yönteminin KÇO yöntemine göre daha yüksek C- indeks ortalama değerlerine ve daha düşük standart hata değerlerine sahip olduğu görülmüştür. Örneklem büyüklüğü bakımından incelendiğinde, örneklem büyüklüğündeki artış ile birlikte her iki senaryo ve her iki yöntem için C-indeks ortalama değerlerinin arttığı ve standart hata değerlerinin düştüğü gözlenmiştir. $\hat{G}$ tahmin edicileri olan parametrik olmayan tahmin edici, Cox tahmin edicisi ve Aalen tahmin edicisi

dikkate alındığında, RSO yönteminde Aalen tahmincisinin en iyi sonuç verdiği, parametrik olmayan tahmincinin ise Aalen ve Cox tahmincisinden daha düşük değerlere sahip olduğu gözlenmiştir. KÇO yönteminde ise parametrik olmayan tahmincinin daha düşük C-indeks ortalama değerlerine sahip olduğu, Cox ve Aalen tahmincisinde ise benzer düzeyde sonuçlar elde edildiği görülmüştür. RSO yöntemi kendi içinde kullanılan iki farklı ayırma kriteri bakımından incelendiğinde logrank ayrımının daha yüksek C-indeks ortalama değerlerine ve daha düşük standart hata değerlerine sahip olduğu saptanmıştır. Çalışmada incelenen orantısal hazard varsayımının sağlandığı durum ile orantısal hazard varsayımının sağlanmadığı durum karşılaştırıldığında, orantısal hazard varsayımının sağlanmadığı durumda her iki yönteminde daha iyi performans gösterdiği görülürken, RSO'da orantısal hazard varsayımının sağlandığı durumda KÇO'ya göre daha fazla bir düşüş olduğu gözlenmiştir.

Çalışmada IBS ölçütüne göre; bütün durumlar için her iki senaryoda ve bütün $\hat{G}$ tahmin yöntemleri için RSO yönteminin KÇO yönteminden daha düşük IBS ortalama ve standart hata değerleri verdiği görülmüştür. Örneklem büyüklüğünün artışı ile birlikte bütün durumlarda IBS ölçütüne göre model performansının arttığı gözlenmiştir. $\hat{G}$ tahmin edicileri olan parametrik olmayan tahmin edici, Cox tahmin edicisi ve Aalen tahmin edicisi dikkate alındığında, bütün yöntemler ve her iki senaryo için Aalen tahmincisinin daha düşük hata değeri verdiği görülmüştür. RSO kendi içinde incelendiğinde ise logrank ayrımının daha düşük IBS ortalama değerlerine ve standart hata değerlerine sahip olduğu saptanmıştır. Çalışmada incelenen orantısal hazard varsayımının sağlandığı durum ile sağlanmadığı durum karşılaştırıldığında, orantısal hazard varsayımının sağlanmadığı durumda bütün yöntemlerin daha iyi performans gösterdiği gözlenmiştir.

Mogensen ve ark. (2012) yaptıkları çalışmada pec paketinde yer alan COST veri setini kullanarak RSO, KÇO ve Cox regresyon modellerinin performanslarını IBS ölçütüne incelemişler ve bazı çapraz geçerlilik yöntemlerine göre yöntemlerin performanslarını benzer bulmuşken bazılarına göre RSO yönteminin performansını daha yüksek bulmuşlardır. Mogensen ve ark. (2012) yaptığı çalışma bir simülasyon çalışması olmamakla birlikte, bu tez çalışmasındaki simülasyon çalışmasına göre RSO yöntemi daha yüksek performans göstermiştir.

Gerds ve Schumacher (2007) IBS değerlerinin hesaplanmasında kullanılan $\hat{G}$ tahmincisi için marjinal Kaplan-Meier, Cox, Aalen ve parametrik olmayan tahmincileri kullanmışlardır. Ancak Kaplan-Meier tahmincisinin sansürlenme mekanizmasının ortak

değişkenlere bağlı olduğu durumlarda potansiyel bir hata vereceğini, bu nedenle sansürlü sağkalım fonksiyonunun ortak değişkenlere bağlı olduğu durum için diğer üç tahmin edicinin kullanılmasını önermişlerdir. Yaptıkları simülasyon çalışmasında ise Aalen tahmincisinin Cox tahmincisine göre daha iyi olduğunu belirtmişlerdir. Bu tez çalışmasındaki simülasyon sonuçlarına görede Aalen tahmincisinin her iki yöntemde de daha iyi performans gösterdiğine saptanmıştır.

1986 yılında Ciampi oluşturulan karar ağaçlarındaki iki çocuk düğümü karşılaştırmak için logrank test istatistiğinin kullanılmasını önermiştir. Ishwaran ve arkadaşları (2008a), 11 veri seti üzerinde RSO yöntemini farklı ayırma kurallarına göre uyguladıklarında logrank ayrımı kullanarak elde edilen modelin daha yüksek C-indeks değeri verdiğini belirtmişlerdir. Bu tez çalışmasında elde edilen simülasyon çalışması sonuçlarına göre RSO yönteminde orantısal hazard varsayımının sağlandığı ve sağlanmadığı durumda logrank ayrımının logrank skor ayrımına göre daha yüksek performans gösterdiğine saptanmıştır.

Sonuç olarak, RSO yönteminin KÇO'ya göre daha iyi performans gösterdiği ortaya konulmuştur. Her iki yöntem için Aalen tahmin edicisinin diğer tahmin edicilere göre daha iyi bir performans gösterdiği söylenebilir. Orantısal hazard varsayımının sağlanmadığı durumda her iki yöntemin de performansı daha iyi bulunmuştur. Ayrıca RSO yönteminin, kendi içinde kullanılan iki farklı ayırma kriterinden biri olan logrank ayrımının, logrank skor ayırma kriterine göre daha iyi performans gösterdiği görülmektedir.

KAYNAKLAR

1) Aalen OO (1982) Practical applications of the nonparametric statistical theory for counting processes. Preprint series Statistical Research Report http://urn nb no/URN: NBN: no-23420.
2) Adler W, Lausen B (2009) Bootstrap estimated true and false positive rates and ROC curve. Computational Statistics & Data Analysis 53**,** 718-729.
3) Akpınar H (2000) Veri tabanlarında bilgi keşfi ve veri madenciliği. İÜ İşletme Fakültesi Dergisi 29**,** 1-22.
4) Altman DG (1990) Practical statistics for medical research. CRC press.
5) Andersen MN, Andersen KK, Kammersgaard LP et al. (2005) Sex differences in stroke survival: 10-year follow-up of the Copenhagen stroke study cohort. Journal of Stroke and Cerebrovascular Diseases 14**,** 215-220.
6) Andersen PK, Borgan O, Gill RD et al. (2012) Statistical models based on counting processes. Springer Science & Business Media.
7) Breiman L (1994) Bagging predictors. TechnicalReport421, Department of Statistics. University of California, Berkeley.
8) Breiman L (1996) Bagging predictors. Machine learning 24**,** 123-140.
9) Breiman L (2001) Random forests. Machine learning 45**,** 5-32.
10) Brieman L, Friedman J, Olshen R et al. (1984) Classification and regression trees. Wadsworth & Brooks. Cole Advanced Books & Software.
11) Ciampi A, Thiffault J, Nakache J-P et al. (1986) Stratification by stepwise regression, correspondence analysis and recursive partition: a comparison of three methods of analysis for survival data with covariates. Computational statistics & data analysis 4**,** 185-204.
12) Cox D (1972) Regression models and life tables. JR Statist. Soc. B 34, 187-202. Cox18734J. R. Statist Soc B 1972.
13) Davis RB, Anderson JR (1989) Exponential survival trees. Statistics in Medicine 8**,** 947-961.
14) Dietterich TG (2000) Ensemble methods in machine learning. In: International workshop on multiple classifier systems. Springer. pp. 1-15.
15) Efron B, Tibshirani R (1997) Improvements on cross-validation: the 632+ bootstrap method. Journal of the American Statistical Association 92**,** 548-560.
16) Gerds TA, Kattan MW, Schumacher M et al. (2013) Estimating a time-dependent concordance index for survival prediction models with covariate dependent censoring. Statistics in Medicine 32**,** 2173-2184.
17) Gerds TA, Schumacher M (2007) Consistent estimation of the expected Brier score in general survival models with right-censored event times. Biometrical Journal 48**,** 1029-1040.
18) Gordon L, Olshen R (1985) Tree-structured survival analysis. Cancer treatment reports 69**,** 1065-1069.
19) Graf E, Schmoor C, Sauerbrei W et al. (1999) Assessment and comparison of prognostic classification schemes for survival data. Statistics in medicine 18**,** 2529-2545.
20) Ho TK (1995) Random decision forests. In: Document Analysis and Recognition, 1995, Proceedings of the Third International Conference on. IEEE. pp. 278-282.
21) Ho TK (1998) The random subspace method for constructing decision forests. IEEE transactions on pattern analysis and machine intelligence 20**,** 832-844.
22) Hothorn T, Bühlmann P, Dudoit S et al. (2006a) Survival ensembles. Biostatistics 7**,** 355-373.

23) Hothorn T, Hornik K, Zeileis A (2006b) Unbiased recursive partitioning: A conditional inference framework. Journal of Computational and Graphical statistics 15**,** 651-674.
24) Hothorn T, Lausen B (2003) On the exact distribution of maximally selected rank statistics. Computational Statistics & Data Analysis 43**,** 121-137.
25) Hothorn T, Lausen B, Benner A et al. (2004) Bagging survival trees. Statistics in medicine 23**,** 77-91.
26) Hothorn T, Hornik K, Zeileis A (2005) party: A Laboratory for Recursive Part(y)itioning. R package version 0.2-8, http://CRAN.R-project.org.
27) Ishwaran H, Blackstone EH, Pothier CE et al. (2004) Relative risk forests for exercise heart rate recovery as a predictor of mortality. Journal of the American Statistical Association 99**,** 591-600.
28) Ishwaran H, Kogalur UB, Blackstone EH et al. (2008a) Random survival forests. The annals of applied statistics**,** 841-860.
29) Ishwaran H, Kogalur UB, Blackstone EH et al. (2008b) RandomForestSRC:Random Forests for Survival, Regression and Classification. R package version 2.4.1, http://CRAN.R-project.org.
30) Kaplan EL, Meier P (1958) Nonparametric estimation from incomplete observations. Journal of the American statistical association 53**,** 457-481.
31) Kass GV (1980) An exploratory technique for investigating large quantities of categorical data. Applied statistics**,** 119-127.
32) Keles S, Van Der Laan M, Dudoit S (2004) Asymptotically optimal model selection method with right censored outcomes. Bernoulli**,** 1011-1037.
33) Kim H, Loh W-Y (2003) Classification trees with bivariate linear discriminant node models. Journal of Computational and Graphical Statistics 12**,** 512-530.
34) Kim H, Loh W-Y (2011) Classification trees with unbiased multiway splits. Journal of the American Statistical Association.
35) Kim JS, Dailey RJ (2008) Biostatistics for oral healthcare. John Wiley & Sons.
36) Klein JP, Moeschberger ML (2005) Survival analysis: techniques for censored and truncated data. Springer Science & Business Media.
37) Kleinbaum DG, Klein M (2006) Survival analysis: a self-learning text. Springer Science & Business Media.
38) Krus DJ, Fuller EA (1982) Computer-assisted multicross-validation in regression analysis, Educational and Psychological Measurement, 42, 187-193.
39) Kurtz AK (1948) A research test of Rorschach test. Personnel Psychology, 1, 41-53.
40) LeBlanc M, Crowley J (1992) Relative risk trees for censored survival data. Biometrics**,** 411-425.
41) LeBlanc M, Crowley J (1993) Survival trees by goodness of split. Journal of the American Statistical Association 88**,** 457-467.
42) Lee ET, Wang J (2003) Statistical methods for survival data analysis. John Wiley & Sons.
43) Lim T-S, Loh W-Y, Shih Y-S (1998) An empirical comparison of decision trees and other classification methods.
44) Loh W-Y (2002) Regression tress with unbiased variable selection and interaction detection. Statistica Sinica**,** 361-386.
45) Loh W-Y, Shih Y-S (1997) Split selection methods for classification trees. Statistica sinica**,** 815-840.
46) Mantel N, Haenszel W (1959) Statistical aspects of the analysis of data from retrospective studies of disease.

47) Mingers J (1987a) Expert systems—rule induction with statistical data. Journal of the operational research society 38**,** 39-47.
48) Mingers J (1987b) Rule induction with statistical data—a comparison with multiple regression. Journal of the Operational Research Society 38**,** 347-351.
49) Mogensen UB, Ishwaran H, Gerds TA (2012) Evaluating random forests for survival analysis using prediction error curves. Journal of statistical software 50**,** 1.
50) Mogensen UB, Ishwaran H, Gerds TA (2012a) pec: Prediction Error Curves for Risk Prediction Models in Survival Analysis. R package version 2.4.9, http://CRAN.R-project.org.
51) Mosier (1951) The need and means of cross validation. Problems and designs of cross-validation, Educational and Psychological Measurement.
52) Molinaro AM, Dudoit S, Van der Laan MJ (2004) Tree-based multivariate regression and density estimation with right-censored data. Journal of Multivariate Analysis 90**,** 154-177.
53) Morgan JN, Sonquist JA (1963) Problems in the analysis of survey data, and a proposal. Journal of the American statistical association 58**,** 415-434.
54) Safavian SR, Landgrebe D (1990) A survey of decision tree classifier methodology.
55) Segal MR (1988) Regression trees for censored data. Biometrics**,** 35-47.
56) Stevenson M (2009) An Introduction to survival analysis. EpiCentre, IVABS, Massey University. Unpublished manuscript.
57) Team RC (2014) R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. 2013. ISBN 3-900051-07-0.
58) Therneau TM, Atkinson EJ (1997) An introduction to recursive partitioning using the RPART routines. Technical Report 61. URL http://www. mayo. edu/hsr/techrpt/61. pdf.
59) Therneau TM, Grambsch PM (2000) Modeling survival data: extending the Cox model. Springer Science & Business Media.
60) Therneau TM, Grambsch PM, Fleming TR (1990) Martingale-based residuals for survival models. Biometrika 77**,** 147-160.
61) Van Belle G, Fisher LD, Heagerty PJ et al. (2004) Biostatistics: a methodology for the health sciences. John Wiley & Sons.
62) Van Der Laan MJ, Dudoit S (2003) Unified cross-validation methodology for selection among estimators and a general cross-validated adaptive epsilon-net estimator: Finite sample oracle inequalities and examples.
63) Van der Laan MJ, Robins JM (2003) Unified methods for censored longitudinal data and causality. Springer Science & Business Media.
64) Wei L J (1992) The accelerated failure time model: a useful alternative to the Cox regression model in survival analysis. Statistics in medicine, 11(14-15), 1871-1879.
65) Zhao G (2008) Nonparametric and parametric survival analysis of censored data with possible violation of method assumptions. ProQuest.
66) Zhu R, Kosorok MR (2012) Recursively imputed survival trees. Journal of the American Statistical Association 107**,** 331-340.

SİMGELER ve KISALTMALAR

$a_i(\boldsymbol{x})$	:	i'nci birim için tahmin ağırlığı
$AppErr(t, \hat{S})$	:	Tahmin hatasının görünen tahmini
$BS(t, \hat{S})$	:	Brier Skoru
$BootCvErr(t, \hat{S})$	:	Bootstrap çapraz geçerlilik tahmini
$Boot632Err(t, \hat{S})$	:	Tahmin hatasının bootstrap.632 tahmini
C	:	Sansürlenme zamanı
C_{index}	:	Uyum indeksi
$c_{max}(\boldsymbol{t}, \mu, \Sigma)$	:	Gözlenen çok değişkenli doğrusal istatistik olan $t \in \mathbb{R}^{pq}$ 'yu gerçel doğruya haritalayan tek değişkenli test istatistiği
$crossvalErr(t, \hat{S})$	:	Tahmin hatasının çapraz geçerlilik tahmini
Δ	:	Durum değişkeni
$\Delta_i(t)$	:	t zamanı için i'nci birimin gerçek durumu
$\mathbb{E}$	:	Beklenen değer
$\mathcal{F}_{\boldsymbol{T\|X}}$	:	Bağımlı değişkenin koşullu dağılımı
ψ	:	Aday tahminciler
Ψ	:	Aday tahmincilerin fonksiyon uzayı
D_n	:	Brier skorun hesaplanmasında kullanılan n birimlik veri seti
D_b	:	D_n veri setinden elde edilen b'inci bootstrap örneklemi(öğrenme örneklemi)
$D_n \backslash D_b$	:	D_n veri setinden D_b veri seti çıkartıldığında elde edilen veri seti
g_j	:	X_j ortak değişkeninin rasgele olmayan dönüşümü
$\hat{G}$	:	Koşullu sansürlü sağkalım fonksiyonu tahmincisi
$G(c \mid \mathbf{X})$	:	Koşullu sansürleme sağkalım fonksiyonu
$\hat{G}_{Nonpar}$	:	Parametrik olmayan tergresyon tahmin edici
$\hat{G}_{Cox}$	:	Cox regresyon tahmin edicisi
$\hat{G}_{Aalen}$	:	Nelson Aalen tahmin edicisi
h	:	Etki fonksiyonu
$h(t)$	:	Hazard fonksiyonu
$H(t)$	:	Kümülatif hazard fonksiyonu
$H_0(t)$	:	Başlangıç kümülatif hazard fonksiyonu
$H_{\mathrm{e}}(t\|\mathrm{x})$	:	Topluluk koşullu kümülatif hazard fonksiyonu
$H_{\mathrm{e}}^{*}(t\|\mathrm{x})$	:	OOB için topluluk kümülatif hazard fonksiyonu
$\widehat{H}_m(t \mid x)$	:	Nelson-Aalen tahmin edicisi ile koşullu kümülatif hazard fonksiyonu
Σ	:	Varyans-Kovaryans matrisi
$\otimes$	:	Kroneker çarpım
$\mathcal{L}$	:	Öğrenme örneklemi
$L_{tam}(T, \psi(\mathbf{X}))$	:	Tam veri kayıp fonksiyonu
$L = (T, \psi(\mathbf{X}) \mid \eta)$	:	Gözlenen veri kayıp fonksiyonu
$LogRank(X_j, c)$	:	log-rank test istatistiği
$LogRankskor(X_j, c)$	:	log-rank skor istatistiği
$loocvErr(t, \hat{S})$	:	Tahmin hatasının birini dışarıda bırak (leave-one-out) tahmini
$NoInfErr(t, \hat{S})$	:	Bilgisizlik tahmini (no information estimate)

OOB	:	Out of Bag data
P_j	:	H_0^j hipotezi için p değeri
s_α^2	:	Rankların örneklem varyansı
$S(t)$	:	Sağkalım fonksiyonu
$\hat{S}(t \mid X_i)$	:	i'nci birim için t zamanında tahmin edilen sağkalım olasılığı
$\hat{S}^{cforest}$	:	Koşullu çıkarsama ormanları için topluluk sağkalım fonksiyonu
$\hat{S}^{rsf}$	:	Rasgele sağkalım ağaçları için topluluk sağkalım fonksiyonu
$\tilde{T}$	:	Gerçek ölüm zamanı
$\mathcal{T}_m$	:	m'nci sağkalım ağacı
$\mathcal{T}_m(\boldsymbol{x})$	:	$\boldsymbol{x}$ ortak değişken değeri ile m'nci ağaçtaki terminal düğümü
$\boldsymbol{T}_j(\mathcal{L}, w)$	:	$\boldsymbol{T}$ ile X_j , $(j = 1, \dots, p)$ arasındaki ilişkiyi ölçen doğrusal test istatistiği
$\boldsymbol{T}_{j^*}^A(\mathcal{L}, w)$	:	X_j^* 'ın mümkün tüm A alt gruplarının test istatistiği
$\boldsymbol{w}$	:	Birim ağırlıkları vektörü
$w_.$	:	Birim ağırlıklarının toplamı
$\boldsymbol{w}_{sağ}$	:	Sağ birim ağırlığı
$\boldsymbol{w}_{sol}$	:	Sol birim ağırlığı
$\boldsymbol{X}$	:	Ortak değişkenler matrisi
$\boldsymbol{X}_{j^*}$	:	T ile en güçlü ilişkiye sahip j^*'ncı ortak değişken
T	:	Bağımlı değişken
T_i	:	i'nci birimin sağkalım zamanı

TEŞEKKÜR

Yüksek lisans eğitimim boyunca ve tezimi gerçekleştirmem sırasında sonsuz özverisini, desteğini ve sabrını benden hiç esirgemeyen değerli danışmanım Doç. Dr. Deniz SIĞIRLI'ya bilimsel gelişimime verdiği emek ve katkılarından dolayı sonsuz teşekkürlerimi sunarım.

Yüksek lisans öğrenimim boyunca eğitimime katkıda bulunan değerli hocam Prof. Dr. İlker ERCAN' a ve anabilim dalımızdaki öğretim üyelerine teşekkürlerimi sunarım.

Ayrıca yüksek lisans eğitimim boyunca ve tezimi gerçekleştirmem sırasında her zaman yanımda hissettiğim aileme maddi ve manevi desteklerinden dolayı teşekkürlerimi sunarım.

ÖZGEÇMİŞ

29 Ağustos 1991 tarihinde Gemlik/Bursa'da doğdum. İlk ve Orta öğrenimimi Şehit Cemal İlköğretim Okulu'nda bitirdim. Lise öğrenimimi Gemlik Lisesi'nde tamamladım. 2009 yılında Yıldız Teknik Üniversitesi Fen – Edebiyat Fakültesi İstatistik bölümünü kazandım ve aynı yıl İngilizce hazırlık okudum. 2010 yılında İngilizce hazırlık dönemini tamamladım ve Lisans eğitimime başladım. 2014 yılında Yıldız Teknik Üniversitesi Fen – Edebiyat Fakültesi İstatistik bölümünde lisans eğitimimi Onur Öğrencisi olarak tamamladım. 2015 yılında Uludağ Üniversitesi Tıp Fakültesi Biyoistatistik Anabilim Dalında yüksek lisans eğitimime başladım.

ULUDAĞ ÜNİVERSİTESİ
TEZ ÇOĞALTMA VE ELEKTRONİK YAYIMLAMA İZİN FORMU

Yazar Adı Soyadı	Ayşegül YABACI
Tez Adı	Sağkalım Verilerinde Kullanılan Ağaç Tabanlı Yöntemlerin Karşılaştırılması
Enstitü	Sağlık Bilimleri Enstitüsü
Anabilim Dalı	Biyoistatistik Anabilimdalı
Bilim Dalı	
Tez Türü	Yüksek Lisans Tezi
Tez Danışman(lar)ı	Doç. Dr. Deniz SIĞIRLI
Çoğaltma (Fotokopi Çekim) İzni	☐ Tezimden fotokopi çekilmesine izin veriyorum ☐ Tezimin sadece içindekiler, özet, kaynakça ve içeriğinin % 10 bölümünün fotokopi çekilmesine izin veriyorum ☐ Tezimden fotokopi çekilmesine izin vermiyorum
Yayımlama İzni	☐ Tezimin elektronik ortamda yayımlanmasına izin veriyorum ☐ Tezimin elektronik ortamda yayımlanmasının ertelenmesini istiyorum 1 yıl ☐ 2 yıl ☐ 3 yıl ☐ ☐ Tezimin elektronik ortamda yayımlanmasına izin vermiyorum

Hazırlamış olduğum tezimin yukarıda belirttiğim hususlar dikkate alınarak, fikri mülkiyet haklarım saklı kalmak üzere Uludağ Üniversitesi Kütüphane ve Dokümantasyon Daire Başkanlığı tarafından hizmete sunulmasına izin verdiğimi beyan ederim.

Tarih: 19/01/2017
İmza: