T.C.
EGE ÜNİVERSİTESİ
Sağlık Bilimleri Enstitüsü

SAĞLIK VERİLERİNDE VERİ MADENCİLİĞİ TEKNİKLERİ İLE SAĞKALIMI ETKİLEYEN FAKTÖRLERİN SEÇİMİ, PERFORMANSLARININ DEĞERLENDİRİLMESİ

Yüksek Lisans Tezi

Büşra Ecem GÜNAYDIN

Biyoistatistik ve Tıbbi Bilişim Anabilim Dalı

İzmir

2022

T.C.

EGE ÜNİVERSİTESİ SAĞLIK BİLİMLERİ ENSTİTÜSÜ

SAĞLIK VERİLERİNDE VERİ MADENCİLİĞİ TEKNİKLERİ İLE SAĞKALIMI ETKİLEYEN FAKTÖRLERİN SEÇİMİ, PERFORMANSLARININ DEĞERLENDİRİLMESİ

Büşra Ecem GÜNAYDIN

Danışman

Prof. Dr. Soner DUMAN

Biyoistatistik ve Tıbbi Bilişim Anabilim Dalı

Biyoistatistik Yüksek Lisans Programı

İzmir

2022

TEZ ONAY SAYFASI

Kurum Adı **:**Ege Üniversitesi

Anabilim Dalı **:**Biyoistatistik ve Tıbbi Bilişim Anabilim Dalı

Program **:** Biyoistatistik Yüksek Lisans Programı

Tez Konusu **:** Sağlık Verilerinde Veri Madenciliği Teknikleri İle Sağkalımı Etkileyen Faktörlerin Seçimi, Performanslarının Değerlendirilmesi

Danışman **:** Prof. Dr. Soner DUMAN

Tezi Hazırlayan **:** Büşra Ecem Günaydın

Değerlendirme Kurulu Üyeleri Adı Soyadı

Başkan(Danışman) : Prof. Dr. Soner DUMAN

Üye / İmza : Prof.Dr. Mehmet Nurullah ORMAN

Üye / İmza : Prof.Dr. Pembe KESKİNOĞLU

Tezin Kabul Edildiği Tarih :

Önsöz

"Sağlık Verilerinde Veri Madenciliği Teknikleri İle Sağkalımı Etkileyen Faktörlerin Seçimi, Performanslarının Değerlendirilmesi" çalışmamda, sağkalım analizinde kullanılan makine öğrenim yöntemlerinden birisi olan Rastgele sağkalım orman yöntemi ve geleneksel istatistiksel yaklaşımlardan olan Cox regresyon yöntemi üzerine çalıştım.

Tüm öğrenim hayatımda her zaman arkamda olan, desteklerini hiçbir zaman esirgemeyen canım babam Muharrem Günaydın ve canım annem Şen Günaydın'a ve canım kardeşim Dilara Günaydın'a sonsuz teşekkürlerimi borç bilirim.

Aynı zamanda araştırma süresince üstümden birçok yükü alıp bana destek olan sevgili canım eşim Uğur Şakar'a da teşekkürü borç bilirim.

Ege Üniversitesi Biyostatistik Bölümü hocalarıma ve Sayın Prof. Dr. Pembe Keskinoğlu'na katılarından dolayı teşekkür ederim.

İzmir, 2022 **Büşra Ecem GÜNAYDIN**

Özet

Sağlık Verilerinde Veri Madenciliği Teknikleri İle Sağkalımı Etkileyen Faktörlerin Seçimi, Performanslarının Değerlendirilmesi

GÜNAYDIN, Büşra Ecem

Yüksek Lisans Tezi, Biyoistatistik ve Tıbbi Bilişim Anabilim Dalı

Tez Danışmanı: Prof. Dr. Soner DUMAN

2022

Giriş ve Amaç: Bu çalışmada, ağaç tabanlı veri madenciliği yöntemi Rastgele Sağkalım Orman'ın araştırılması ve tıp alanındaa toplanan veri üzerinde analizler yaparak sonuçların incelenmesi amaçlanmıştır. **Yöntem:** Çalışmada Mendeley Data açık kaynak platformunda yayınlanan Awodutire, Kolawole, Ilori tarafından 2017 yılındaki çalışmalarında uygulanmış, Osogbo'daki Ladoke Akintola Teknoloji Eğitim Hastanesi'ndeki meme kanseri hastasından alınan klinik verilerle Cox Regresyon ve Rastgele Sağkalım Orman yöntemleri ile analizler yapılmıştır. Analizlerde RStudio Version 1.4.1717 programı kullanılmıştır. **Bulgular:** Çalışmada 89 hasta vardır. Hastaların yaş ortalaması 50.29±10.848, ortalama menarş yaşı 15.72±2.326 ve ortalama emzirme yılı 1.39±0.547'dir. Hastalara ait sağkalım sürelerinin %42.7'si (n=38) sağdan sansürlüdür. Hastaların %53.9'u (n=48) doğum kontrol hapı kullanmakta, teşhis anında %67.4'ünün (n=60) erken evre(Evre I. Ve II), %32.6'sının (n=29) geç evredir (III.Evre ve IV.Evre). Tedavi sırasında %43.8'ine (n=39) neoadjuvant uygulanmıştır. Hastaların medyan sağkalım süresi 251± 52.82 [%95 G.A. 147.5-354.5] gün olarak hesaplanmıştır. Cox regresyon yönteminde tedavi süresinde neoadjuvan kullanımının tehlike oranı üzerindeki etkisi istatistiksel olarak anlamlıdır ($p<0.05$). Cox regresyon modeli uyum indeksi olan C-Index 0.648 olarak hesaplandı. Rastgele Sağkalım orman yönteminde veri seti %70 test, %30 analiz dışı tutulmuş, out-of-box veri olarak ayrılmıştır. Buna göre önemli değişkenler ortalama emzirme yılı ve doğum kontrol hapı kullanımı olarak belirlenmiştir. Rastgele sağkalım yöntemine ait C-index 0.60 olarak hesaplanmıştır. **Sonuç:** Her iki yöntemde benzer performans göstermiştir, Örneklem ve değişken sayısı arttırılarak analiz tekrarlanmalıdır.

Anahtar Kelimeler: Sağkalım Analizi; Veri Madenciliği; Rastgele Sağkalım Orman

Abstract

Selection Of Factors Affecting Survival And Performance Evaluation With Data Mining Techniques İn Health Data

Günaydın, Büşra Ecem

MSc Thesis, Department of Biostatistics and Medical Informatics

Supervisor: Prof. Dr. Soner DUMAN

2021

Introduction and Aim: In this study, it is aimed to examine the Random Survival Forests (RSF) method, which is one of the tree-based data mining methods, and to discuss the results by applying it to a data set obtained from the health field. **Methods:** In the study, analyzes were made by Cox Regression and Random Survival Forest methods with clinical data obtained from a breast cancer patient at Ladoke Akintola Technology Teaching Hospital in Osogbo, which was applied by Awodutire, Kolawale, Ilori, published on the Mendeley Data open source platform, in their study in 2017. RStudio Version 1.4.1717 program was used in the analysis. **Results:** There were 89 patients in the study. The mean age of the patients was 50.29±10.848, the mean age at menarche was 15.72±2.326, and the mean breastfeeding year was 1.39±0.547. 42.7% (n=38) of the patients' survival times were right-censored. 53.9% (n=48) of the patients were using birth control pills, at the time of diagnosis, 67.4% (n=60) were in the early stage (Stage I and II), and 32.6% (n=29) was in the late stage (III. Stage and Stage IV). Neoadjuvant was applied to 43.8% (n=39) of them during treatment. Median survival of patients 251±52.82 [95% G.A. 147.5-354.5] days. In the Cox regression method, the effect of neoadjuvant use on the hazard ratio was statistically significant ($p<0.05$). The C-Index, which is the Cox regression model fit index, was calculated as 0.648. In the Random Survival forest method, the data set was divided as 70% tested, 30% excluded, out-of-box data. Accordingly, the important variables were determined as the mean year of breastfeeding and the use of birth control pills. The C-index of the random survival method was calculated as 0.60. **Conclusion:** Both methods showed similar performance. The analysis should be repeated by increasing the number of samples and variables.

Keywords: Survival Analysis; Data Mining; Random Survival Forest

İçindekiler

Tablolar Listesi

Şekiller Listesi

Giriş

Günümüzde artık çoğu veri dijital ortamlarda kaydedilmeye başlamıştır. Kaydedilen bilgiler her geçen gün daha da artmakta, onların kaydedildiği veri tabanları da katlanarak artmaktadır. Bu veri tabanlarında kaydedilen veriler, bir yığın olarak örneklendirilirse, bu büyük veri yığını tek başına değersiz olacaktır. Fakat bu veri yığını, bir amaç için düzenli olarak işlenir ve analizleri yapılırsa, artık değersiz olarak görünmez ve belirli bir amaca yönelik anlamlı ve değerli bilgilere ulaşılabiliyor olacaktır (Özekes, 2003).

Sağlık sistemi yönetiminin ve kararlarının altında, veri ve ondan çıkarılmış bilgiler yatar. Sağlık yönetiminin ve kararlarının amaca uygun ve etkili olması için güncel, doğru ve güvenilir veriler elde edilmesi gerekmektedir (Koyuncugil & Özgülbaş, 2009). Sağlık bilgi sistemlerinin hedefi yüksek düzeydeki sağlık verilerinden amaca yönelik anlamlı bilgi üretmektir. Üretilen bu bilgiler ile sağlık hizmetinin kalitesi arttırmaya, sağlık kurumlarının yönetimlerinin iyileştirilmesinde, sağlık sisteminin kaynaklarının daha verimli kullanılmasında ve sağlık politikalarının oluşturulmasında kullanılmaktadır.

Karar verici, çalışmasının temel amacına uygun ve etkili olabilecek bir karar elde etmek için birçok veriyi analiz etmek ve birçok değişkeni göz önüne almalıdır. Günümüzde dijital verilerin hacmindeki artış ile birlikte bu çok büyük, birden fazla boyuta sahip ve oldukça karmaşık verileri incelemek için yöntemler ya da yeni çalışma sistemleri geliştirmek gerekmiştir. Aynı zamanda bu sistemler için metotlar, politikalar ya da sistem altyapıları geliştirmek amacıyla verilerin nasıl kullanılması gerektiği ve veri güvenliği ile ilgili yeni modellere ulaşmak da araştırmacının karşılaştığı en büyük sorunlardır (Koyuncugil & Özgülbaş, 2009). Hastalıkların hem tıbbi anlamda hem de sağlık yönetimi açısından doğru yönetilmesi gerekmektedir. Bu ihtiyaçlar ile birlikte veri madenciliği yöntemleri ortaya çıkmıştır. Veri madenciliği yöntemleri kullanılarak, çok boyutlu verilerden amaca uygun ve önemli bilgiler daha kolay ve hızlı elde edilmektedir.

Cox regresyon modeli, belli bir hastalığa yakalanan hastaların sağkalım süresi ile bu sağkalım süresini etkilen değişkenleri tespit etmek ve bu değişenlerin etkilerini incelemek amacıyla literatürde sıklıkla kullanılan bir regresyon modelidir. Günümüzde artan veri sayısı ve bu nedenle oluşan karmaşık veri yapıları ile sağkalım

analizlerinde uygulanabilecek yeni yöntemlere ihtiyaç duyulmuştur. Veri madenciliği yöntemlerinden olan Rastgele Orman yöntemi Cox Regresyon yöntemi yerine kullanılmaktadır.

Son yıllarda rastgele orman teknikleri hastalık teşhisi ve tahmini için çoğunlukla kullanılmaktadır. Akman ve ekibinin (2010) yaptığı çalışmada, rastgele orman yöntemini sağlık verisinde kullanılmıştır. DNA veri seti gibi birbirinden farklı ve birden fazla değişkenin olduğu büyük boyutlu veri seti olan gen setleri arasından sağkalım için etkili olabilecek setleri elde edebilmek için uyguladıkları bu rastgele orman yönteminin uygun olduğunu önermişlerdir (Akman, Genç, Ankaralı, 2010). Exarchos ve ekibi (2011) ağız kanserine yakalanan sağkalım analizindeki hastaların oral skumoz hücreli karsinomun (OSCC)'un yeniden ortaya çıkmasını tahmin edebilme için araştırma yapmıştır. Bu çalışmalarında Bayes network, yapay sinir ağları, destek vektör, karar ağaçları, rastgele orman yöntemlerini veriye uygulayıp, yöntemlerin performansları birbiriyle karşılaştırılmış ve yorumlanmıştır (Exarchos, Goletsis, Fotiadis,2011). Weathers ve Cutler'in (2017), rastgele orman, koşullu çıkarım orman (Conditional Inference Forest) ve Cox regresyon modelleri ile çalışmış ve bu modelleri üç farklı veri kullanarak birbiriyle karşılaştırmıştır. Uyum indeksi ve hata kestirimleri ile bu yöntemler karşılaştırılmıştır. Çalışmanın sonucunda da Cox regresyon modelinin performansının rastgele orman yönteminden daha iyi olduğu söylenememiştir. (Weathers, Cutler, 2017)

Veri madenciliği, tanımlayıcı (Descriptive) ve tahmin edici (Predictive) olarak iki temel alanda ayrılır. Tıp alanında veri madenciliği çoğunlukla tahmin edici (Predictive) alanında uygulanmaktadır. Bu çalışmada da, ağaç tabanlı veri madenciliği yöntemlerinden birisi olan Rastgele Sağkalım Orman yönteminin incelenmesi ve Tıp alanından oluşturulan veri ile uygulaması gerçekleştirilip, uygulama sonuçlarının tartışılması hedeflenmektedir. Aynı zamanda da tıpta veri madenciliğinin uygulanma alanı hakkında bir düzenlenme yaratmak ve veri madenciliğinin sağkalım süresini etkileyen değişkenlerin seçimi açısından kullanılmasına yeni bir bakış açısı daha kazandırmaktır.

1. Genel Bilgiler

1.1 Veri Madenciliği

Veri madenciliği, çok büyük genişlikte veri topluluklarını araştırmak ve analiz etmek, ardından da verilerden anlamlı sonuçlar çıkarmak için uygulanan bilgisayar destekli bir prosestir. Veri, anlamlı bir sonuç elde edebilmek için elde edilen ilk bilgi ya da nicelik, olay, kayıt ve sayı kümeleri olarak tanımlanabilmektedir (Akpınar, 2000).

Günlük yaşamda birçok veri setleri vardır ve bu veriler ile birçok bilgiye ulaşılabilmektedir. Günümüzde oluşan veriler her geçen gün büyümeye devam etmektedir. Öyleki oluşan bu veriler zettabytelar hatta yottabytelar boyutuna ulaşmıştır.

Veri tabanı sistemlerine ihtiyacın ve kullanımının artması işlenen verinin her gün artmasına neden olmaktadır. Bu artış karşısında karşılaşılan en büyük problem, bu verilerin nasıl analiz edileceği ve nasıl çıkarımlar yapılabileceğidir. Klasik yöntem olan veri tabanlarından sorgu (Query) ile raporlama yapılması veya daha farklı raporlama araçlarının performansının büyük veri ile daha düşük ve zor olması, Veri Tabanlarında Bilgi Keşfi-VTBK (Knowledge Discovery in Databases) kavramını oluşturmuştur. VTBK sürecinde, en önemli adım modelin değerlendirilmek üzere kurulmasının aşaması olan Veri Madenciliği (Data Mining)'dir (Akpınar, 2000). Veri madenciliği adımı bu sürecin ilk temelini yaratan keşf etme uygulandığı adım olduğu gibi bu veri sürecinden ayrı bir şekilde de farklı bir adım gibi değerlendirmektedir. (Koyuncugil, 2007).

Şekil.2.1'de veri madenciliğinin ilişki içerisinde olduğu diğer bilimler gösterilmektedir. Veri madenciliği, birden çok disiplinli bir alandır. (Sayyad, https://www.saedsayad.com/data_mining.html, Erişim Tarihi:12.05.2021)

Şekil 1.1 Veri madenciliğinin ilişkili olduğu bilimler

1.2 Veri Tabanlarında Bilgi Keşfi Aşamaları

Fayyad ve arkadaşlarına göre, VTBK temel olarak 5 adımdan oluşur (Fayyad, Piatetsky-Shapiro, Smyth, Uthurusamy, 1996).

1. Veri Seçimi (Data Selection): Bu adımda çalışmanın yapılacağı veri tabanından verilerin uygulamada geçerli olacak bir veri seti seçilip, veri dosyasının oluşturulmasıdır. Araştırmada analiz edilecek tüm verileri içermesi gerektiği için bu önemli bir adımdır. Eksik verilerin içirmesi durumunda tekrar bu adıma dönülmesi gerekebilir.
2. Veri Temizleme ve Önişleme (Data Cleaning & Preprocessing): Sürecin içerisinde yer alacak örneklemdeki eksik verilerin ayrıştırılıp ve eksik olan özelliklerin eklendiği adımdır. Veri Temizleme ve Önişleme adımında analizin kalitesi artar. Verilerin temizlenip özetlendiği veya Önişleme işlemlerinin gerçekleştirilerek madencilik için uygun olan forma getirildiği ve son halinin edildiği adımdır (Han, Kamber, Pei 2011).
3. Veri İndirgeme (Data Reduction): Çalışmanın yapılacağı örneklemde yer alan çalışmayla ilgili olmayan değişkenlerin ve tekrarlı kayıtların belirlendiği adımdır.
4. Veri Madenciliği (Data Mining): Veri madenciliği adımında, çalışmaya uygun olan bir algoritma seçilerek hazırlanmış veri setine analiz uygulanır.

5. Değerlendirme (Evaluation): En son adım olan Değerlendirme adımında, Analiz sonucunda keşfedilen çıktının geçerliliği, yeniliği, yararlılığı ve basit olmasına göre yorumlanır.

VTBK sürecinde öncelikli olan veri madenciliği adımıdır. Bununla birlikte, süreçte yer alan diğer adımlar, VTBK'nın pratikte başarılı bir şekilde uygulanması için gereklidir. Bu çalışmada da, literatürde en çok dikkat çeken veri madenciliği konusuna odaklanmaktadır.

1.3 Veri Madenciliği Yöntemleri

Veri madenciliği yöntemleri iki ana başlıkta toplanır. Bunlar tahmin edici (Predictive) ve tanımlayıcı (Descriptive) olarak ikiye ayrılmıştır (Zhong, Zhou, 2003). Tahmin edici modellerde, daha önceden sonuçları hakkında bilgi sahibi olunan veriler ile model geliştirilmesi, geliştirilen bu modelden yararlanılarak sonuçları hakkında bilgi sahibi olunmayan veriler için sonuçların tahmin edilmesi hedeflenir. Örneğin, geçmişe dönük satış dataları bilinen bir ürünün, geçmişe dönük olan sonuçları ile gelecekteki satışlarını tahmin edilebilir. Tanımlayıcı modellerde ise sahip olunan veriler ile karar vermeye öncülük etmek için veri düzenlerinin tanımlanması sağlanmaktadır.

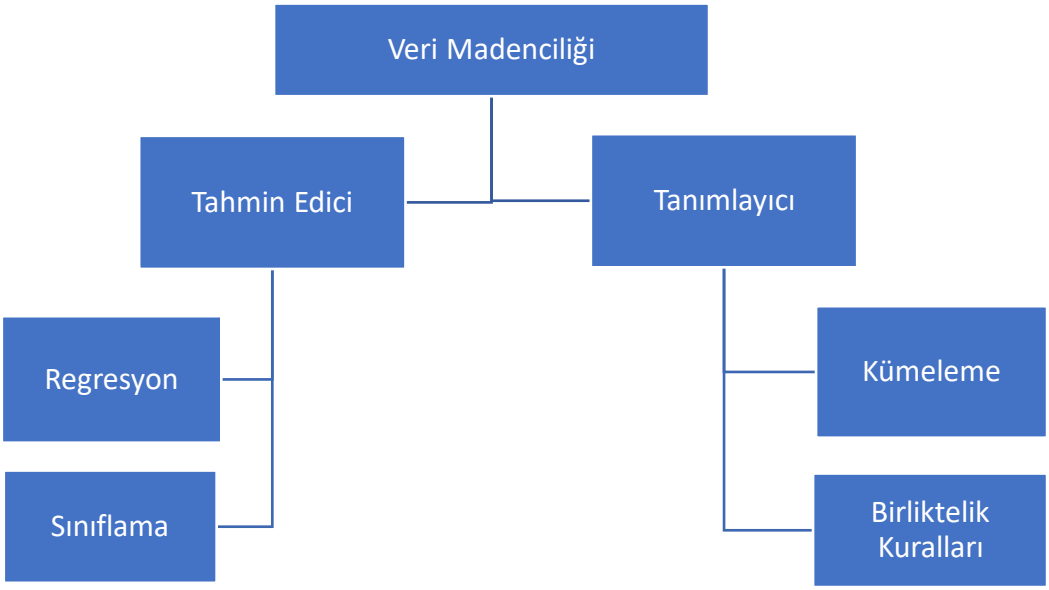
Veri madenciliği modelleri,

1- Sınıflama (Classification) ve Regresyon (Regression)

2- Kümeleme (Clustering)

3- Birliktelik Kuralları (Association Rules)

olmak üzere üç ana başlık altında incelenir (Akpınar, 2000). Sınıflama (Classification) ve Regresyon (Regression) modelleri tahmin edici, Kümeleme (Clustering) ve Birliktelik Kuralları (Association Rules) modelleri tanımlayıcı modellerdir (Akpınar, 2000).



Şekil 1.3. Veri Madenciliği Yöntemleri

1.3.1 Sınıflama ve Regresyon

Sınıflama ve regresyon, verileri sınıflandırarak önemini tespit eden veya sonuç verileri tahmin edebilen modelleri kuran veri analizi türüdür (Han & Kamber, 2000). Sınıflama ile kategorik veriler tahmin edinebilir iken, regresyon sürekli olan verilerin tahmin edilmesinde uygulanır (Han & Kamber, 2000). Sınıflama ve regresyon modellerinde karar ağaçları, yapay sinir ağları, genetik algoritmalar, K-en yakın komşu, Lojistik Regresyon (Logistic Regression) ve naïve-bayes gibi yöntemler uygulanmaktadır.

1.3.1.1 Karar Ağaçları (Decision Trees)

Karar ağaçlarının oluşturulmasında uzmanlık alan bilgisine gerek duyulmaz. Karmaşık yapıların düzensizliğini en aza indirir ve karmaşık yapıdaki bu yapıların çıktılarında kesin değerler sunar. Boyutu yüksek olan verileri kolaylıkla işleyebilir. Analizin sonucunda çıktıların yorumlanması kolaydır. Karar ağaçları sadece sayısal veriler değil kategorik verilerin de işlenmesini sağlar. Karar ağaçları, ağaç formunda olan, tahmin edici (Predictive) olan çalışma yöntemidir (Berry Michael and Linoff Gordon, 1999). Karar ağacı karar düğümlerinden, onlardan oluşan dallardan ve dalların sonucu olan yapraklardan oluşur (Han & Kamber, 2000). Uygulanacak test karar düğümü ile belirtilir. Testin sonucundada karar ağacı herhangi bir veri kaybetmeden dallara ayrılır. Karar düğümü üst seviyedeki özelliklere uygun olacak şekilde, her oluştuğunda

tekrardan teste ve dallara bölünür. Karar ağacının karar düğümü oluşuncaya, sınıflama işlemi tamamlanıncaya, kadar tüm dalları kendi içinde sınıflara ayrılmaya devam eder. Eğer bir dalın sonunda sınıflama işlemi oluşmuyorsa, o dalın sonucunda artık bir karar düğümü oluşmuştur.

Karar ağacı tekniğini ile verinin sınıflanması iki adım olacak şekilde gerçekleşir (Han & Kamber, 2000). Öğrenme basamağı karar ağacı tekniğinin ilk basamağıdır. Önceden elde edilmiş olan bir eğitim verisi, tahmin modelini elde etmek amacıyla sınıflama algoritması tarafından analiz edilir. Karar ağacı tekniğinin ikinci basamağında ise sınıflama yer alır. Bu basamakta önceden belirlenen test verisi, eğitim verisiyle oluşturulmuş tahmin modelini test eder. Eğer test sonucunda doğruluğu kanıtlanırsa, karar ağacı tekniği yeni verilerle tekrardan sınıflanabilir halde olur.

Karar ağacı yöntemi çıktı olarak tek bir çıktı verir. Ve yalnızca kategorik olarak çıktı üretir. Kararsız bir sınıflandırıcıdır, yani sınıflandırıcının doğruluğu veri kümesinin türüne bağlı olur. Eğer veri kümesinin türü sayısal ise, karar ağacının yapısı karmaşık hale gelir.

1.3.1.2 Yapay Sinir Ağları (Artifical Neural Networks)

Yapay sinir ağları (YSA), sırasıyla bir şeyi öğrenme, onu hatırlama ve öğrenilen şeyi genelle atayarak çevresinde topladığı yeni verilerden veri üretme gibi temel uygulamaların, tıpkı insan beynindeki öğrenme yöntemine benzer olacak şekilde gerçekleştirilen bir bilgisayar yazılımıdır. Yapay sinir ağları; insan beyni gibi, öğrenme işlemlerinin matematiksel bir şekilde modellenmesi sonucunda gelişmiştir (Kabalcı, 2014). Yapay Sinir Ağları teknikleri tahminleme yapmak, veriyi sınıflandırmak, veriyi ilişkilendirmek, veriyi filtrelemek ve veriyi yorumlamak için uygulanmaktadır (Ağyar, 2015).

Bağımlı ve bağımsız türdeki değişkenlerde, Yapay Sinir Ağları ile birbiri arasındaki karmaşık düzeydeki ilişkiler kolaylıkla tespit edilebilir. YSA tekniği ile gürültülü veriler işlenebilir. YSA tekniğine yerel minimum değerler oluştuğunda, modelden alınabilecek performans düşmektedir. Bunun dışında bu teknik aşırı uyum gösterebilir. Yapay sinir ağı tekniğinde işlemleri yorumlamak zordur ve eğer modelde büyük sinir ağları varsa yüksek boyutlu işlem süresi gerekmektedir.

1.3.1.3 Genetik Algoritmalar (Genetic Algorithms)

Genetik Algoritmalar (GA), birçok gerçek hayatta da kullanılabilen yaygın olan optimizasyon tekniklerinden biridir. Bu yöntem, gelişen tekniklerden biri olarak optimizasyon problemlerine daha iyi çözümleri geliştirip daha iyi bir çözüm arar (Popa, 2012). Genetik algoritmalar biyolojik evrimin ve doğal seçimin temel amaçlarını örnek alan rastlantısal arama algoritmalarındandır. Genetik algoritmalar, mutasyon ,seçim, çaprazlama gibi canlı evrimin oluşumu sağlayan adımları baz alan, güçlü genlerin zayıf olan genlerden üstün olduğu biyolojik evriminin simülasyonudur. Genetik algoritmalar, hem sürekli hem de kategorik verilerde optimizasyon sorunları için uygulanmaktadır (Scrucca 2013).

1.3.1.4 K-En Yakın Komşu (K-Nearest Neighbor)

K En Yakın Komşu yöntemi, verileri sınıflandırmak için kullanılan yöntemler arasında yer alır. Sınıflandırma yapılırken, öğrenme kümesinde yer alan verilere benzerliklerine göre; en yakın olan k adet verinin ortalaması ile hesaplanan eşik değere göre sınıflandırma yapılır. (Shah & Kusiak, 2004). Bu yöntemde, oluşturulan sınıfların özelliklerinin daha önceden açıkça belirlenmiş olması önemlidir. K en yakın komşu yönteminin performansı eşsizlik değeri, eşik değeri ve öğrenme sınıfındaki normal davranışların uygun düzeyde olması ile belirlenir. K En Yakın Komşu yöntemi uygulaması kolay ve hızlı bir yöntemdir. Bunun yanı sıra gürültülü verilere duyarlı aynı zamanda uygulama hızı yavaştır. Örneklem bilinmediği durumda, örneklemin gerçek değerini tespit etmede de kullanılabilir (Han & Kamber, 2000).

1.3.1.5 Naïve-Bayes

Thomas Bayes'in sunduğu bir sınıflandırma algoritması olan Naïve Bayes sınıflandırma algoritması, olasılık teorilerine dayanarak oluşturulmuş verilerin en olası sınıfını tespit etmeyi amaçlar. Naïve Bayes, sınıflandırma yöntemi ile kalp hastalığına sahip hastalarının teşhisinde başarılı sonuçlar üreten veri madenciliği tekniğidir (Sitar-Taut, 2009). Naïve Bayes, en olası ve uygun olan sınıflandırmaları oluşturmak için olasılık teorisine dayanır (Yadav, 2012).

Naïve Bayes sınıflandırmasında sisteme genellikle %80 oranında eğitim, %20 oaranında test verileri belirlenir. Eğitim verilerinin yer aldığı bir sınıfı yer alır. Eğitim setinde yer alan veriler ile olasılık hesaplanır. Sonrasında test verileri eğitim

verilerinden elde edilen olasılıklara göre işlenir ve bulunduğu sınıf belirlenir. Analizde eğitim verisi ne kadar büyük boyutlu olursa test verisine en uygun sınıfı tespit etmek kolaylaşır. Navies Bayes, veri madenciliğinde kullanılan en yaygın tekniktir (Kaur & Bawa, 2015).

1.3.1.6 Regresyon

Regresyon, çoğunlukla değişkenler arasındaki belirli ilişkiyi incelemek için kullanılır (Parvez Ahmad, 2015). Analizde kullanılan Bağımlı değişken sürekli değişken yapısında ise, genellikle doğrusal regresyon modeli uygulanır. Doğrusal regresyon modelinde en önemli varsayım hata terimlerinin normal dağılıma sahip olmasıdır. Bununla birlikte doğrusal regresyon modelinde yer alan bağımsız değişkenlerin kesikli veya sürekli değişken yapısında olmaları, model tahmini için seçilecek yöntemi etkilemez. Kısaca doğrusal regresyon modelinde yer alacak bağımsız değişkenler hem kesikli hem de sürekli değişken yapısında olabilir. Bağımlı değişkenin kategorik değişken yapısında olması istatistiksel normallik varsayımı dışında olmasına sebep olur. Bu durumda doğrusal regresyon modelinin uygulanması doğru olmayacaktır. Bağımlı değişkenin iki ya da çok kategorik değişken olması halinde uygulanabilecek modeller çok çeşitlidir. Bu modeller arasında en çok tercih edilen model lojistik regresyon modelidir. Lojistik regresyon modeli ile doğrusal regresyon modeli arasındaki en büyük fark, lojistik regresyon analizinde bağımlı değişkenin iki ya da çok kategorik veri olmasıdır.

Bağımlı değişkenin değeri doğrusal regresyon analiziyle hesaplanırken, bağımlı değişkenin belirli bir olasılığı lojistik regresyonla elde edilir.

1.3.2 Kümeleme

Kümeleme yönteminde veri sınıflara veya kümelere ayrılır (Karypis, Han, Kumar, 1999). Eş kümede yer alan değişkenler birbirleriyle aynı iken farklı kümelerde yer alan değişkenlerden birbirlerinden ayrıdır. Sınfılandırma modellerine yer alan veri sınıfları kümeleme modelinde yer almaz (Ramkumar, Swami, 1998). Sınıflama modellerinde, verilerin olası sınıfları önceden tespit edilmekte ve yeni bir verinin hangi sınıfta yer alabileceği tahmin edilirken, kümeleme modellerinde ise herhangi bir sınıfı olamayan veriler kümelere dağıtılır. Bazı analizlerde kümeleme modeli, sınıflama modelinin önişlemini gerçekleştirebilmektedir (Ramkumar, Swami, 1998).

Literatürde yer alan birçok kümeleme algoritması yer amaktadır. Kümeleme algoritması veri tiplerine ve amaca göre değişebilir. Kümeleme yöntemleri, Bölme (Partitioning methods), Hiyerarşik (Hierarchical methods),Yoğunluk tabanlı (Density-based methods), Izgara tabanlı (Grid-based methods) ve Model tabanlı yöntemler (Model-based methods) olarak farklılaşır (Han, Kamber, 2000).

Bölme yöntemlerinde, n veri tabanında yer alan değişken sayısı ve k oluşturulan küme sayısıdır. Bölme algoritmasında n adet nesne, k adet kümeye ayrılır ($k \leq n$). Aynı kümede yer alan nesneler birbirleriyle benzer iken diğer kümelerde yer alan elemanlarından farklıdır. (Han & Kamber, 2000). Kullanımı en çok olan ve en iyi bilinen bölme yöntemi k-means yöntemidir (Fayyad, 1998).

Hiyerarşık kümeleme yönteminde, bölünme eğer aşağıdan yukarı yönlü ise agglomerative, yukarıdan aşağı yönlü ise divisive olarak hiyerarşik bölünme adlandırılır (Han, 2001).

1.3.3 Birliktelik

Birliktelik kuralları uygulaması Agrawal, Imielinski ve Swami tarafından 1993 yılında veri madenciliğinde kullanılmış ilk uygulamadandır (Agrawal, Imieliński, Swami, 1993). Olayların aynı anda gerçekleşmesi olayını belli olasılıklarla ortaya koyarak analiz eden veri madenciliği uygulamasıdır. Birliktelik kuralları, geçmiş verilerin analizi sonucunda elde edilen birliktelik olasılıklarının tespiti sonrasında gelecekte kullanılacak verilerin analizini sağlar.

1.4 İstatistiksel Açıdan Veri Madenciliği veya İstatistiksel Öğrenme

İstatistik birçok alanı ile veri madenciliğinin birçok alanı içeçe geçmiştir (Zhao, Luan 2006). Veri madenciliği için istatistik, verilerin modellenmesi, indirgenmesi gibi temelde yer alan veri ön işleme adımlarında ve analiz sonuçlarının değerlendirilmesinde kullanılması gereken alandır. Verinin bilgiye dönüştürülmesi veya öğrenilmesi istatistik ve veri madenciliğinin en önemli ortak paydasıdır (Ganesh, 2002:1, Kuonen, 2004:5)

Veri madenciliği yöntemlerinin çoğunda analizin temeli istatistiksel yöntemleri temel alır. Veri madenciliğinde istatistik olmadan madencilikten bahsedilemez, temelinde istatistik yer almaktadır (Kumar, Bhardwaj, 2011). Bazı diğer araştırmacılara göre de istatistiksel olarak veri madenciliği yöntemleri, istatistiğe nazaran daha esnek

yöntemlerdir ve veri madenciliği istatistikte yer alan çok değişkenli analizle benzer bulunmuştur (Kuonen, 2004:5).

İki yöntemde yer alan analizler içerisinde kümeleme, diskriminant, regresyon ve korelasyon uygulamalarının hem veri madenciliği hem de istatistik alanlarında yer alan ortak yöntemler olduğu söylenebilir (Tüzüntürk, 2010).

Veri madenciliğinde tümden gelim yaklaşımı varken, istatistik tüme varım yaklaşımına dayanır (Tüzüntürk, 2010). Veri madenciliğimde hipotez testleri yer almaz bununla birlikte istatistikte hipotezin anlamı büyüktür. Bu nedenle de istatistikte yer alan anlamlılık düzeyinin veri madenciliğinde anlamı yoktur (Tüzüntürk, 2010).

İstatistiksel yöntemler geleneksel yöntemlerden olduğu gibi çalışmacı eğer büyük hacimli veri setleri ile karşılaştığında bu geleneksel yöntem analiz açısından yetersiz kalabilmektedir. Büyük hacimli verilerin analizi için veri madenciliği yöntemleri ile analiz edilmesi daha uygun ve kullanılışlı olmaktadır.

Büyük verilerle analizlerin performansı istatistik ve veri madenciliğinde farklılık gösterir. Veri madenciliğinde istatistiğe göre daha kolay bir şekilde analizler yapılabilir (Ganesh, 2002).

1.5 Sağlık Alanında Veri Madenciliği Uygulamaları

Veri madenciliğinde yer alan algoritmaları tıbbi verilere uygulamak için, araştırmacılar veri madenciliği algoritmalarındaki yöntemlerin türünü ve işlevlerini net olarak anlamalıdır. Veri madenciliği algoritmaları tanımlayıcı ve tahmine dayalı olarak iki kategoriye ayrılıdır (Fayyad, Shapiro, Smyth, Uthurusamy, 1996). Tanımlayıcı veri madenciliğinde verilerin benzerliğini belirleyerek ve bilinmeyen modelleri veya verilerdeki ilişkilendirmeleri tespit eder. Tanımlayıcı veri madenciliği araştırıcı bir yöntemdir. Sınıflandırma, regresyon, zaman serisi analizi ve tahmini içeren tahmin veri madenciliği, eğitim verilerinden tahminleme kurallarını oluşturur ve bu kurallarla tahmini yapılamayan verileri tahminleyebilmek için kullanılır (Kharya, 2012).

Veri madenciliği teknikleri gelecekteki modeller için etkili ve öngörücüdür. Kullanımı kolay, tahminin geçmiş koşullara dayanması yani geçmiş verilerden öğrenerek çalışması ve bunları çok sayıda kaynaktan gelen veriler ile yönetip, çıktıları çıkarması, onu güvenilir ve pratik yapan özellikleridir (Fayyad, Shapiro, Smyth, Uthurusamy, 1996).

Günümüzde hastane bilgi sistemlerinin gelişmesi ile tıp ve sağlık alanında, hastalara ait demografik, tanı ve tedavi durumları, test sonuçları gibi birçok hastaya ait veri büyük veri ambarlarında saklanabilmekte ve ulaşımı kolay olmaktadır. Tıp alanı günümüzde en önemli bilimsel araştırmaların gerçekleştirildiği bir alan olduğu için bu alandaki bilgi sistemleri her zaman önemlidir. Bilgi sistemlerinde yer alan bu verilerden yararlanılarak aynı hastalığa sahip olanların ortak özelliklerinin belirlenmesi ve tahmin edilmesi, herhangi bir hastalığa yapılan müdahaleler sonra hastane maliyetinin tahmin edilmesi gibi birçok analizler yapılabilir (Kudyba, 2004).

1.5.1 Tedavi etkinliği

Veri madenciliği yöntemleri ile hastalığa uygulanan çalışmaların verimliliği ölçülebilir. Hastalığın analizlerindeki tüm adımlarını, yani hastalığın sebebi, bulguları, tedavi yöntemi gibi aşamaları analizleyip, sonuçları değerlendirilerek, en etkili tedavi planı ortaya çıkarılabilir (Milley, 2000). Tedavi ile ilgili diğer veri madenciliği uygulamaları, tedavi yönteminin yan etkilerini ilişkilendirmede, tanıda gerekli olan ortak semptomları birleştirmede, ana popülasyondan ayrı olarak tedaviye farklı yanıt veren alt popülasyonların tedavisindeki etkili olan tedavi içeriğini belirlemede uygulanır (Milley, 2000).

Örneğin, iki farklı tedavi yönteminden hangisinin en iyi sonuç veren ve en uygun maliyeli olduğuna, aynı hastalığa sahip bireylerde ayrı ayrı bu tedavi yöntemleri uygulanıp sonuçları karşılaştırılarak varılabilir (Kincade, 1998).

Hastanelerde ilaçla tedavisi zor olan oldukça fazla enfeksiyon yer almaktadır. Bu enfeksiyonlardan etkilenen hastaların sayısı oldukça fazladır. Veri madenciliği aracılığıyla ile enfeksiyon analizleri yapılır ve analiz çıktılarına göre enfeksiyonların incelenmesi sağlanır (Elmaghraby, Kantardzic, Wachowiak, 2006).

1.5.2 Sağlık Yönetimi

Veri madenciliği teknikleri ile hastanelerin tüm özellikleri incelenerek, derecelendirilir ve ciddi hastalıkları olan hastaları tedavi edebilme özelliğine göre, numaralandılır ve yüksek dereceye sahip olan hastaneler için diğer hastalara göre durumu ciddi olan hastaların tedavisinde uygun olabilir (Obenshain, 2004).

Sağlık sektöründe tıbbi cihazlar tedavi ve teşhis koyma durumunda çok önemlidir. Veri madenciliği bu cihazlarda yaygın olarak kullanılır.

2.5.1 Dolandırıcılık ve sahteciliğin önüne geçmek

Birçok alanda olduğu gibi sağlık sektöründe de dolandırıcılık ve sahteciliğe rastlanmaktadır. Veri madenciliği uygulamaları sayesinde dolandırıcılık ve kötüye kullanım, sahte reçeteleme, sahte sigorta ve uygun olmayan tıbbi yöntemler tespit edilebilir (Durairaj, Ranjani, 2013).

2. Gerekçe ve Yöntem

2.1 Sağkalım Analizi

Sağkalım analizi ilgili olayın sağkalım süresini ele alarak, bu hastalığın tekrar meydana çıkıp çıkmayacağını, yaşam ve ölüm ihtimalini, ortalama yaşam süresini gibi parametreleri tahmin ederve sağkalıma etkisi olabileceği düşünülen diğer faktörlerin sağkalıma etkisini analiz eder (Özdamar, 2015).

Sağkalım analizi temel amacı, gözlenmesi istenen olgu oluşuncaya kadarki zamanı modelleştirmek için kullanılır. Analiz sonucunda da ortaya çıkan model ile sağkalım süresi analiz edilebilir. Burada yer alan süre, hastanın takibinin başladığı andan olayın gerçekleştiği ana kadar geçen günler, haftalar, aylar veya yıllardır. Gerçekleşmesi izlenen olay ise, hastalığın tekrardan nüks etmesi, hastanın iyileşmesi, ölüm veya bunların dışında hastanın yaşayabileceği herhangi bir olayın başlaması olabilir ve bu durum başarısızlık olarak nitelendirilir.

Sağkalım analizi sadece sağlık alanında insanların yaşam sürelerinde değil, aynı zamanda mühendislikte mekanik eşyaların dayanma sürelerinde de uygulanabilir. Aynı zamanda sağkalım analizi hayvanların kullanıldığı laboratuvar testlerinde de uygulanabilir. Fakat sağkalım analizi, çoğunlukla hastaların yaşam sürelerini analiz etmek için kullanılmaktadır (Lee ve Go, 1997).

Sağkalım analizinde, bir olayın başlangıç anından başarısızlıkla sonuçlandığı ana kadar geçen süreye "yaşam süresi" veya "sağkalım süresi" adı verilir (Johnson & Johnson,1980). Sağkalım analizinde yaşam süresi sürekli bir değişkendir.

Sağkalım analizinde yer alan "başarısızlık" olayı, bir kez ölçümlenir ve olayın denekte tespit edilmiş olmasıdır. İnsanlar ve hayvanlarda bu olay genelde ölüm, hastalık veya nüks, mekanik aletlerde ise bozulma anlamına gelir (Nelson, 2003).

Sağkalım analizinde yaşam süresinin ölçümlenebilmesi için,

• İlgilenilen olayın başlangıç zamanı, tüm bireyde veya birimlerde açık bir şekilde tanımlanmalıdır.

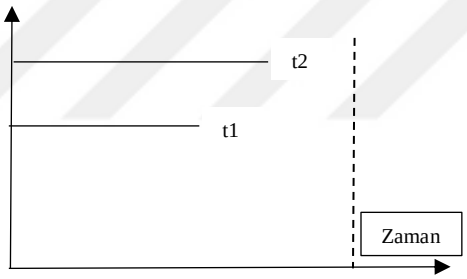
• Yaşam süresi ölçümlenebilir olmalıdır. (yıl, ay, hafta, gün.. vb.)

• Sağ kalım analizindeki tüm deneklerde başarısızlık olayı açık bir şekilde bilinmelidir (Cox ve Oakes, 2018).

2.1.1 Veri Türleri

Sağkalım analizinde, sağkalım süresi bilinmeyen olaylar sansürlü (censored) olarak isimlendirilir (Kleinbaum ve Klein, 2010). Yaşam analizinde sansürlü verilerin kullanılabilir olması onu diğer yöntemlerden ayırmaktadır. Çalışmaya katılan deneklerin gözlem esnasında başka bir olaydan dolayı ölmesi, deneğin gözlem devam ederken herhangi bir sebepten ulaşılabilir olmaması, gözlem sonunda olay veya deneğin gözlenemesi durumu gibi durumlarda sansürün meydana geldiği söylenmektedir.

2.1.1.1 Sansürsüz Veriler



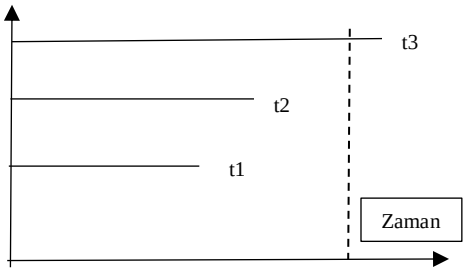
Şekil 2.1 Sansürsüz veri

Şekil.3.1 incelendiğinde, grafikte yer alan üç denek, sağkalım süresi içerisinde başarısızlıkla sona ermiştir ve başarısızlık zamanı kesindir. Bu tip veriler sansürsüz veri olarak adlandırılmıştır (Kleinbaum ve Klein, 2010).

Sansürlü veriler, sağdan sansürlü veri, soldan sansürlü ve aralık sansürlü olmak üzere temel olarak 3 gruba ayrılmaktadır.

2.1.1.2 Sağdan Sansürlü veriler

Sağdan sansürlü olayda, denek gözlem süresi sonunda takip dışı kalmış veya gözlem süresinin sonrasında yani sağında tamamlanmıştır. Sağdan sansürlü olan deneğin yaşam süresinin uzunluğu, gözlemin tamamlanmış olduğu zamandan daha fazladır.



Şekil 2.2 Sağdan sansürlü veriler

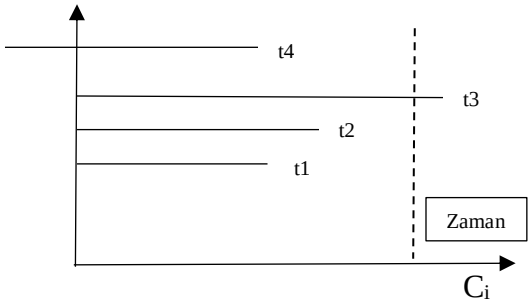
C_i sansürleme anı, T_i deneğin sağkalım süresi iken, $T_i > C_i$ ise deneğe ait sağkalım süresi için sağdan sansürlüdür denilir. Grafikte t_3 sağdan sansürlü veriye örnektir.

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

δ_i=0 olduğunda deneğin sağkalım süresi sansürlü, δ_i=1 olduğunda ise sağkalım süresi bilinmektedir yani sansürsüzdür (Nelson, 1982).

2.1.1.3 Soldan sansürlü veriler

Sağkalım analizinde bazı denekler sansür zamanından önce olayı deneyimlediği biliniyor ancak olayın tam zamanı bilinemiyorsa soldan sansürlü veri olarak adlandırılır.



Şekil 2.3 Soldan sansürlü veriler

C_i sansürleme anı, T_i deneğin sağkalım süresi iken $T_i < C_i$ olursa bu denek için sağkalım zamanı soldan sansürlüdür denir. Grafikte t_4 soldan sansürlü veriye örnektir.

2.1.1.4 Aralıklı sansürlü veriler

Denekler, gözlem süresinde tanımlanan olaylara mağruz kalır. Aralıklı sansürlü veriler bu gözlem süresince başarısızlığın kesin olarak ortaya çıktığı zamanın bilinmediği

durumlardır. Aralıklı sansürlü veriler genel bir sansür türüdür ve yaşam süresi yalnızca belirli bir aralıkta gözlemlenir. Genellikle takip gerektiren çalışmalarda kullanılmaktadır. Buna örnek olarak, bazı mekaniklerin sağkalım süresi analiz edilmekte ve belirli aralıklarla deneklerin kontrollerin yapıldığı bir durumda aralıklı sansür doğal olarak meydana çıkabilir. Burada denekler belirli aralıklarla kontrol edildiğinden, mekaniklerin çalışma süresi muhtemelen aralıklı sansürlü olacaktır (Klein, Moeschberger, 2003).

2.1.2 Sağkalım analizi fonksiyonları

2.1.2.1 Sağkalım fonksiyonu

Sağkalım fonksiyonu S(*t*) olarak tanımlanır. S(t), *T* rastgele değişkeninin belirli olan bir zamanı(t) aşma olasılığıdır.

T sürekli bir rastgele değişken olup, 0 ile ∞ değerleri arasında değerler alır. Sağkalım fonksiyonu, t'in farklı değerleri için sağkalım olasılığı hesaplar (Kleinbaum, Klein, 2010).

Sürekli rastgele bir değişken olan *T* değişkeni için 0<t<∞ zaman aralığında, sağkalım fonksiyonu aşağıdaki gibi formülize edilir.

$$S(T) = P(T > t) = \int_t^{\infty} f(t)dt \qquad \text{Denklem 2.1}$$

Formülü ile hesaplanır.

Sağkalım analizinde gözleme başlanılan zaman yani t=0 anında, S(t)=S(0)=1 olacaktır. Gözlemin başında (t=0), ilgilenilen olay henüz ortaya çıkmadığı için, sağkalım olasılığı bir olarak hesaplanacaktır. Aynı şekilde t=∞ zamanında, S(t)=S(∞)=0 olacaktır. Yani eğer gözlem süresi sonsuz olsaydı, gözlem sonunda hayatta kalan kimse olmazdı (Kleinbaum ve Klein, 2010).

T'in birikimli olasılık yoğunluk fonksiyonu, F(t) ile gösterildiğinde, sağkalım fonksiyonu,

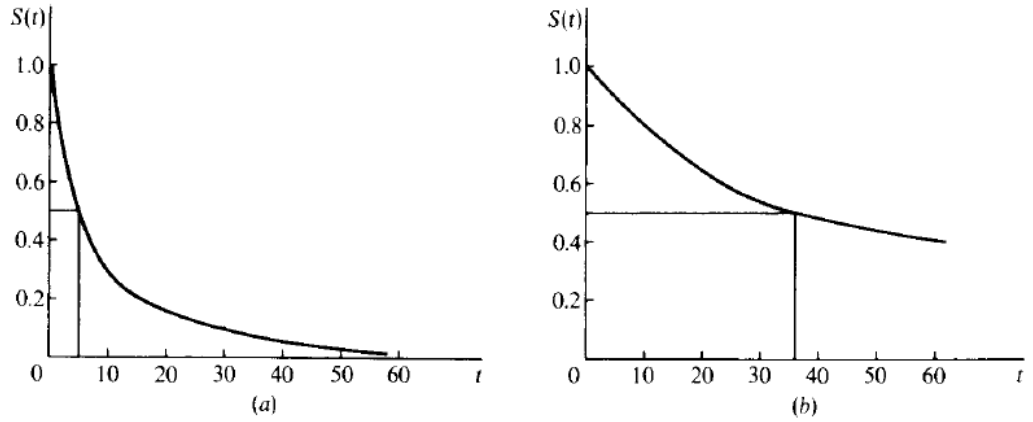
$$S(t) = F(t) = 1 - S(t) \qquad \text{Denklem 2.2}$$

S(t) azalan bir fonksiyondur.

$$S(t) = \begin{cases} 1, & t = 0 \; iken \\ 0, & t = \infty \; iken \end{cases}$$

Yani, zaman 0 olarak düşünüldüğünde sağkalım olasılığı en yüksek değerde yani 1, zamanın sonsuza eşit olduğu düşünüldüğünde de 0 olarak hesaplanır.

Sağkalım olasılığının grafiğine sağkalım eğrisi denir. Şekil 2.4'de a numaralı eğri dik bir eğridir ve burada sağkalım oranının düşük veya sağkalım süresinin kısa olduğunu temsil ederken, b numaralı eğri ise aşamalı veya sabit bir sağkalım eğridir ve sağkalım olasılığının yüksek, sağkalım süresinin uzun olduğunu gösterir (Lee, Wang, 2003).



Şekil 2.4 Lee ve Wang (2003), Sağkalım Eğrisi örnekleri

Sağkalım analizinde eğer sansürlü veri yer almıyorsa, sağkalım fonksiyonu t zamandan daha uzun süre hayatta kalan hastaların oranı olarak tahmin edilebir. Buna göre sağkalım fonksiyonu tahmini,

$$\hat{S}(t) = \frac{\text{t zamanından daha uzun sağkalan hasta sayısı}}{\text{toplam hasta sayısı}} \qquad \text{Denklem 2.3}$$

Bu tahminlemede eğer sansürlü gözlemler yer alsaydı, t zamandından daha uzun sağkalan hasta sayısını tahmin etmek güç olacaktı. Bu nedenle sansürlü gözlemlerin yer aldığı sağkalım analizlerinde Denklem 2.3 kullanılmamaktadır.

2.1.2.2 Olasılık yoğunluk fonksiyonu

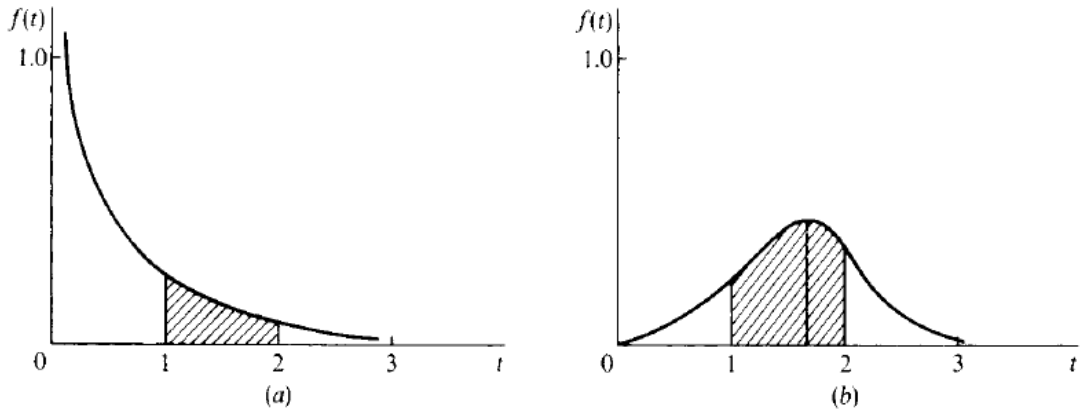
Olasılık yoğunluk fonksiyonu (f(t)), t ve t+Δt aralığındaki kısa aralıklardaki başarısız deneklerin olasılıksal limiti veya bu aralıkların başarısızlık olasığıdır. Olasılık yoğunluk fonksiyonu Denklem 2.4' de formülize edilmiştir.

$$f(t)=\frac{\lim_{\Delta t\to\infty} P(t\leq T\leq t+\Delta t)}{\Delta t} \qquad \text{Denklem 2.4}$$

Olasılık yoğunluk fonksiyonunun eğrisi Şekil 2.5'de 2 farklı şekilde verilmiştir ve yoğunluk eğrisi olarak adlandırılır.

Olasılık yoğunluk fonksiyonu iki özelliğe sahiptir;

1. İlk olarak f(t) negatif fonksiyon değildir ve t>= 0 için f(t)>0 iken, t<0 için ise f(t)=0'dır.
2. Olasılık yoğunluk eğrisi ile t ekseni arasındaki alan her zaman 1'e eşittir (Lee ve Wang 2003).



Şekil 2.5 Lee ve Wang (2003), yoğunluk eğrisi örnekleri

Olasılık yoğunluk fonksiyonu tahmini sağkalım fonksiyonu tahminine benzer şekilde sansürlü gözlemlerin olmadı zamanda kullanılır ve sansürlü gözlem olduğunda kullanılamaz.

$$\hat{f}(t)=\frac{t\ zamanında\ başlangıç\ aralığıdaki\ ölen\ hasta\ sayısı}{(Toplam\ hasta\ sayısı)*(aralık\ denişliği)} \qquad \text{Denklem 2.5}$$

Şekil 2.5'de yer alan a numaralı grafikte başarısızlık olasılığının yüksek ve zaman arttıkça azalan başarısızlık oranı modeli, b numaralı grafikte ise başarısızlık olasılığının en yüksek olduğu tepe noktasını yaklaşık olarak 1.7'inci zamanda meydana geldiğini göstermektedir.

Her iki şekilde de 1 ve 2 birim zaman aralığındaki başarısız olan hastaların oranı yoğunluk eğrisi ile eksen arasındaki taralı olarak belirtilen alana eşittir. Olasılık yoğunluk fonksiyonu, mutlak başarısızlık oranı olarak da bilinmektedir (Lee, Wang, 2003).

2.1.2.3 Hazard (Risk) fonksiyonu

Hazard fonksiyonu, deneğin başlangıçta sağ olduğu bilindiğinde (t), t ve t+Δt zamanda başarısızlık (hayatının sonlanmasının) olasılığın tanımıdır. Denklem 2.6'da formülize edilmiştir.

$$h(t) = \frac{\lim_{\Delta t \to 0} P(t \leq T \leq t+\Delta t / T \geq t)}{\Delta t} \qquad \text{Denklem 2.6}$$

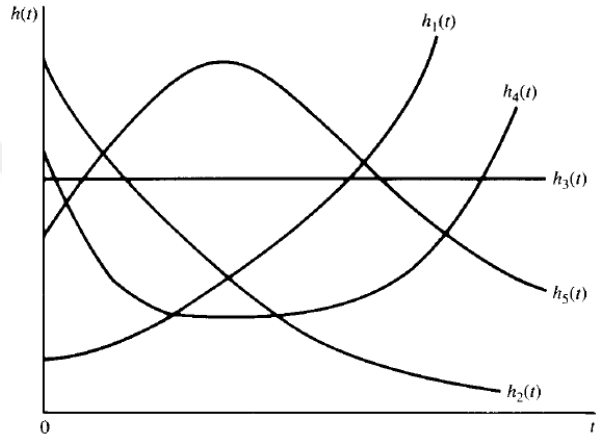
Hazard fonksiyonu aynı zamanda birikimli dağılım fonksiyonu F(t) ve olasılık yoğunluk fonksiyonu f(t) türünden de Denklem 2.7'de belirtildiği gibi tanımlanabilir.

$$h(t) = \frac{f(t)}{1-F(t)} \qquad \text{Denklem 2.7}$$

Hazard fonksiyonu, karmaşık davranışlarda bulunabilir. Bir olasılık fonksiyonu değildir. Artma, azalma durumunda olduğu gibi sabit de olabilir. Şekil 2.6'da farklı tipte hazard fonksiyonu belirtilmiştir. Bu fonksiyon litaratürde koşullu ölüm oranı, ani başarısızlık oranı, ölüm kuvveti (gücü) olarak da geçer. Yaşlanma süresince sağkalım analizinde önemli yeri olan hazard fonksiyonu birim zamana düşen başarısızlık riskini hesaplar (Lee, Wang, 2003). Hazard fonsiyonu olasılık fonksiyonu değildir, çıktısında oran elde edilir ve oran her zaman pozitif değerdedir (Klein, Moeschberger, 2003).

Hazard fonksiyonu, sağkalım analizinde, eğer sansürlü gözlem yoksa birim zamana düşen ölüm olasılığını tahmin eder ve belirli zaman aralığında başlangıçtaki hastaların sağ olduğu kabul edildiğinde hazard fonksiyonunun tahmini Denklem 2.8 ile hesaplanır.

$$\hat{h}(t)=\frac{\text{belirli zaman aralığında ölen hasta sayısı}}{\text{t zamanında sağkalan hasta sayısı}} \qquad \text{Denklem 2.8}$$



Şekil 2.6 Lee ve Wang (2003), yoğunluk eğrisi örnekleri

Şekil 2.6'da belirtilen farklı türdeki hazard fonksiyonlarından $h_1(t)$ ve $h_2(t)$ artan tehlike oranını, $h_3(t)$ sabit tehlike oranını göstermektedir. Burada $h_4(t)$ fonksiyonu insan yaşamının sürecini başlangıçta yüksek bebek ölümü ile başlayıp belirli zamana kadar sabit kalıp sonrasında zamanla yaşlanmaya bağlı olarak tehlike oranı artmaktadır. $h_5(t)$ fonksiyonu ise başlangıçta tehlike oranı başlangıçta artan sonrasında etkin tedavi ile tehlike oranı azalan durumu göstermektedir.

2.1.2.4 Birikimki (kümülatif) hazard fonksiyonu

Birikimli hazard fonksiyonu H(t), t zamanındaki başarısızlık olasılıklarının birikimli fonksiyonudur. Denklem 2.9'da hazard fonksiyonu, denklem 2.10'da ise birikimli hazard fonksiyonu formülize edilmiştir.

$$H(t) = \int_0^t h(x)dx \qquad \text{Denklem 2.9}$$

$$H(t) = -\log S(t) \qquad \text{Denklem 2.10}$$

Hazard fonksiyonu,

• limt→∞ H(t) = ∞ olarak hesaplanır.

• H(t) artan bir fonksiyondur.

• Sağdan süreklidir (Lee ve Wang 2003).

2.1.3 Sağkalım analizinde bazı önemli parametrik dağılımlar

Sağkalım analizinde parametrik model ile çalışmak için modelin oransal hazard varsayımını sağlaması gerekir. Bu varsayım sağlandığında parametrik modeller ile sağkalım analizi yapılabilir. Parametrik modellere ait dağılımlarda Üstel, Weibull, Lognormal, Gamma ve Loglojistik dağılımları sıklıkla kullanılır (Lawless, 2011).

2.1.3.1 Üstel Dağılım

Üstel dağılım çoğunlukla güvenilirlik ve sağkalım analizlerinde kullanılır ve bu dağılımda tek bir parametre yer alır. Bu parametre sabit hazard oranıdır ve λ olarak gösterilir. Sağkalım analizinde λ arttıkça risk artar, sağkalım azalır aynı şekilde λ azaldıkça risk azalır, sağkalım artar ve λ=1 olduğu durumda ise üstel dağılım gösterir.

Sağkalım zamanı *T,* parametresi λ olan üstel dağılıma sahip ise, olasılı yoğunluk fonksiyonu,

$$f(t) = \begin{cases} \lambda e^{-\lambda t}, t \geq 0, \lambda > 0 \\ 0\ ,\ t < 0 \end{cases}$$ Denklem 2.11

Kümülatif dağılım fonksiyonu,

$$F(t) = 1 - e^{-\lambda t}\ t \geq 0$$ Denklem 2.12

Sağkalımfonksiyonu,

$$S(t) = e^{-\lambda t}\ t \geq 0$$ Denklem 2.13

olarak tanımlanabilir. Buna göre, hazard fonksiyonu denklemdeki eşitlikle hesaplanabilir.

$$h(t) = \lambda\ \ t \geq 0$$ Denklem 2.14

Üstel dağılımda, sağkalım analizinde yer alan hastaların yaşı dikkate alınmadan sabit hazard oranına sahip olduğu kabul edildiği için, bu dağılımda ölüm veya başarısızlık olayı rastgele, zamandan bağımsız olduğu kabul edilir.

2.1.3.2 Weibull Dağılım

Üstel dağılımın genelleştirilmiş hali Weibull dağılımıdır. Burada Üstel dağılımdaki gibi sabit hazard oranı varsayımı gerektirmez ve bu nedenden dolayı daha uygulanabilir bir yöntemdir. Üstel dağılımda γ ve λ olarak gösterilen iki adet parametre yer alır. Burada Weibull dağılımının şeklini γ, ölçeklenmesini λ değeri tanımlar.
Olasılık yoğunluk fonksiyonu,

$$f(t) = \lambda\gamma(\lambda t)^{\gamma-1}\ e^{-(\lambda t)^{\gamma}} t \geq 0, \qquad \gamma, \lambda > 0$$ Denklem 2.15

Kümülatif dağılım fonksiyonu,

$$F(t) = 1 - e^{-(\lambda t)^{\gamma}}$$ Denklem 2.16

Sağkalım fonksiyonu ise,

$$S(t) = e^{-(\lambda t)^{\gamma}}$$ Denklem 2.17

ile hesaplanır. Bu durumda hazard fonksiyonu denklem eşitliğindeki gibi olacaktır.

$$h(t) = \lambda\gamma(\lambda t)^{\gamma-1} \;\; t \geq 0$$ Denklem 2.18

2.1.3.3 Lognormal Dağılımı

Sağkalım süresi T ise, logaritmik T değerinin normal dağılım göstermesi Lognormal Dağılım olarak adlandırılır. Lognormal dağılımında iki adet parametre yer alır.

μ ortalaması, σ^2 varyansa sahip T'nin olasılık yoğunluk fonksiyonu denklem 2.19'da belirtildiği gibidir.

$$f(t) = \frac{1}{t\sigma\sqrt{2\pi}} exp\left[-\frac{1}{2}\left(\frac{logt-\mu}{\sigma}\right)^2\right], \; t > 0$$ Denklem 2.19

Sağkalım fonksiyonu,

$$S(t) = \frac{1}{t\sigma\sqrt{2\pi}} \int_t^{\infty} \frac{1}{x} exp\left[-\frac{1}{2\sigma^2}(logx - \mu)^2\right] dx$$ Denklem 2.20

Hazard fonksiyonu,

$$h(t) = \frac{f(t)}{S(t)}$$ Denklem 2.21

2.1.3.4 Gamma Dağılımı

Gamma dağılımında iki adet parametre yer almaktadır. Bu parametreler λ and k olarak belirtilir ve k=1 olduğu durumda Gamma dağılımı Üstel dağılımı içerir.

Olasılık yoğunluk fonksiyonu,

$$f(t) = \frac{\lambda\gamma(\lambda t)^{k-1}\, e^{-\lambda t}}{\Gamma(k)} \qquad t \geq 0,\ k > 0,\ \lambda > 0 \qquad \text{Denklem 2.22}$$

Sağkalım fonksiyonu,

$$S(t) = \frac{\lambda\gamma(\lambda t)^{k-1}\, e^{-\lambda t}}{\Gamma(k)} \qquad \text{Denklem 2.23}$$

Buna göre hazard fonksiyonu,

$$\lambda(t) = \frac{f(t)}{S(t)} \qquad \text{Denklem 2.24}$$

eşitliğinden hesaplanabilir. Ve hazard fonksiyonu,

• k > 1 ise monoton şekilde artar,

• k = 1 ise sabittir,

• k < 1 ise, monoton olarak azalır. (Rodrıguez, 2010).

2.1.3.5 Loglojistik Dağılım

Loglojistik dağılımı weibul hazard fonksiyonu kısıtlı oldupu zamanlarda kullanılabilir. Bu dağılımda iki adet parametre yer alır.

t≥0 , α>0 ve $\gamma > 0$ iken, Olasılık yoğunluk fonksiyonu,

$$f(t) = \frac{\alpha\,\gamma\,t^{\gamma-1}}{(1+\alpha t^{\gamma})^2}$$ Denklem 2.25

Sağkalım fonksiyonu,

$$S(t) = \frac{1}{1+\alpha t^{\gamma}}$$ Denklem 2.26

Hazard fonksiyonu,

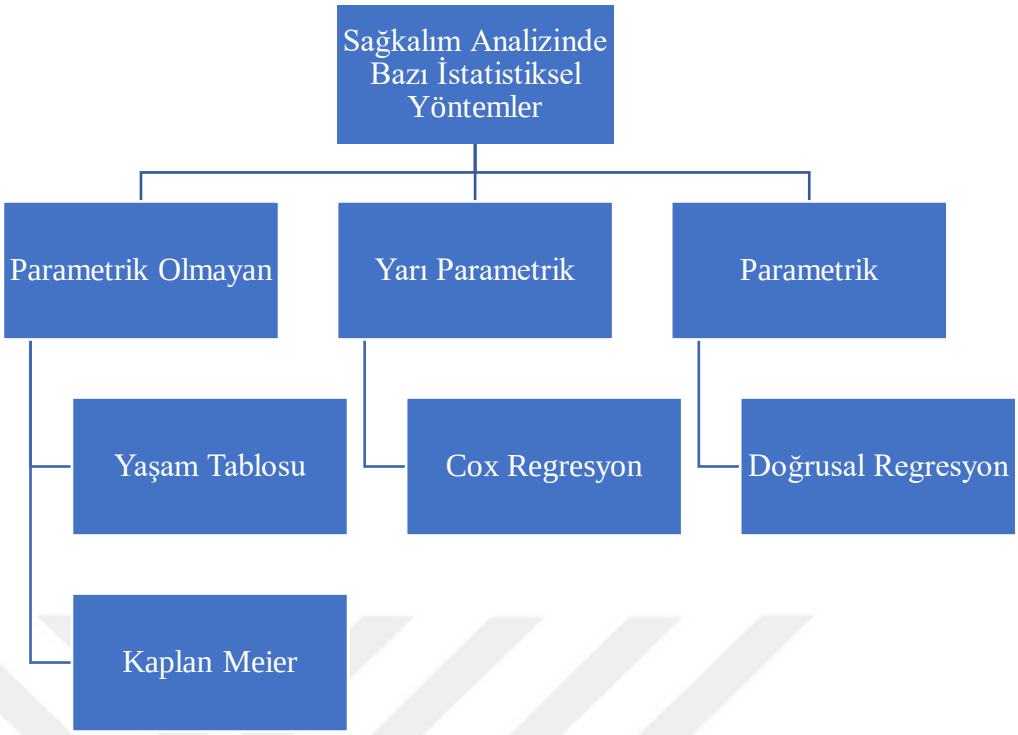
$$h(t) = \frac{\alpha\,\gamma\,t^{\gamma-1}}{1+\alpha t^{\gamma}}$$ Denklem 2.27

2.1.4 Sağkalım analizinde istatistiksel yöntemler

Sağkalım analizinde yer alan bazı istatistiksel yöntemler şekil 2.7'de gösterilmiştir.

Parametrik olmayan yöntemler, sağkalım süresinin dağılımı belirli değilse veya orantısal hazard varsayımlarını karşılamıyorsa uygulaması etkili olur. Kaplan-Meier (KM) yöntemi, Yaşam Tablosu (YT) yöntemlerinden biri uygulanarak sağkalım fonksiyonu tahmin edilir (Lee ve Wang 2003).

Yarı parametrik yöntemlerde sıklıkla uygulanan yöntem cox regresyon, parametrik yöntemlerde sağkalım süresinin dağılımı bilindiği durumda en sık doğrusal regresyon uygulanmaktadır (Lee ve Wang 2003).



Şekil 2.7 Sağkalım Analizinde yer alan bazı istatistiksel yöntemler

2.1.4.1 Sağkalım Analizinde parametrik olmayan yöntemler

2.1.4.1.1 Yaşam Tablosu

Berkson ve Gage (1950) ekibi ile başlatılan, Cutler ve Ederer (1958) ekibi ile de uygulamaya koyulan yaşam tablosu yöntemi, çalışmada yer alan denek sayısı fazla ise uygulanır (Matthes ve Farewell, 1988).

Yaşam Tablosu analizinde, sansürlü veriler dikkate alınır. Çalışmada yer alan veriler adından da anlaşıldığı gibi tabloya girilip, analiz edilir ve tabloya girilen veriler aynı zamanda eşit aralıklarla gruplandırılmış olmalıdır (Pagano, Gauvreau, 2018). Aynı zamanda yaşam tablosu analizinde her deneğin eşit ölüm riskine sahip olduğu varsayılır.

Veri sayısı 100'den büyük ise yaşam tablosu yöntemi tercih edilir ve bu yöntemde sağkalım süresi eşit aralıklı, sağkalım sürelerinin eşit ve tekrarlı olduğu durumlarda kullanılan yöntemdir. Bu yöntem ile sağkalım ve ölüm olasılıkları, yaşam süresi ortalaması hesaplanır.

Yaşam tablosunda sınıfların aralıklarının eşit olduğu ve aynı zamanda eşit ölüm riskine gösterdiği varsayıldığında, i'inci aralıkta yer alan istenen olayın (ölüm) gerçekleştiği

hastaların dışındaki bireyler, bir sonraki i+1'inci sınıfa taşınırken aynı sağkalım olasılığı ile yerleşirler. i'inci aralıktaki hasta sayısı r_i Denklem 2.28 ile hesaplanır.

$$r_i = n_i - \frac{c_i}{2} \qquad \text{Denklem 2.28}$$

n_i = i.inci aralığın başlangıcındaki hasta sayısı

c_i = i.aralıkta yaşayan hasta sayısı (sansürlü veya sansürsürz)

Yaşam tablosu yönteminde eşit aralıktaki hastaların yer aldığı aralıkta eşit sağkalım olasınığına sahip olduğu varsayımına göre, sağkalan hastaların sınıf orta noktasında riske uğradığı da varsayılmıştır. Buna göre i'inci aralıkta yer alan hastaların ölüm riski $\hat{q}_i$ Denklem 2.29'da belirtildiği gibi hesaplanır. Burada d_i = i'inci aralıktaki ölen hasta sayısıdır.

$$\hat{q}_i = \frac{d_i}{n_i} \qquad \text{Denklem 2.29}$$

Aynı aralıktaki yaşam olasılığı ise Denklem 2.30 ile hesaplanır.

$$\hat{p}_i = 1 - \hat{q}_i \qquad \text{Denklem 2.30}$$

i'inci aralıkta sağkalan hastaların bir sonraki i+1'inci aralığa yaşayan hasta olarak geçtiği varsayılıp araştırmaya dâhil edildiği durumda, $\hat{S}(t_0) = 1$ olarak hesaplanır ve $\hat{S}(t_i)$'nin standart hatası Denklem 2.31 ile hesaplanır.

$$Std.Hata(\hat{S}(t_i)) = \hat{S}(t_i)\sqrt{\sum_{j=1}^{i} \frac{\hat{q}_i}{\hat{p}_j r_j}} \qquad \text{Denklem 2.31}$$

Bu yöntemde her bir aralıkta hazard fonksiyonu Denklem 2.32 ile hesaplanır.

$$\hat{h}(t_{mi}) = \frac{2\hat{q}_i}{b_i(1+\hat{p}_i)}$$ Denklem 2.32

b_i= sınıf aralığı

$\hat{q}_i$= ölüm olasılığı

$\hat{p}_i$= yaşam olasılığı

t_{mi}= aralığın orta noktası'dır.

Ölüm olasılık yoğunluk fonksiyonu Denklem 2.33 ile hesaplanır.

$$\hat{f}(t_{mi}) = \frac{\hat{S}(t_i) - \hat{S}(t_{i+1})}{b_i}$$ Denklem 2.33

Bu durumda medyan yaşam süresi Denklem 2.34 ile hesaplanabilir.

$$\theta_i = t_i + \frac{\hat{S}(t_{i-1}) - 0.5}{\hat{f}(t_{mi})}$$ Denklem 2.34

2.1.4.1.2 Kaplan Meier

Parametrik olmayan yöntemlerde en pratik yöntem Kaplan-Meier yöntemidir. Sebebi ise, Yaşam Tablosu'nuda çeşitli sebeplerden sansürlü olan verilerin olasılıkları işlemlere dahil edilirken, Kaplan-Meier yönteminde, sarsürlü verilerin olasılıkları hesaplamalara dahil edilmez (Cox, 1972).

Kaplan-Meier sağkalım eğrisinde, sağkalım süresi birçok küçük aralıktaki zamanı göz önünde bulundurarak belirli bir süre içinde hayatta kalma olasılığı olarak tanımlanır (Altman, 1990).

Kaplan Meier analizinde üç adet varsayım vardır,

- Sansürlü denekler, gözlemlenmeye devam eden deneklerle benzer hayatta kalma olasılığına sahip olduğu varsayılır.
- Sağkalım analizine erken veya geç giren deneklerin sağkalım olasılıklarının eşit olduğu varsayılır.
- İlgilenilen olay kesindir ve açıkça belirlenmiş bir zamanda gerçekleşir (Lawless, 2011).

Gerçekte, koşulsuz bir şekilde sağkalım fonksiyonunu kesin olarak hesaplayamayız. Bu nedenle de Kaplan-Meier tahmincisi ile çalışmada yer alan deneklerden elde edilen verilerle gerçek hayatta kalma fonksiyonunu tahminleriz.

K adet birimin gözlenen sağkalım süresi $t_1 < t_2 < t_3 < \cdots < t_k$ ile belirtilmiş olsun. Bu durumda sağkalım fonksiyonu Denklem 2.35 ile hesaplanır (Lawless, 2011).

$$\hat{S}(t) = \prod_{i:t_i \leq t} \left(1 - \frac{d_i}{n_i}\right) \qquad \text{Denklem 2.35}$$

Denklemde,

t_i olayın meydana geldiği zaman,

d_i, t_i anında oluşmuş olayların adedi,

n_i, t_i anına kadar hayatta kaldığı kesin olan deneklerin adedi olarak denkleme dâhil edilir. Buna göre t_i=0 anında S(0)=1 olacaktır.

Kaplan-Meier sağkalım fonksiyonunun standart hatası ise Denklem 2.36 ile hesaplanır (Lawless, 2011).

$$Std.Hata\ \left(\hat{S}(t)\right) = \hat{S}(t_i)\sqrt{\sum_{j=t_i}^{t} \frac{d_i}{n_i(n_i+1)}} \qquad \text{Denklem 2.36}$$

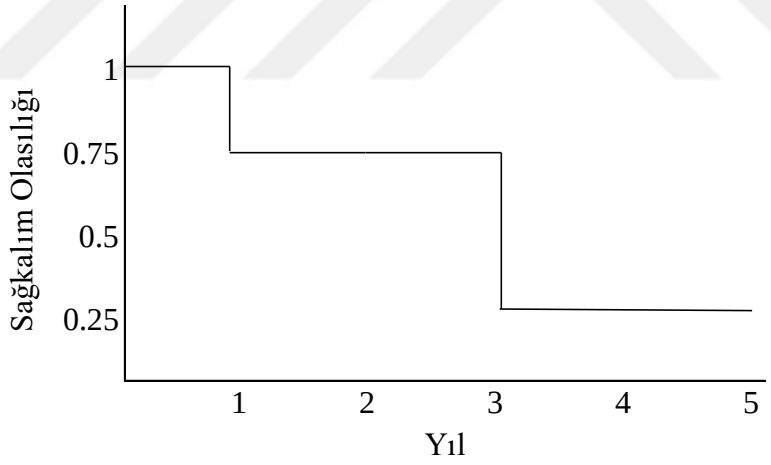
Kümülatif hazard fonksiyonu ise Denklem 2.37'de belirtildiği gibi hesaplanır (Lawless, 2011).

$$\widehat{H}(t) = \sum_{i:t_i \le t} \frac{d_i}{n_i}$$ Denklem 2.37

Sağkalım izlem süresince izlenen hastalardan süresi $t_1 < t_2 < t_3 < \cdots < t_N$ sırasıyla N adet ölüm gerçelleştiği varsayıldığında, Ortalama sağkalım süresi tahmin edicisi Denklem 2.38 ile hesaplanır.

$$\mu = t_1 + \sum_{i=1}^{K-1} Y(t_i)(t_{i+1} - t_i)$$ Denklem 2.38

Kaplan-Meier tahminleri, yaşam tablosu sağkalım eğrisine benzer şekilde deneklerin sağkalım olasılıklarını çizmek için kullanılabilir. Sağkalım eğrisi, olayın gerçekleştiği kesin anda düşer ve bir sonraki olayın gerçekleşeceği zamana kadar sabit kaldığı için düzgün bir eğri eğiliminde değil, kademeli çizgi biçimindedir. Kaplan-Meier sağkalım grafiği eğrisi Şekil 2.8'de gösterilmiştir.



Şekil 2.8 Kaplan-Meier sağkalım grafiği

Şekil 2.8 incelendiğinde, 1. yılda sağkalım olasılığı 1'e eşit, 5.yılsonunda ise sağkalım olasılığı 0.25 olmuştur. Grafikte düşüşün gözlendiği her zamanda ölüm gerçekleşmiştir.

2.1.4.2 Sağkalım eğrilerinin karşılaştırılması

Farklı tedavi yöntemleri uygulanan hasta gruplarında sağkalım süresini tahmin etmek kadar aynı hastalığa sahip bireylerde hastalığın sağkalım süresini tahmin etmek de önemlidir. Örneğin, kanser hastalarında hastalığın evreleri açısından sağkalım oranları karşılaştırılmak istenebilir.

Sağkalım analizinde eğer sansürlü veriler olmasaydı, iki grup karşılaştırılmasında için Mann-Whitney U testi ve 2'den fazla gruplar için ise Kruskal-Wallis testi kullanılabilirdi. Fakat sağkalım analizinde sağkalım süreleri sansürlü veriler içerdiği için bu yöntemler uygulanamayacaktır. Bu nedenle böyle durumlarda sağkalım eğrilerinin karşılaştırılması için en yaygın kullanılan testler Log-rank testi, Breslow-Wilcoxon testidir ve her ikisi de iki ve ikiden fazla gruplara ait sağkalım eğrilerini karşılaştırmada kullanılır (Lee & Wang, 2003).

2.1.4.2.1 Long-Rank Testi

Kaplan Meier yöntemi sağkalıma ait olasılıkları grafik yardımıyla görselleştirilir. Birden fazla grubun sağkalım olasılıkları karşılaştırılmak istendiğinde Log-rank testi uygulanır.

Bu testde karşılaştırılan gruplara ait hazard oranlarının her sağkalım döneminde aynı olduğu varsayılmıştır (Yayla, 2013). Log-rank testi Ki-Kare testinin genişletilmiş hali olduğu söylenebilir. Ki-kare istatistiği, bu test yönteminde Kaplan Meier testi eğrilerini karşılaştırmada kullanılır (Kleinbaum, Klein, 2010). Bu durumda,

H_0 : Grupların sağkalım eğrileri benzerdir.

H_1 : En az bir sağkalım eğrisi farklıdır, hipotezleri test edilirken, İkiden fazla grup için Log-Rank test istatistiği,

$$X^2 = \sum_i^k \frac{(\text{Gözlenen}_i - \text{Beklenen}_i)}{\text{Beklenen}_i} \qquad i=1,2,\ldots,k \qquad \text{Denklem 2.39}$$

Hesaplanan X^2 değeri, $X^2_{i-1,\alpha}$ değerinden küçük ise H_0 hipotezi reddedilemez, grupların sağkalım eğrileri benzerdir denilir.

2.1.4.2.2 Breslow-Wilcoxon testi

Logrank testine benzer şekilde, bu test her gözleme bir puan vermektedir. Wilcoxin testinin genelleştiril halidir ve benzer şekilde gözlemelere verilen puanların toplamı test istatistiği değeridir. Hesaplanan istatistik değeri eğer Z_α değerinden büyük ise "İki grup arasındaki medyan farkı sıfırdır" H_0 hipotezi reddedilir. Yani sağkalım eğrileri benzer değildir sonucuna varılır (Lee ve Wang, 2003).

$$Test\ istatistiği = \frac{[\sum jw(t_j)*(m_{ij}-e_{ij})]^2}{var[\sum jw(t_j)*(m_{ij}-e_{ij})]} \qquad \text{Denklem 2.40}$$

ile hesaplanır. Formülde w_{tij} ile belirtilen ağırlıklandırma Wilcoxon testinde, j zamanındaki risk altında olan birey sayısı kadar olmaktadır. Bu yöntemde eğer uygulanan tedavi yönteminin etkinliği zamanla azalıyorsa kullanılması daha uygun olmaktadır.

2.1.4.3 Sağkalım Analizinde yarı parametrik ve parametrik yöntemler

Sağkalım analizinde parametrik yöntemler ile çeşitli dağılımlar uygulanarak sağkalım ve hazard (risk) fonksiyonu tahmin edilir (Lee & Wang, 2003).

Bazı önemli parametrik dağılımlar Üstel, Weibull, Gamma, Lognormal, Loglojistik, Gompertz ve Genelleştirilmiş Gamma dağılımıdır. (Klein ve Moeschberger, 2003).

2.1.4.3.1 Cox Regresyon

Parametrik modellere göre daha az varsayıma sahip olan Cox regresyon modeli, yarı parametrik bir tekniktir. (Breslow, 1975).

Cox regresyon yönteminde, denekler birbirinden bağımsız, risk oranının sabit iken,

(1) Bağımsız olan değişkenlerin hazard fonksiyonu üstündeki etkileri loglineerdir.

(2) Bağımsız olan değişkenlere ait loglineer fonksiyonu ile hazard fonksiyonu arasındaki yer alan ilişki çarpımsaldır (Cox, 1972). Cox regresyon modeli çoklu değişken olması durumunda Denklem 2.41'deki gibi yazılabilir,

$S(t|X) = h_0(t)exp(\beta^t X) = h_0(t)exp\left(\sum_{k=1}^{p} \beta_k X_k\right)$ Denklem 2.41

Burada $h_0(t)$, temel hazard oranı, β_k regresyon analizindeki bağımsız değişkenlere ait katsayılar, X_k ise bağımlı değişkendeki değişmelerin açıklanması için modelde yer alan, bağımsız değişkenlerdir.

Hazard oranı, bir gözlemin tehlikesinin başka bir gözlemin tehlikesine oranıdır (Kleinbaum & Klein, 2012) ve Denklem 2.42'deki formülle hesaplanır.

$$\frac{h(t|X)}{h(t|X^{\cdot})} = \frac{h_0(t)exp\left[\sum_{k=1}^{p} \beta_k X_k\right]}{h_0(t)exp\left[\sum_{k=1}^{p} \beta_k X_k^{\cdot}\right]} = \exp\left[\sum_{k=1}^{p} \beta_k (X_k - X_k^{\cdot}\right] \quad \text{Denklem 2.42}$$

Sağkalım verilerinde sağkalım süresi arttıkça bağımsız değişkenlerin değerleri de değişir ve bu doğrultuda da sağkalım hazard oranı değişir. Tabakalandırılmış veya genişletilmiş Cox regresyon modeli, orantılı hazard oranı varsayımı sağlanmadığı durumda Cox regresyon yerine kullanılır (Kleinbaum & Klein, 2012).

2.1.4.3.1.1 Orantısal hazard değerlendirilmesi

Cox regresyon yönteminin en önemli varsayımlarından biri olan "oransal hazard varsayımı", hazard oranının zamana karşı sabit veya sağkalım analizinde yer alan bir bireyin hazardı ile diğer bireyin hazardı birbiriyle orantılı olması anlamına gelmektedir.

Cox regresyon modelinin ana varsayımı, hazard oranının sağkalım süresi içinde sabit olduğu anlamına gelen orantılı hazard oranı varsayımıdır. Cox regresyon modelinin orantılı hazard oranı varsayımını karşıladığının testi için kullanılan yöntemler aşağıda belirtilmiştir (Collett, 1994).

Grafiksel yöntem'de iki grup için tahmini -log(-log(survival)) hesaplanır ve buna göre sağkalım süresinin grafiğini çizilir. İki gruba ait hazard oranı orantılıysa paralel eğriler oluşması beklenir (Kleinbaum, Klein, 2012). Fakat bu yöntem, eğer veriler kategorik ise daha karmaşık hale geldiğinden, düzgün çalışmayabilir.

Ölçeklendirilmiş Schoenfeld artıkları yöntemi'nde kovaryans matrisli ve matrise dayalı ölçeklenmiş olan Schoenfeld artıklarının grafiği çizilir. Çizilen grafik eğer yatay bir doğru etrafında dağılım gösteriyorsa orantısal hazard oranı varsayımını karşılamaktadır (Schoenfeld, 1982).

Zamana bağlı ortak değişken ekleme yönteminde Cox modelinde yer alan değişkenler, zamana bağlı olarak genişletilir. Cox regresyon analizi ile zamana bağlı olarak genişletilmiş ortak değişkenin katsayısı istatistiksel olarak anlamlı belirlenemez ise orantısal hazard oranı varsayımı karşılanır (Kleinbaum, Klein, 2012).

2.1.4.3.2 Doğrusal Regresyon

Doğrusal regresyon, lojistik regresyon ve Poisson regresyonu, sağlık alanında sıklıkla tercih edilen parametrik modellerdendir. Bu modellerde sonucun normal, binom veya Poisson dağılımı gibi bazı dağılımları takip ettiği varsayılır. Parametrik regresyon modelleri için, veriler tipik olarak, o dağılımı tam olarak belirten parametrelerin değerlerini tahmin etmek için kullanılır. (Kleinbaum, Klein, 2010)

2.1.5 Model Seçim Kriterleri

Sağlık bilimleri çalışmalarda, sağkalım analizine ilk başlanıldığında birçok değişken analizde yer alabileceği gibi tahmin hatasının artmasına neden olacak sahte ortak değişkenlerde yer alabilmektedir. İstatistiksel sağkalım modellemesinde ortak değişkenlerden hangisinin seçileceğine karar vermek, sağkalım analizi için her zaman zordur. Bu nedenle AIC (Akaike, 1974), BIC (Schwarz, 1978) ve Cp (Mallows, 1973) gibi geleneksel uygulamalar, sağkalımda en uygun bir modeli seçmek için sıklıkla kullanılmaktadır. Bu AIC ve BIC uygulamaları değerlendirirken, değişken seçiminde 3 farklı yöntem olan, ileriye doğru, geriye doğru ve adımsal seçme yöntemleri uygulanabilir (Lee, Wang, 2003).

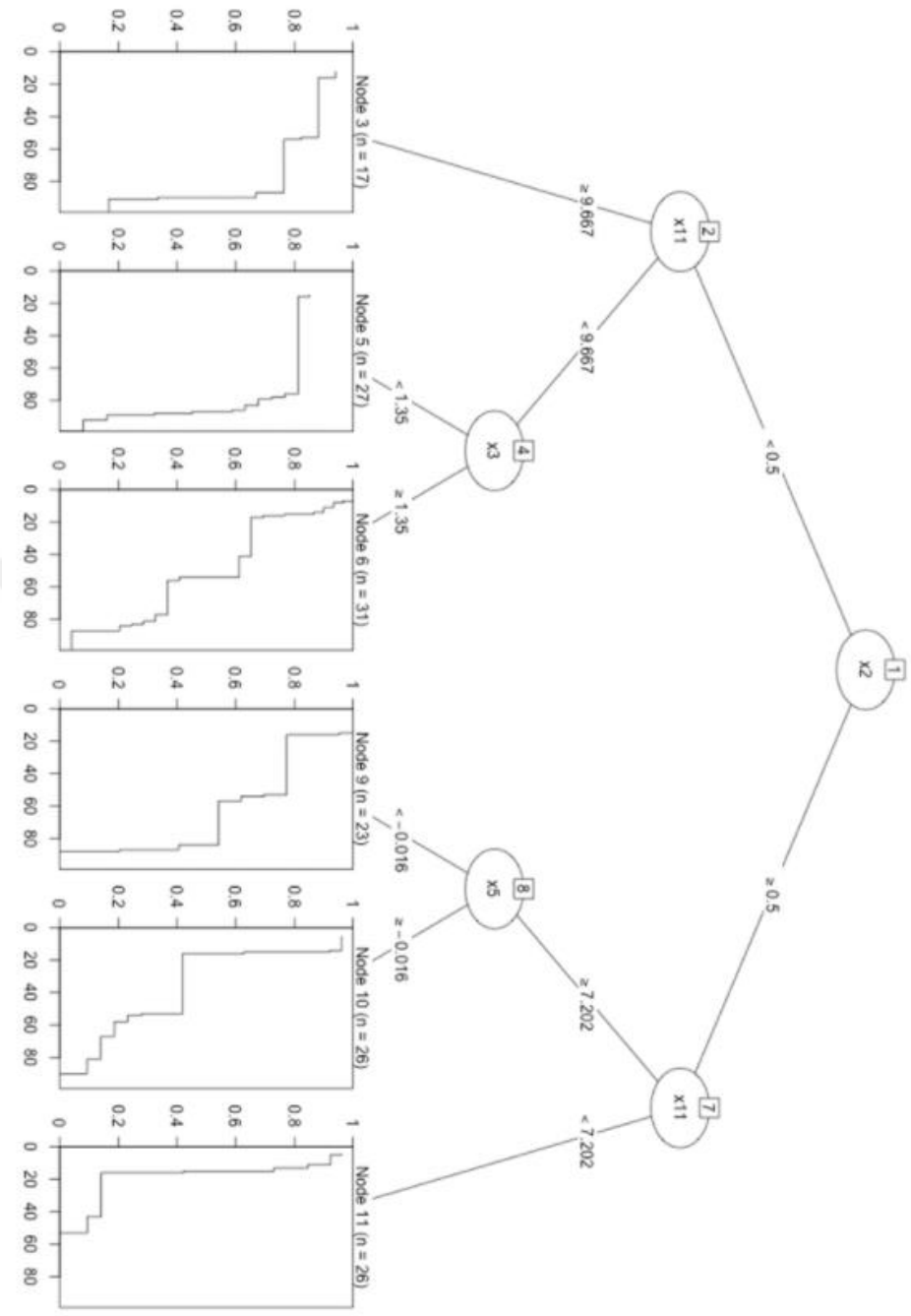
Sağkalım analizi için model seçiminde herhangi bir koşul yoksa en düşük -2logL veya AIC sonuçlarının yer aldığı model uygun olabilir. Modelde hangi açıklayıcı değişkenin yer almasının uygun olacağının belirlenmesinde ileriye doğru seçim yöntemi (forward selection), geriye doğru seçim yöntemi (backward elemination), giriş (Enter) yöntemi olmak üzere üç yöntem vardır (Kaygısız, 2010).

2.2 Rasgele Sağkalım Ormanlar (Random Survival Forest) Yöntemi

Sağkalım analizine rastgele orman yaklaşımı (Breiman 2001), oransal hazard gibi kısıtlayıcı varsayımları olan klasik yöntemlere alternatif yöntem sağlar. Bu yaklaşım, parametrik veya yarı parametrik temelindeki dağılımlardaki kısıtlamalara ihtiyacı gerektirmez ve değişkenlerin birbiriyle olan etkileşimlerle otomatik olarak başa çıkar ve doğru tahminini sağlar (Ishwaran, Kogalur, Blackstone, Lauer,2008).

Rastgele Orman yöntemi, sağkalım analizinde yer alan sağdan sansürlü verilerle ağaç yöntemi kullanılarak yapılan sağkalım analizidir.

Rastgele sağkalım ormanları yönteminde, rasgeleleştirme iki şekilde sağlanır. İlk olarak, bir ağaç büyütmek için verilerin rastgele çizilmiş bir önyükleme örneği kullanılır. İkinci olarak ise, ağacın her bir düğümünde aday değişken olarak rastgele seçilen değişkenleri bölünen düğümlere yerleştirir.



Şekil 2.9 Rastgele sağkalım orman örneği
doi: https://doi.org/10.1371/journal.pone.0250963.g001

Şekil 2.9'da yer alan sağkalım ağacı örneğinde, n = 150 adet gözlem vardır, 1. Düğüm kök düğümüdür, algoritma gözlemleri ikili bir değişken (0,1) olan x2 ortak değişkenini iki çocuk düğüme böler. Bu çocuk düğümlerinde karar verme oluşmaz. Daha sonrasında ortaya çıkan çocuk düğümlerin her biri, sırasıyla 9.667 ve 7.202'de x11 ortak değişkeni tekrar bölünür. Çocuk düğümler kök düğüme göre daha homejendir.

Bu işlem ayırma kriterine göre terminal düğümlere bölünür. Şekil 1'deki ortaya çıkan hayatta kalma ağacında, 6 adet terminal düğüm oluşmuştur. Şekil 1'de verilen sağkalım eğrileri, birbirinden farklı ortak değişkenlerin, sağkalım olasılığı üzerindeki etkisi görselleştirilmiştir. Parametrik olmayan hayatta kalma eğrileri bu 6 adet terminal düğümlerin karşılaştırılmasında kullanılır (Whetten, Stevens, Cann, 2021).

2.2.1 Bölme Kuralları

Rastgele sağkalım orman yöntemi, belirli bir olay meydana gelinceye kadar geçen sağdan sansürlü zaman verilerinin analizi için ağaçların bir araya getirilmesi ve analiz edilmesi yöntemidir. Bu yöntem, sağkalım analizlerinde yarı parametrik veya parametrik modellere alternatif olarak uygulanan en popüler parametrik olmayan yöntemlerden birisidir.

Rastgele sağkalım orman yönteminde, sağkalım ağaçları sınıflandırma ve regresyon ağaçlarında benzer şekilde oluşturulur. İlk olarak önyüklemede yer alan tüm verileri kapsayan sağkalım ağacının tepesindeki kök düğümle başlanılır. Bu önyükleme örneğinden, p adet bağımsız değişkenler belirlenir ve daha önceden belirlenen sağkalım kriterine göre bu kök düğümü iki çocuk düğüme bölmek için kullanılır. İki çocuk düğümün her biri için, öncesinden farklı p tahmin değişkeni kümesi rastgele belirlenir ve iki çocuk düğümü kendi içlerinde tekrardan iki çocuk düğüme bölmek için kullanılır. Bu işlem, her düğüm için en az bir adet gözlenen olay gerçekleşene kadar tekrarlanır (Weathers, 2017).

Rastgele sağkalım orman yönteminde h düğümü için önerilen bir bölme bulmak istediğimizi varsayalım. Rastgele seçilmiş p yordayıcı değişkenler kümesinden seçilen, belirli bir kategorik x tahmin değişkeni seçilir, iki eşitsizlik arasındaki sağkalım farkları, $x \leq c$ (bir bireyin çocuk düğüm 1'e yerleştirilmesi) ve $x > c$ (bir bireyin çocuk düğüm 2'ye yerleştirilmesi) arasındaki hayatta kalma farklarını maksimize edecek bir c değeri bulmak istiyoruz. Bunun için öncelikle rastgele seçilen p yordayıcı değişkenler kümesinden bir x seçilir, bölünmüş değer (c) belirlendikten sonra iki kardeş düğüme bireyler atanır. Önceden belirlenmiş bir bölme yöntemi kullanarak ve x için hayatta kalma farklarını en üst düzeye çıkaran c değerini bulana kadar işlemi başka bir c ile tekrarlayarak iki grup arasındaki hayatta kalma farkı hesaplanır. Bu

işlem iki çocuk düğüm arasındaki hayatta kalma farkını maksimize eden x' ve c' bulana kadar diğer p-1 adet değişkenler için tekrarlanır (Ishwaran and Kogalur, 2008).

Rastgele Ormanlar yaklaşımı, ilk adımında, bölme değeri tamamen rastgele seçilir. Rasgele sağkalım orman yöntemi ilk adımdaki bölme işleminde kullanılan kurallar Log-rank, Olayları koruma, Log-rank skor ve Yaklaşık Log-rank bölme kuralladır.

Bölme kurallarında yer alan notasyonlar,

- h: Bir ağacın h. Düğümü
- n: h düğümündeki bireylerin sayısı
- T_i: i.inci birey için sağkalım süresi
- σ_i: i. birey için sansürleme durumu, i.inci birey sağdan sansürlenirse $\sigma_1 = 0$ ve i.inci birey öldüyse $\sigma_1 = 1$.
- *x*: Düğüm bölme için bir aday tahmincisi
- *c*: x tahmin değişkeni için bölme değeri
- x^*: Çocuk düğümler arasındaki sağkalım farklarını maksimize eden tahmin edici değişken
- c^*: Tahmin değişkeni x için çocuk düğümler arasındaki sağkalım farklarını maksimize eden bölünmüş değer
- j: Çocuk düğüm, j € {1, 2}
- t_N: h düğümündeki farklı olay süreleri
- $Y_{i,j}$: J düğümünde t_i zamanında risk altında olan (canlı) veya olay (ölüm) olan bireyler
- $d_{i,j}$: j çocuk düğümündeki t_i zamanındaki olay sayısı (Ishwaran, Kogalur, 2008).

2.2.1.1 Log-rank bölme kuralı

Log-rank bölme kuralında, log-rank test istatistiğini maksimize ederek düğümlere bölünür. Bu yöntemin olumsuz yönü, tahmin edicileri tercih ederken daha fazla sayıda bölünme noktasına olanları tercih etmesidir. Bağımsız değişkenlerinin $n_2<n_1$ iken x_1 ve x_2, bağımlı değişkeni y olan bir veri kümesinde, x_1 daha büyük bir bölme noktasına sahip ve y üzerinde daha büyük bir etkiye sahip olması olası olacaktır. Bir bölmede x bağımsız değişkeni için log-sıra istatistiği c değeri Denklem 2.43 ile hesaplanır ve bu

yöntemde amaç, log rank testinin en büyük büyüklüğünü veren x ve c'yi bulmaktır. (Ishwaran, Kogalur, 2008).

$$L(x,c) = \frac{\sum_{i=1}^{N}(d_{i,1} - Y_{i,1}\frac{d_i}{Y_i}}{\sqrt{\sum_{i=1}^{N}\frac{Y_{i,1}}{Y_i}\left(1-\frac{Y_{i,1}}{Y_i}\right)\left(\frac{Y_i-d_i}{Y_i-1}\right)d_i}} \qquad \text{Denklem 2.43}$$

2.2.1.2 Yaklaşık Log-rank bölme kuralı

İkinci yöntem, Log-rank yönteminin bir yaklaşımdır ve bu nedenle, yaklaşık log-rank bölme olarak adlandırılır. Log-rank yöntemindeki L(x, c) denkleminin payını yaklaşık olarak bulmak için, üst düğüm için Nelson-Aalen birikimli tehlike tahmincisi kullanılarak bir düzenleme yapılır. (Ishwaran, Kogalur, 2008).

Nelson-Aalen tahmincisi,

$$\widehat{H}(t) = \sum_{t_i<t}\frac{d_i}{Y_i} \qquad \text{Denklem 2.44}$$

LeBlanc ve Crowley (1993)'de gösterildiği gibi,

Log-rank yöntemindeki $L(x,c)$ denkleminin payında $D_J = \sum_{i=1}^{N} d_{i,j}$ j =1,2 için Denklem2.45'deki gibi hesaplama yapılır.

$$\sum_{i=1}^{N}(d_{i,1} - Y_{i,1}\frac{d_i}{Y_i}) = D_J - \sum_{l=1}^{n} I\{x_1 \le c\}\,\widehat{H}(T_l) \qquad \text{Denklem 2.45}$$

D = $\sum_{i=1}^{N} d_i$ olarak ayarlandığında, log-rank testi L(x, c) için aşağıdaki yaklaşıklığı elde ederiz:

$$\frac{D^{1/2}\left(D_1 - \sum_{l=1}^{n} I\{x_l \le c\}\hat{H}(T_l)\right)}{\sqrt{\{\sum_{l=1}^{n} I\{x_l \le c\}\hat{H}(T_l)\}\{D - \sum_{l=1}^{n} I\{x_l \le c\}\hat{H}(T_l)\}}} \qquad \text{Denklem 2.46}$$

2.2.1.3 Log-rank skor bölme kuralı

Log-rank skor bölme kuralı, log-rank bölme kuralının geliştirilmiş versiyonudur. Verilen sağkalım sürelerinin sıra vektörünü $r = (r_1,r_2,...,r_N)$ varsayılsın, bu verilerin indikatör değişkenleri $(T,\sigma)=((T_1,\sigma_1),(T_2,\sigma_2)...(T_N,\sigma_N))$ ve r vektöründeki sıralara bağlı olarak puan vektörü de $a = a(T,\sigma)=(a_1(r),a_2(r),...,a_N(r))$ şeklinde ifade edilir. Sıralamaların tahmin değişkenlerini $x_1 < x_2 < ... < x_N$ .olacak şekilde sıraladığını varsayalım, T_l 'deki bir gözlem için log-rank puanı (Ishwaran, Kogalur, Blackstone, Lauer,2008).

$$a_i = a_i(T,\sigma) = \sigma_i - \sum_{k=1}^{Y_i(T)} \frac{\sigma_k}{N-\gamma_k(T)+1} \qquad \text{Denklem 2.47}$$

2.2.1.4 Olayları koruma bölme kuralı

Bu yöntemde tahmin edilen birikimli hazard fonksiyonu, sağkalım analizinde süresince gözlemlenen tüm olaylar ve sansürlü zamanlar üzerinden toplandığında fonksiyondaki toplam olay sayısına eşit olması gerekmektedir. Bu durum, ana düğüm yerine her bir çocuk düğüm için hesaplanan Nelson Aalen tahmincisinin değiştirilmiş bir türü kullanılarak yapılır. Bu tahmin edici aşağıdaki formülle verilir:

$$\widehat{H}_j(t) = \sum_{t_{i,j} \le t} \frac{d_{i,j}}{Y_{i,j}} \qquad \text{Denklem 2.48}$$

$t_{i,j,}$ burada j çocuk düğümü için verilen sıralı olay zamanıdır. Her bir j çocuk düğümü için toplam olay sayısı 2.4 formülü ile hesaplanır.

$$\sum_{l=1}^{n_j} \widehat{H}_j(T_{l,j}) = \sum_{l=1}^{n_j} \delta_{l,j} \qquad \text{Denklem 2.49}$$

Daha sonra j çocuk düğümü için sıralı olay sürelerini $T_{(1),j} \le \ldots \le T_{(nj),j}$ olacak şekilde sıralanır. Olayları koruma bölme kuralının doğruluğunun bir ölçüsünü elde etmek için,

$$R_{k,j} = \sum_{l=1}^{k} \widehat{H}_j\left(T_{(l),j}\right) - \sum_{l=1}^{k} \delta_{(l),j} \ , \ \text{k=1, … , } n_j. \qquad \text{Denklem 2.50}$$

$$Conserve(x, c) = \frac{1}{Y_{1,1}+Y_{1,2}} \sum_{j=1}^{2} Y_{1,j} \sum_{k=1}^{n_j-1} \left|R_{k,j}\right| \qquad \text{Denklem 2.51}$$

Kısaca olayları koruma bölme kuralında, her bir j çocuk düğümü için, $R_{k,j}$ 'nin büyüklükleri toplanır ve her bir çocuk düğüm içindeki risk altındaki bireylerin sayısı ile ağırlıklandırılır.

Formül sonucunda elde edilen test istatistik değerleri azaldıkça iki çocuk düğüm arasındaki ayrım farkı arttığından, en uygun bölünmenin elde edilmesi için bu değerin en aza indirilmesi veya dönüştürülmüş değeri 1/(1 + Conserve(x, c)) 'inin en yükseğe çıkarılması gerekmektedir (Ishwaran, Kogalur, Blackstone, Lauer,2008).

2.2.2 Rastgele sağkalım orman yöntemi algoritması

Rastgele sağkalım orman yöntemindeki en genel algoritma adımları aşağıdaki gibidir,

1: Orijinal veri kümesinden B adet önyükleme (bootstrap) örnekleri alınır. Her önyükleme örneklemi, verilerin yaklaşık %30'unu hariç tutar ve bu verilere torba dışı (out-of-bag, OOB) veriler denir.

2: Her bir önyükleme (bootstrap) örneklemi için her düğümünde rastgele √p değişkenleri seçilen bir sağkalım ağacı yaratılır. Önceden belirlenmiş bir bölme kuralı kullanarak ek düğümler arasındaki maksimize eden aday değişken seçilerek düğüm bölünür.

3: Her bir uçbirim düğümde $d_o > 0$ tane ilgilenilen olay, gözlenen veri elde edilinceye kadar bölme işlemi sürdürülür.

4: Birikimli hazard fonksiyonunu her ağaç için hesaplanır. Daha sonrasında ortalaması alınarak topluluk kümülatif hazard fonksiyonu tahmin edilir.

5: 4.adımda hesaplanan kümülatif hazard fonksiyonu tahminine ait eğriler Torba dışı (OOB) verilere ile hesaplanır (Ishwaran, Kogalur, Blackstone, Lauer,2008).

2.3 Modellerin performanslarının değerlendirilmesinde kullanılan indeksler

2.3.1 Brier Skoru

Brier skoru, belirli bir t zamanında tahmin edilen sağkalım fonksiyonunun doğruluğunu analiz etmek için kullanılır ve t_i, gözlemlenen sağkalım olasılığı ile tahmin edilen sağkalım olasılığı arasındaki ortalama farkın karesini temsil eder ve Brier skoru her zaman 0 ile 1 arasında değer alır ve Brier skorunun 0 olması durumu mümkün olan en iyi durumdur (Fotso, 2018).

Sağdan sansürlü verilerin olduğu bir data setinde, t sağkalım anında i'nci verinin durumu $\Delta i = I(\tilde{T} \leq t)$, ve X ortak değişkeni ele alındığında i'nci veri için t sağkalım zamanında tahmin edilen sağkalım olasılığı $\hat{S}(t|X_i)$ olarak,

$$BS(t) = \frac{1}{N}\sum_{i=1}^{N} \begin{pmatrix} \left(\hat{S}(t|X_i)\right)^2 I(\tau_i \leq t \text{ and } \delta_i = 1)\hat{G}^{-1}(\tau_i) \\ + \\ \left(1 - \left(\hat{S}(t|X_i)\right)^2 I(\tau_i > t)\hat{G}^{-1}(t)\right) \end{pmatrix} \quad \text{Denklem 2.52}$$

Burada $\hat{G}(t|x) \approx P(C > t|X = x)$ sansürleme sürelerinin koşullu hayatta kalma fonksiyonunun Kaplan-Meier tahminidir (Nasejje, Mwambi, Dheda, Lesosky, 2017)

2.3.2 IBS skoru

İntegrali alınmış Brier Skoru (IBS), sağkalım modelinin tüm sağkalım sürelerindeki tahminlenmesi çin genel bir ölçüdür (Kronek, Reddy, 2008).

İntegrali alınmış brier skor (IBS),

$$IBS = \int_0^{\max(t)} BS(t)dt \quad \text{Denklem 2.53}$$

2.3.3 Harrell'in uyum indeksi (C Index)

Harrell concordance index (C-index) sağkalım analizlerinin performansını ölçmede kullanılmaktadır. C-endeksi, deneklerin sansürlenmesi de dikkate alarak, rastgele seçilmiş bir çift deneğin sağkalım risklerine bağlı olarak tahmin edilen ve gerçek gözlemlenen değerler arasındaki ilişkinin parametrik olmayan bir tahminini sağlar. Tahmin hatası oranı 1-C olarak hesaplanır ve C, Harrell uyum indeksi'dir. Tahmin hata oranları 0 ile 1 arasındadır ve 0 değerine ne kadar yakında o kadar iyidir. (Ishwaran, Kogalur, Blackstone, Lauer,2008).

2.4 Tez Çalışması Detayları

2.4.1 Tez Konusu

Genelde veri madenciliği teknikleri tanımlayıcı ve tahmin edici (descriptive-predictive) olarak ikiye ayrılır. Bu teknikler çoğunlukla tıpta tahmin edici (predictive) amacıyla kullanılır. Bu çalışmada da, ağaç tabanlı Rastgele Sağkalım Orman veri amdenciliği yönteminin araştırılması ve tıp alanındaa toplanan veri üzerinde analizler yaparak sonuçların incelenmesi amaçlanmıştır

Aynı zamanda sağlıkta Veri Madenciliğinin uygulanımı ile ilgili bir düzenlenme yaratmak ve Veri Madenciliğinin sağkalım süresini etkileyen değişkenlerin seçimi açısından kullanılmasına katkı sağlamaktır.

2.4.2 Verilerin toplanması

Çalışmada meme kanseri hastalığına ait sağkalım süresini etkileyen değişkenlerin tespiti için toplanmış, Mendeley Data açık kaynak platformunda yayınlanan Awodutire, Kolawole, Ilori tarafından 2017 yılındaki çalışmalarında uygulanmış, Osogbo'daki Ladoke Akintola Teknoloji Eğitim Hastanesi'ndeki 89 adet meme kanseri hastasından alınan klinik verilerle analizler yapılmıştır. Bu veri setinde, hastaların sağkalım süresi, tanı konuldan bir yıl sonra sağdan sansürlenerek alınmıştır. Datada yer alan survival süresi, hastanın kabul edildiği günden, son temasına kadar geçen süredir ve gün olarak hesaplanmıştır. Hastanın yaşam süresini etkileyebilecek

değişkenleri, hasta yaşı, menarştaki (ovulasyonun ilk görüldüğü) yaşı, doğum kontrol hapı kullanımı, ortalama emzirme yılı, hastalığın tespiti anındaki tümör gelişim evresi ve neodjuvan tedavi kullanımıdır (Awodutire, Kolawole, Ilori, 2017).

2.4.3 İstatistiksel Analizler

Çalışmada tüm istatistiksel analizler için anlamlılık düzeyi 0.05 olarak belirlenmiş ve uygulamada RStudio Version 1.4.1717 programı kullanılmıştır. R'da uygulanan tüm kodları Ek.1' de yer almaktadır.

Analizlerde, kategorik veriler için frekans tabloları, numerik veriler için tanımlayıcı istatistik değerleri hesaplanmıştır. Sağkalım analizlerinde ilk olarak Kaplan-Meir sağkalım eğrisi oluşturulmuş (Şekil 4.1), Log-rank testi ile gruplara ait sağkalım eğrileri karşılaştırılmış ve tümör tanı evresi (Şekil 4.2), doğum kontrol hapı kullanım durumu (Şekil 4.3) ve tedavide neoadjuvan uygulanım durumu (Şekil 4.4) gruplarına ait sağkalım eğrileri arasında istatistiksel olarak anlamlı fark tespit edilmemiştir.

Çalışmada Cox regresyon için Schoenfeld testi sonuçlarına (Şekil 4.6) ile testin genelinde orantısal hazard varsayımı sağlandığı tespit edilmiş ve cox regresyon modeli oluşturulmuştur (Tablo 4.3). Bu modele göre hastaların tedavi süresinde neoadjuvan kullanımının tehlike oranı üzerindeki etkisi istatistiksel olarak anlamlı analiz edilmiş ve tehlike oranı incelendiğinde ise, tedavi süresinde neoadjuvan kullanımı olmayan bireylerin olan bireylere göre ölüm riski 0.5 kat daha az olduğu tespit edilmiştir.

Rastgele orman sağkalım yöntemi sağkalımda önemli etki gösteren değişkenler ise Ortalama emzirme yılı en yüksek sonrasında da doğum kontrol hapı kullanımı olarak belirlenmiştir (Şekil 4.7).

Her iki yöntemin performanslarını yorumlamak için Harrel C-Indeks değerleri hesaplanmış (Tablo 4.4), ve grafik halinde gösterilmiştir (Şekil 4.9).

3. Bulgular

Tablo 3.1 Çalışmada yer alan hastalara ait tanımlayıcı istatistik değerleri

	Hastaların Yaşı	Menarş Yaşı	Emzirme Yılı
N	89	89	89
Ortalama	50.29	15.72	1.39
Std.Sapma	10.848	2.326	0.547
Medyan	48.00	15.00	1.50
Min	28.0	12.0	0.0
Maks	77.0	22.0	3.0

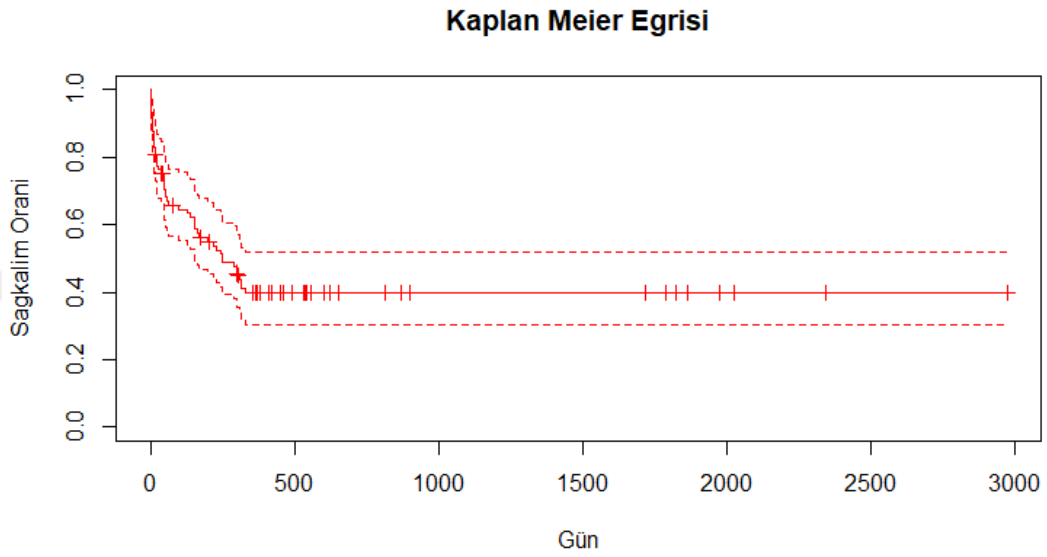
Tablo 4.1'de belirtildiği gibi, çalışmaya katılan hastaların yaş ortalaması 50.29±10.848, ortalama menarş yaşı 15.72±2.326 ve ortalama emzirme yılı 1.39±0.547'dir.

Tablo 3.2 Çalışmada yer alan hastalara ait verilerin frekans dağılımı

		Frekans (n)	%
Olay	Sağdan Sansürlü	38	42.7
	Sansürsüz	51	57.3
Doğum Kontrol Hapı Kullanımı	Evet	48	53.9
	Hayır	41	46.1
Tümör Evresi	Erken Evre(Evre I. Ve II)	60	67.4
	Geç Evre(III.Evre ve IV.Evre)	29	32.6
Tedavi Boyunca Neoadjuvant Uygulanımı	Evet	39	43.8
	Hayır	50	56.2

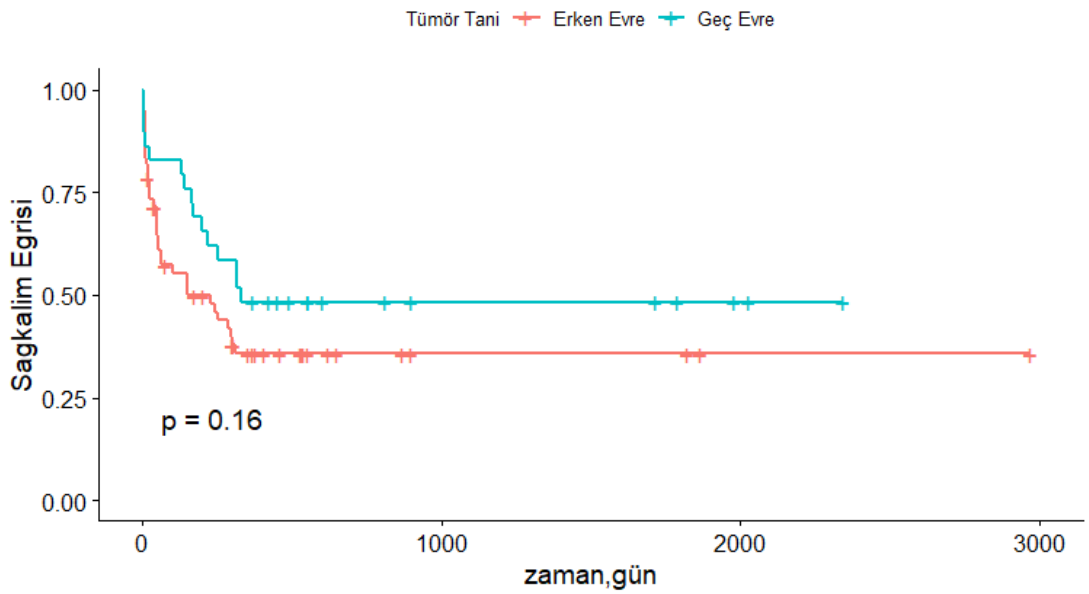
Tablo 4.2'de belirtildiği gibi, Hastalara ait sağkalım sürelerinin %42.7'si (n=38) sağdan sansürlü, %57.3'ü (n=51) sansürsüzdür. Hastaların %53.9'u (n=48) doğum

kontrol hapı kullanırken, %46.1'i (n=41) kullanmamaktadır. Hastaların meme kanseri teşhisi anındaki evresi %67.4'ünün (n=60) erken evre(Evre I. Ve II), %32.6'sının (n=29) geç evredir (III.Evre ve IV.Evre). Tedavi sırasında %43.8'ine (n=39) neoadjuvant uygulanmışken, %56.2'sine (n=50) uygulanmamıştır.



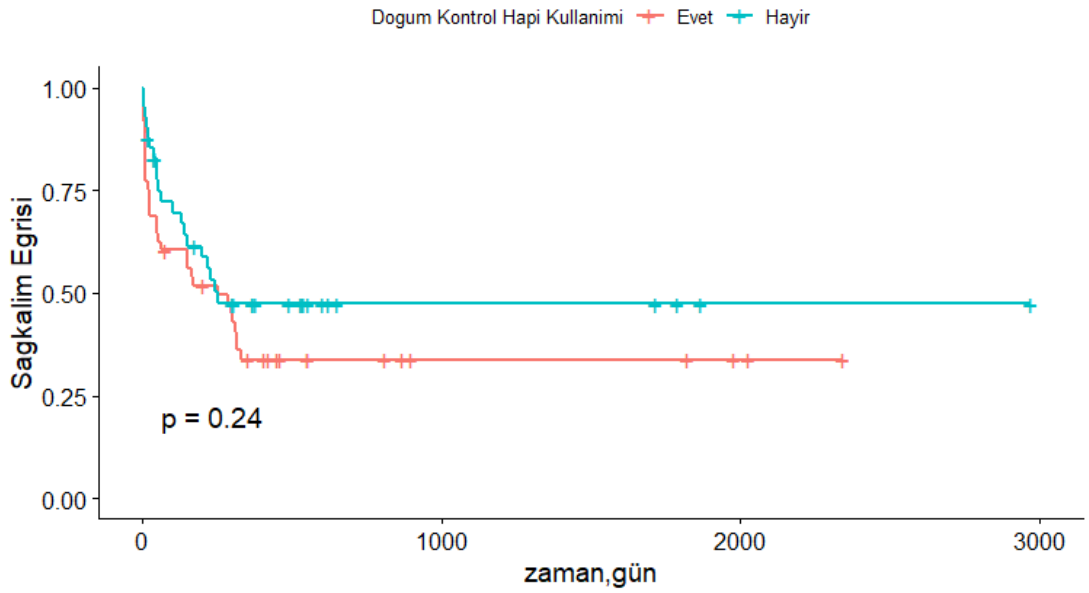
Şekil 3.1 Kaplan-Meier Sağkalım Eğrisi

Şekil 3.1'de Kaplan-Meier sağkalım eğrisi gösterilmektedir. Çalışmada yer alan 89 hastadan %57.3'ünde (n=51) olay(ölüm) gerçekleşmiştir. Sağkalım eğrisinde ilk 329 günde ölümler gerçekleştiği görülürken, sonraki izlem süresinde sağkalım oranında önemli değişimler gözlemlenmemiştir. Çalışmada yer alan hastaların medyan sağkalım süresi 251± 52.82 [%95 G.A. 147.5-354.5] gün olarak hesaplanmıştır. Hastaların ilk 4 ay sağkalım oranı 0.60'ın üzerindedir.



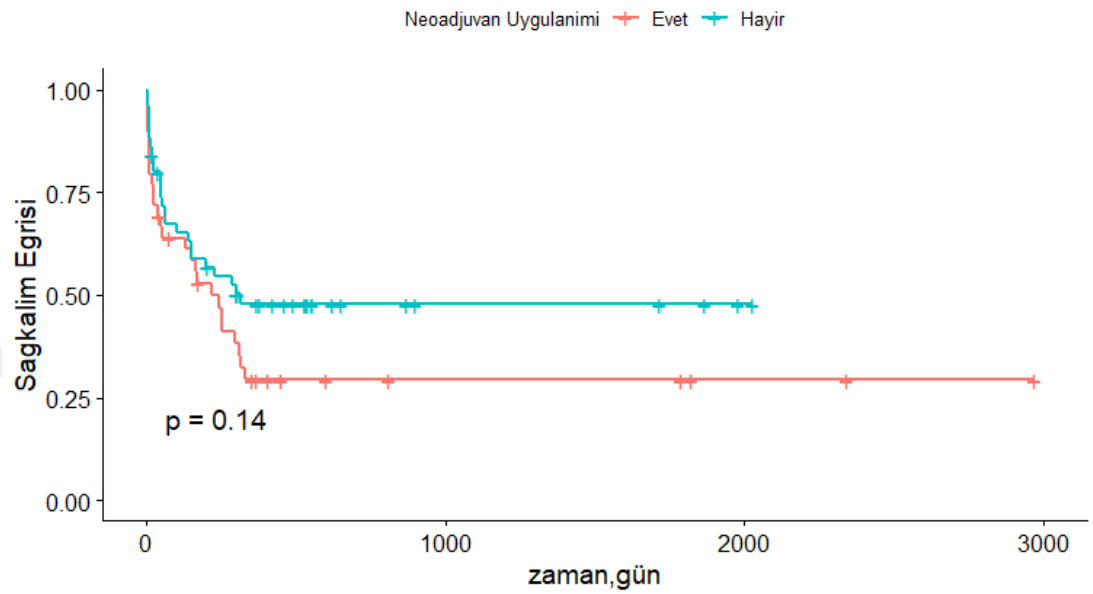
Şekil 3.2 Tümör Tanı grupları arasında sağkalım olasılıkları arasındaki farkın analizi Log-rank Testi

Şekil 3.2'de hastaların tümör tanısını aldığı evreye göre sağkalım olasılıklarının dağılımı gösterilmiştir. Log rank testine göre, Erken evre ve geç evre hasta gruplarının sağkalım olasılıkları arasında istatistiksel olarak anlamlı fark yoktur (p=0.163).



Şekil 3.3 Doğum kontrol hapı kullanımı grupları arasında sağkalım olasılıkları arasındaki farkın analizi Log-rank Testi

Şekil 3.3'de hastaların doğum kontrol hapı kullanıp kullanmadığına göre sağkalım olasılıklarının dağılımı gösterilmiştir. Log rank testine göre, Doğum kontrol hapı kullanan ve kullanmayan hasta gruplarının sağkalım olasılıkları arasında istatistiksel olarak anlamlı fark yoktur (p=0.242).



Şekil 3.4 Tedavide neoadjuvan uygulanımı grupları arasında sağkalım olasılıkları arasındaki farkın analizi Log-rank Testi

Şekil 4.4'de hastalara tedavi süresince neoadjuvan uygulanıp uygulanmadığına göre sağkalım olasılıklarının dağılımı gösterilmiştir. Log rank testine göre, Neoadjuvan uygulanan ve uygulanmayan hasta gruplarının sağkalım olasılıkları arasında istatistiksel olarak anlamlı fark yoktur (p=0.143).

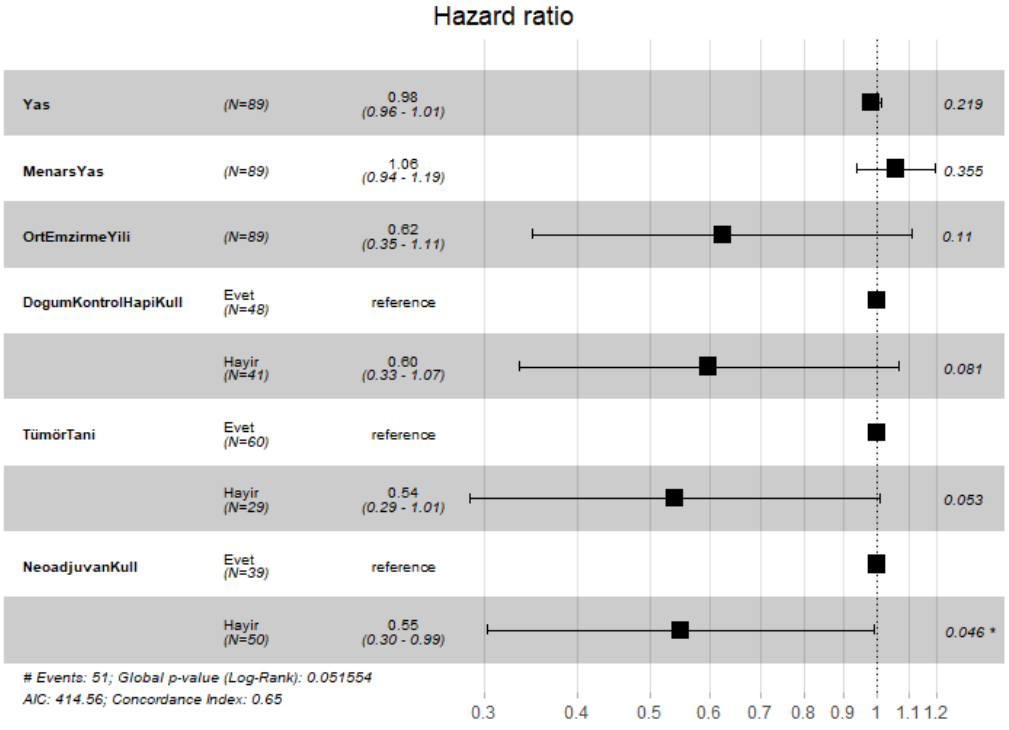
Tablo 3.3 Cox Regresyon Analizi Sonuçları

	β	Hazard Ratio	Std. Hata	p*
Yaş	-0.0177	0.9825	0.0144	0.219
Menarş Yaşı	0.0568	1.0584	0.0614	0.355
Ort. Emzirme Yılı	-0.4728	0.6233	0.2962	0.111
Doğum Kontrol Hapı Kullanımı	-0.5165	0.5966	0.2965	0.081
Tümör Tanı	-0.6194	0.5383	0.3205	0.053
Neoadjuvan Kullanımı	-0.6025	0.5475	0.3024	0.046

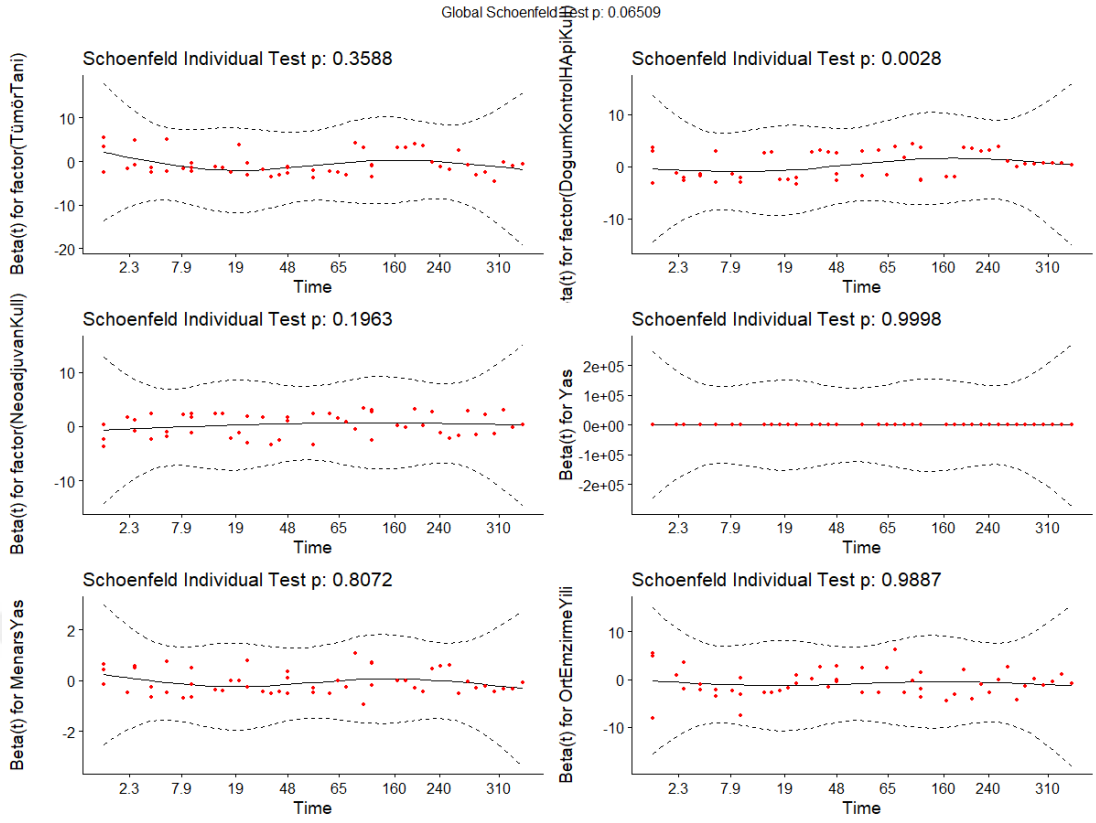
C-Index=0.648

Tablo 4.3'de cox regresyon analizi sonuçları belirtilmiştir. Buna göre, hastanın yaşı, menarş yaşı, ortalama emzirme yılı, doğum kontrol hapı kullanımı ve tümör tanı evresi değişkenleri bakımından tehlike oranı üzerindeki etkileri istatistiksel olarak anlamlı bulunmamıştır (p=0.219, p=0.355, p=0.111, p=0.081, p=0.053) Ancak tedavi süresinde neoadjuvan kullanımının tehlike oranı üzerindeki etkisi istatistiksel olarak anlamlıdır (p<0.05). Tehlike oranı incelendiğinde ise, tedavi süresinde neoadjuvan kullanımı olmayan bireylerin olan bireylere göre ölüm riski 0.5 kat daha azdır. Cox regresyon modeli uyum indeksi olan C-Index 0.648 olarak hesaplanmıştır.

Şekil 4.5'de tüm ortak değişkenler için Cox regresyon modelinden türetilen tahmini tehlike oranlarını (HR) gösterilmektedir. Tehlike oranı (HR) > 1 olması durumunda, ölüm riskinin arttığı, Tehlike oranı (HR) < 1 ise ölüm riskin azaldığı söylenmektedir.

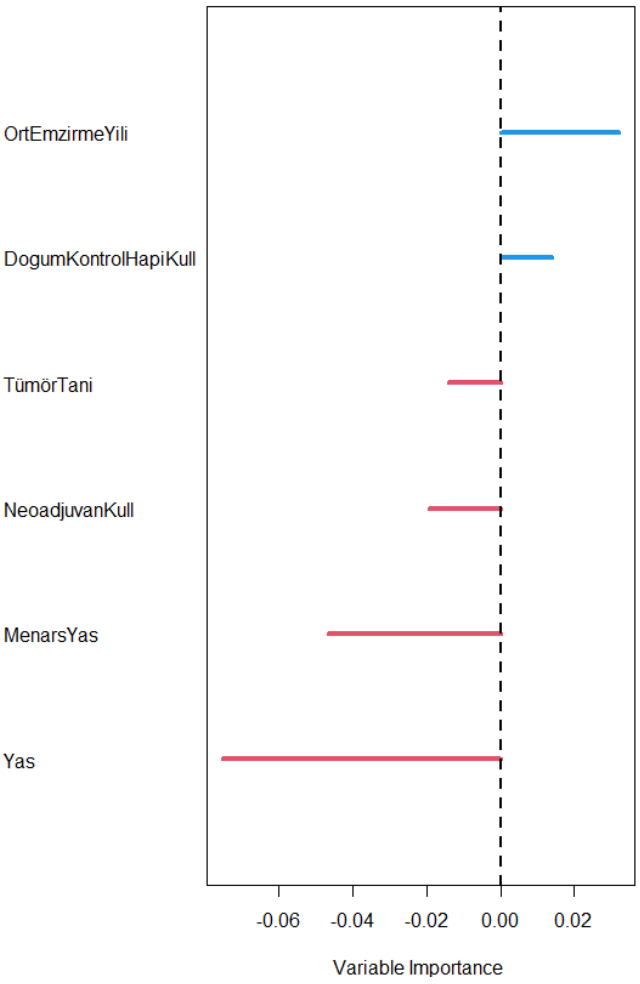


Şekil 3.5 Cox tehlike oranları modeli çıktısı



Şekil 3.6 Değişkenlere ait Schoenfeld Testi Sonuçları

Şekil 3.5 ‘de belirtildiği gibi, Cox regresyon için Schoenfeld testi sonuçlarına göre testin genelinde ve hastanın yaşı, menarş yaşı, ortalama emzirme yılı, neoadjuvan kullanımı ve tümör tanı değişkenleri için orantısal hazard varsayımı sağlanmaktadır ($p>0.05$). Doğum kontrol hapı kullanımı değişkeni için orantısal hazard varsayımı sağlanmamaktadır ($p<0.05$).



Şekil 3.7 Random Forest Yöntemi ile veri setinin önemli değişkenleri

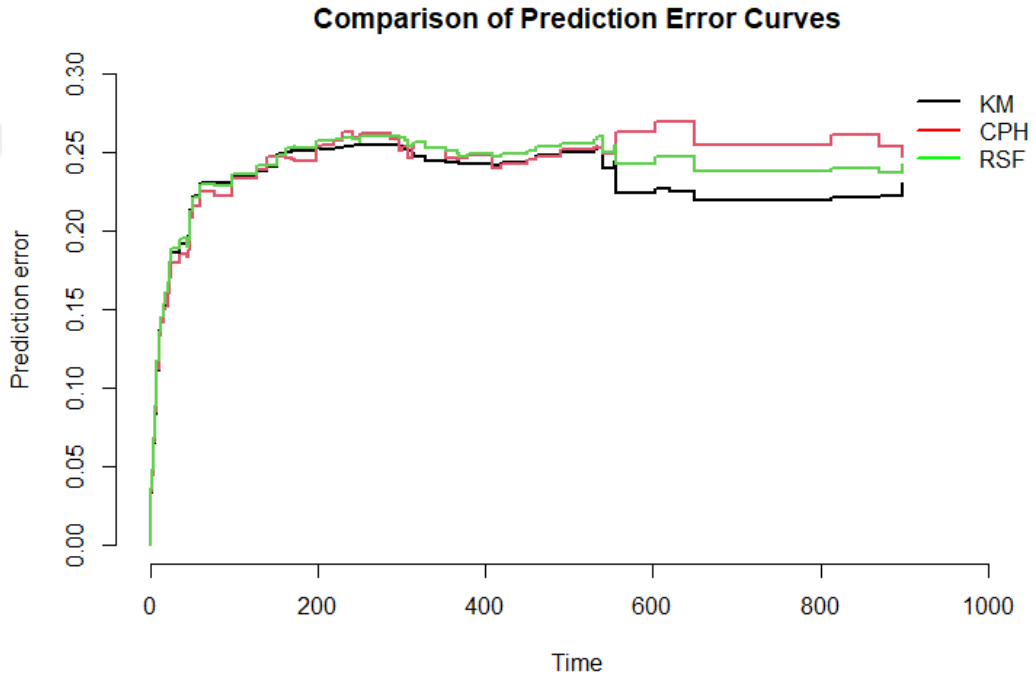
Göğüs kanseri hastalarının ölüm tehlikelerini belirlemek için random forest yöntemi uygulanmıştır. Bu yöntemde veri seti %70 test, %30 analiz dışı tutulmuş yani out-of-bag, OOB veri olarak ayrılmıştır. Logrank bölme kuralına göre n=80 adet sağkalım ağacı yaratılmış ve random forest sağkalım yöntemine göre önemli değişkenler Ortalama emirme yılı en yüksek sonrasında da doğum kontrol hapı kullanımı olarak belirlenmiştir. Random forest yöntemine ait C-index 0.60 olarak hesaplanmıştır.

Tablo 3.4 Modellerin C-Index değerleri

	C-Index

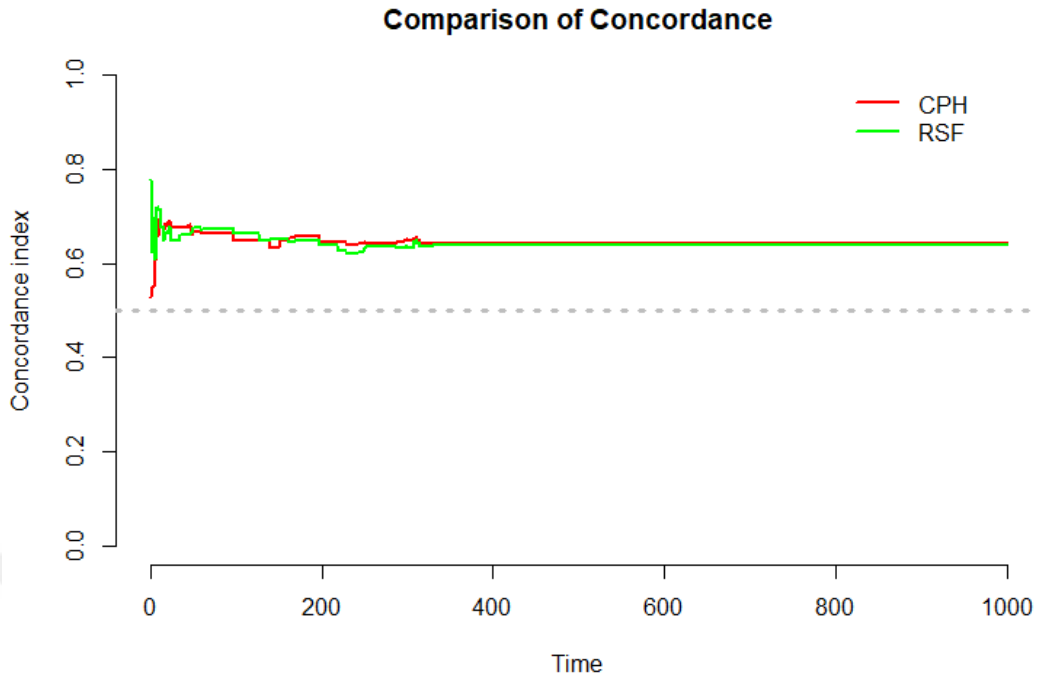
Random Forest Survival	0.596
Cox Regresyon	0.648

Tablo 4.4'de göğüs kanseri hastalara ait verilerle oluşturulan Cox Regresyon ve ratgele sağkalım orman modellerine ait C-Index değerleri verilmiştir. Buna göre Cox Regresyon modeline ait C-Index değerinin rastgele sağkalım orman yöntemi C-Index değerinden daha yüksek hesaplandığı söylenebilir.



Şekil 3.8 Kaplan-Meier, Cox Orantılı Tehlikeler ve Rastgele sağkalım orman yöntemlerinin tahmin hatası grafiği

Şekil 3.8'de belirtildiği gibi, kanser hastalığına sahip hastaların verileri için tahmin hatası grafiği incelendiğinde, tüm yöntemlerin yaklaşık olarak aynı eğriyi takip ettiği söylenebilir. Rastgele sağkalım orman yönteminin, diğer modellerden sürekli olarak biraz daha yüksek bir tahmin hatasına sahip olduğu, 600 gün öncesinde Cox regresyon modelinin en yüksek tahmin hatasıyla devam ettiği söylenebilir.



Şekil 3.9 Cox regresyon ve rastgele sağkalım orman yöntemi uyum indeksi (C-Index) grafiği

Şekil 3.9'da sağkalım analizi modellerini karşılaştırmak için, her modelin zaman içindeki uyum indeksi (c-endeksi) gösterilmektedir. Bu yöntem, sağkalım modellerini değerlendirmek için en yaygın kullanılan yöntemlerden biri olarak kabul edilmektedir. C-endeksi, tahmin edilen ve gözlemlenen sağkalım arasındaki uyum olasılığını verir. C-Indeks değeri 1'e yaklaştıkta modelin tahmin etme performansı yüksek olarak kabul edilir. Grafiğe göre, her iki yöntemin performanslarının benzer olduğu söylenebilir.

4. Tartışma

Bu çalışmada, sağkalım analizindeki etkili değişkenlerin tespitinde sıklıkla kullanılan Cox regresyon ve rastgele orman sağkalım makine öğrenme teknikleri sonuçları analiz edilmiştir. Günümüzde artan veri sayısına bağlı olarak, makine öğrenme tekniklerinin kullanımı giderek yaygınlaşmıştır. Gordon ve Olshen (1985) sağkalım analizinde ağaç tabanlı yöntemi ilk olarak kullanmıştır. Burada ayırma kriteri olarak kaplan-meier eğrileri arasındaki uzaklığı dikkate almıştır. Ciampi ve arkadaşları (1986) ayırma kriteri olarak olabilirlik oranını, Segal (1988) ise logrank test istatistiğini önermiştir.

Çalışmada 89 adet meme kanseri hastalığına ait sağkalım süresini etkileyen değişkenlerin tespiti için Mendeley Data açık kaynak platformunda yayınlanan Awodutire, Kolawole, Ilori tarafından 2017 yılındaki çalışmalarında uygulanmış veri seti ile ilk önce kaplan-meier yöntemi uygulanmış, grupları arasında sağkalım olasılıkları arasındaki farkın analizi için Log-rank Testi uygulanmıştır. Daha sonrasında cox regresyon modeli kurulmuş ve etkisi anlamlı olan değişkenler tespit edilmiştir. Aynı şekilde Rastgele sağkalım orman yöntemi ile de sağkalıma etkisi önemli olan değişkenler analiz edilmiş ve her iki modelin performansı Harrell'in C-Index değerine göre yorumlanmıştır.

Kaplan-Meier sağkalım eğrilerini karşılaştırmada kullanılan Log rank testine göre, Tümör teşhis evresine, Erken evre ve geç evre hasta gruplarının sağkalım olasılıkları, Hastaların doğum kontrol hapı kullanımlarına göre, kullanan ve kullanmayan hasta gruplarının sağkalım olasılıkları, Hastalara tedavi süresince neoadjuvan uygulanıp uygulanmadığına göre, uygulanan ve uygulanmayan hasta gruplarının sağkalım olasılıkları arasında istatistiksel olarak anlamlı fark tespit edilmemiştir.

Çalışmada sonrasında Cox regresyon modeli kurulmuş ve bu modele göre hastanın yaşı, menarş yaşı, ortalama emzirme yılı, doğum kontrol hapı kullanımı ve tümör tanı evresi değişkenleri bakımından tehlike oranı üzerindeki etkileri istatistiksel olarak anlamlı bulunmazken, rastgele sağkalım orman yönteminde önemli değişkenler Ortalama emzirme yılı en yüksek sonrasında da doğum kontrol hapı kullanımı olarak belirlenmiştir. Ancak Cox regresyon modeli için sadece tedavi süresinde neoadjuvan kullanımının tehlike oranı üzerindeki etkisi istatistiksel olarak anlamlı analiz edilmiştir. Cox regresyon yöntemi ve rastgele sağkalım orman yöntemi performanslarını karşılaştımada Harrell'in C-Index'i hesaplanmış ve buna göre, Cox regresyon yönteminin daha yüksek ama aslında her iki yöntemde de çok yakın C-Index değeri hesaplamıştır. Bu durumda iki modelin performanslarının yakın olduğu söylenebilir.

Zhang ve arkadaşları, 2019 yılında 6328 hastanın verileri ile, dislipidemiyi tahmin etmek için Cox regresyon ve rastgele sağkalım yöntemleri uygulanmıştır. İki modelin yorumlanmasında Harrell'in C-Index değeri kullanılmış ve rastgele sağkalım orman modeli, Cox orantılı tehlike regresyonu modelinden daha iyi ayırt edici performansa sahip olduğu tespit edilmiştir.

Rahman ve arkadaşları 2021 yılıda 6198 adet mide kanseri hastası ile rastgele sağkalım orman yöntemi ve cox regresyon modeli ile sağkalım tahminleme analizi yapılmıştır. İki yöntemin performansını değerlendirmede Harrell'in C-Index değeri kullanılmış ve buna göre rastgele sağkalım orman modelinin, küratif özofajektomi sonrası sağkalımı tahmin etmede cox regresyon modeline göre daha iyi performans gösterdiği tespit edilmiştir.

Dateme ve arkadaşları 2012 yılında 1371 hasta verileriyle cox regresyon modellemesi ve rastgele sağkalım modellemesi analiz edilmiş ve Harrell'İn uyum hata oranı karşılaştırılması ile model performansları değerlendirilmiştir. Buna göre

Rastgele sağkalım modeli ve Cox orantılı tehlike regresyonu modeli benzer hata oranlarını gösterdiği tespit edilmiştir.

Qiu ve arkadaşları 202 yılında 82 adet hastanın partikül ışını radyoterapisinden sonra yüksek dereceli gliomanın tümör ilerlemesini tahmin etmede rastgele sağkalım orman yöntemi ve cox regresyon yöntemi uygulanmıştır. Çalışmanın sonucunda her iki yöntemin performansı Harrell'in C-Index'ine göre yorumlanmış ve Cox orantılı tehlike regresyonu modelinin rastgele sağkalım orman yönteminden daha iyi performans gösterdiği tespit edilmiştir.

Hsich ve arkadaşları 2011 yılında 2231 adet kalp yetmezliği olan hastanın 5 yıllık takip süresi analiz edilerek sağkalımı modellenmiştir. Modellemede Cox orantılı tehlike regresyonu ve ratgele sağkalım orman modeli uygulanmış ve C-Index hesaplanmasına göre her iki yöntem de benzer şekilde sağkalımı öngördüğü tespit edilmiştir.

Hamidi ve ark 2016 yılında 378 adet böbrek transplantasyonu olan hastaların sağkalımı modellemesinde rastgele sağkalım yöntemi ve Cox orantılı tehlike regresyon modeli uygulanmış, hesaplanan C-Index değerlerine göre Cox orantılı tehlike modelinin performansının Rastgele sağkalım orman yönteminden daha iyi olduğu tespit edilmiştir.

5. Sonuç ve Öneriler

Çalışmada Mendeley Data açık kaynak platformunda yayınlanan Awodutire, Kolawole, Ilori tarafından 2017 yılındaki çalışmalarında uygulanmış, 89 adet meme kanseri hastasına ait veriler kullanılmış, bu verilerde Hastaların ait sağkalım süreleri,

tanı aldıktan bir yıl sonra sağdan sansürlenerek kaydedilmiştir. Kaydedilen sağkalım süresi, hastanın teşhis edildiği gününden son temas (ölüm, canlı veya takip kaybı) gününe kadar geçen süredir ve gün olarak toplanmıştır. Hastanın sağkalım süresini etkileyebilecek değişkenler, Hastanın Yaşı, Hastanın Menarştaki Yaşı (menarş), Doğum Kontrol Hapı Kullanımı, Ortalam Emzirme Yılı, teşhis anında tümör gelişiminin evresi ve Neoadjuvan Tedavisinin (neoadjuvan) kullanımı olarak yer almaktadır.

Çalışmada yer alan 89 hastadan %57.3'ünde (n=51) olay(ölüm) gerçekleşmiştir ve Kaplan-Meier sağkalım eğrisinde ilk 329 günde ölümler gerçekleştiği görülürken, sonraki izlem süresinde sağkalım oranında önemli değişimler gözlemlenmemiştir. Çalışmada yer alan hastaların medyan sağkalım süresi 251± 52.82 [%95 G.A. 147.5-354.5] gün olarak hesaplanmıştır. Hastaların ilk 4 ay sağkalım oranı 0.60'ın üzerindedir.

Kaplan-Meier sağkalım eğrilerini karşılaştırmada kullanılan Log rank testine göre, Tümör teşhis evresine, Erken evre ve geç evre hasta gruplarının sağkalım olasılıkları arasında istatistiksel olarak anlamlı fark yoktur (p=0.163). Hastaların doğum kontrol hapı kullanımlarına göre, kullanan ve kullanmayan hasta gruplarının sağkalım olasılıkları arasında istatistiksel olarak anlamlı fark yoktur (p=0.242). Hastalara tedavi süresince neoadjuvan uygulanıp uygulanmadığına göre, uygulanan ve uygulanmayan hasta gruplarının sağkalım olasılıkları arasında istatistiksel olarak anlamlı fark yoktur (p=0.143).

Çalışmada sonrasında Cox regresyon modeli kurulmuş ve bu modele göre hastanın yaşı, menarş yaşı, ortalama emzirme yılı, doğum kontrol hapı kullanımı ve tümör tanı evresi değişkenleri bakımından tehlike oranı üzerindeki etkileri istatistiksel olarak anlamlı bulunmamıştır (p=0.219, p=0.355, p=0.111, p=0.081, p=0.053). Ancak tedavi süresinde neoadjuvan kullanımının tehlike oranı üzerindeki etkisi istatistiksel olarak anlamlıdır ($p<0.05$). Tehlike oranı incelendiğinde ise, tedavi süresinde neoadjuvan kullanımı olmayan bireylerin olan bireylere göre ölüm riski 0.5 kat daha azdır. Cox regresyon modeli uyum indeksi olan C-Index 0.648 olarak hesaplanmıştır.

Çalışmada sağkalım analizinde uygulanan makine öğrenme tekniklerinden olan rastgele sağkalım orman yöntemi uygulanmıştır. Uygulama sonrasında önemli değişkenler Ortalama emzirme yılı en yüksek sonrasında da doğum kontrol hapı

kullanımı olarak belirlenmiştir. Random forest yöntemine ait C-index 0.60 olarak hesaplanmıştır.

Çalışmada örneklem sayısının arttırılması hem modelleme için hem de etkileyen değişkenlerin tespiti için önemli rol oynayabilir. Sağkalım süresini etkileyen değişenlerin analiz edilmesinde rastgele orman yönteminin performansının Cox regresyon yönteminden daha iyi olduğu söylenememektedir. Çalışmada örneklem ve değişken sayısı arttırarak tekrardan iki yöntemin analizleri yapılabilir.

6. Kaynakça

Agrawal, R., Imieliński, T., & Swami, A. (1993, June). Mining association rules between sets of items in large databases. In Proceedings of the 1993 ACM SIGMOD international conference on Management of data (pp. 207-216).

Ağyar, Z. (2015). Yapay sinir ağlarının kullanım alanları ve bir uygulama. Mühendis ve Makine, 56(662), 22-23.

Ahmad, P., Qamar, S., & Rizvi, S. Q. A. (2015). Techniques of data mining in healthcare: a review. International Journal of Computer Applications, 120(15).

Akman, M., Genç, Y., & Ankarali, H. (2011). Random forests yöntemi ve sağlık alanında bir Uygulama / Random forests methods and an application in health science. Türkiye Klinikleri Biyoistatistik, 3(1), 36-48.

Akpınar, H. (2000). Veri Tabanlarında Bilgi Keşfi ve Veri Madenciliği, İstanbul Üniversitesi, İşletme Fakültesi Dergisi, C.

Altman, D. G. (1990). Practical statistics for medical research. CRC press.

Awodutire, P. O., Kolawole, O. A., & Ilori, O. R. (2017). Parametric modeling of survival times among breast cancer patients in a teaching hospital, Osogbo. J. Cancer Treat. set.seed(1000) 5(5), 81-85.

Berry, M., & Linoff, G. (1999). Mastering data mining: The art and science of customer relationship management. John Wiley & Sons, Inc..

Breiman, L. (1996). Bagging predictors. Machine learning, 24(2), 123-140.

Breslow, N. E. (1975). Analysis of survival data under the proportional hazards model. International Statistical Review/Revue Internationale de Statistique, 45-57.

Chauhan, D., & Jaiswal, V. (2016, October). An efficient data mining classification approach for detecting lung cancer disease. In 2016 International Conference on Communication and Electronics Systems (ICCES) (pp. 1-8). IEEE.

Ciampi, A., Thiffault, J., Nakache, J. P., & Asselain, B. (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(3), 185-204.

Collett D. (1994). Modeling survival data in Medical research. Chapman & Hall, London.

Cox, D. R. (1972). Regression models and life-tables. Journal of the Royal Statistical Society: Series B (Methodological), 34(2), 187-202.

Cox, D. R., & Oakes, D. (2018). Analysis of survival data. Chapman and Hall/CRC.

Datema, F. R., Moya, A., Krause, P., Bäck, T., Willmes, L., Langeveld, T., ... & Blom, H. M. (2012). Novel head and neck cancer survival analysis approach: random survival forests versus Cox proportional hazards regression. Head & neck, 34(1), 50-58.

Dos Santos Silva, I. (1999). Cancer epidemiology: principles and methods. IARC.

Durairaj, M., & Ranjani, V. (2013). Data mining applications in healthcare sector: a study. International journal of scientific & technology research, 2(10), 29-35.

Elandt-Johnson, R. C., & Johnson, N. L. (1980). Survival models and data analysis (Vol. 110). John Wiley & Sons.

Elmaghraby, A. S., Kantardzic, M. M., & Wachowiak, M. P. (2006). Data mining from multimedia patient records. In Data Mining and Knowledge Discovery Approaches Based on Rule Induction Techniques (pp. 551-595). Springer, Boston, MA.

Emre, İ. E., & EROL, Ç. S. (2017). Veri Analizinde İstatistik mi Veri Madenciliği mi?. Bilişim Teknolojileri Dergisi, 10(2), 161-167

Exarchos, K. P., Goletsis, Y., & Fotiadis, D. I. (2011). Multiparametric decision support system for the prediction of oral cancer reoccurrence. IEEE Transactions on Information Technology in Biomedicine, 16(6), 1127-1134.

Fayyad, U. M. (1998). Mining databases: Towards algorithms for knowledge discovery. IEEE Data Eng. Bull., 21(1), 39-48.

Fayyad, U. M., Piatetsky-Shapiro, G., Smyth, P., & Uthurusamy, R. (Eds.). (1996, February). Advances in knowledge discovery and data mining. American Association for Artificial Intelligence

Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). From data mining to knowledge discovery in databases. AI magazine, 17(3), 37-37.

Fink, S. A., & Brown Jr, R. S. (2006). Survival analysis. Gastroenterology & hepatology, 2(5), 380.

Fotso, S. (2018). Deep neural networks for survival analysis based on a multi-task framework. arXiv preprint arXiv:1801.05512.

Ganesh, S. (2002, July). Data mining: Should it be included in the statistics curriculum. In The 6th international conference on teaching statistics (ICOTS 6), Cape Town, South Africa.

Ganesh, S. (2002, July). Data mining: Should it be included in the statistics curriculum. In The 6th international conference on teaching statistics (ICOTS 6), Cape Town, South Africa.

Gordon, L., & Olshen, R. A. (1985). Tree-structured survival analysis. Cancer treatment reports, 69(10), 1065-1069.

Hamidi, O., Poorolajal, J., Farhadian, M., & Tapak, L. (2016). Identifying Important Risk Factors for Survival in Kidney Graft Failure Patients Using Random Survival Forests. Iranian journal of public health, 45(1), 27–33.

Han, J. (2001). Spatial clustering methods in data mining: A survey. Geographic data mining and knowledge discovery, 188-217.

Han, J., & Kamber, M. (2000) Data Mining: Concepts and Techniques, Simon Fraser University

Han, J., Kamber, M., & Pei, J. (2011). Data mining concepts and techniques third edition. The Morgan Kaufmann Series in Data Management Systems, 5(4), 83-124.

Hothorn, T., & Lausen, B. (2003). On the exact distribution of maximally selected rank statistics. Computational Statistics & Data Analysis, 43(2), 121-137.

Hsich, E., Gorodeski, E. Z., Blackstone, E. H., Ishwaran, H., & Lauer, M. S. (2011). Identifying important risk factors for survival in patient with systolic heart failure using random survival forests. Circulation: Cardiovascular Quality and Outcomes, 4(1), 39-45.

Ishwaran, H., Kogalur, U. B., Blackstone, E. H., & Lauer, M. S. (2008). Random survival forests. The annals of applied statistics, 2(3), 841-860.

Kabalcı, E. (2014). Yapay Sinir Ağları. Ders Notları https://ekblc. files. wordpress. com/2013/09/ysa. pdf.

Kantardzic, M. (2011). Data mining: concepts, models, methods, and algorithms. John Wiley & Sons.

Karypis, G., Han, E. H., & Kumar, V. (1999). Chameleon: Hierarchical clustering using dynamic modeling. Computer, 32(8), 68-75.

Kaur, S., & Bawa, R. K. (2015). Future trends of data mining in predicting the various diseases in medical healthcare system. International Journal of Energy, Information and Communications, 6(4), 17-34.

Kaygisiz, G. (2010) Cox oransal hazard regresyon modeli ve trafik verilerine uygulanması, Yüksek Lisans Tezi, Gazi Üniversitesi, Fen Bilimleri Enstitüsü, Ankara, 115s.

Kharya, S. (2012). Using data mining techniques for diagnosis and prognosis of cancer disease.

Kincade, K. (1998). Data mining: digging for healthcare gold. Insurance & Technology, 23(2), 2-7.

Klein, J. P., & Moeschberger, M. L. (2003). Survival analysis: techniques for censored and truncated data (Vol. 1230). New York: Springer.

Kleinbaum, D. G., & Klein, M. (2010). Survival analysis (Vol. 3). New York: Springer.

Koyuncugil, A. S. (2007). Veri Madenciliği ve Sermaye Piyasalarına Uygulanması. Sermaye Piyasası Kurulu Araştırma Raporu, 1-17.

Koyuncugil, A., & Özgülbaş, N. (2009). Veri madenciliği: Tıp ve sağlık hizmetlerinde kullanımı ve uygulamaları. Bilişim Teknolojileri Dergisi, 2(2).

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

Kudyba, S. (Ed.). (2004). Managing data mining: advice from experts. IGI Global.

Kumar, D., & Bhardwaj, D. (2011). Rise of data mining: current and future application areas. International Journal of Computer Science Issues (IJCSI), 8(5), 256.

Kuonen, D. (2004). Data mining and Statistics: What is the connection?. The Data Administration Newsletter, 30, 1-6.

Lawless, J. F. (2011). Statistical models and methods for lifetime data (Vol. 362). John Wiley & Sons.

Lee, E. T., & Go, O. T. (1997). Survival analysis in public health research. Annual review of public health, 18(1), 105-134.

Lee, E. T., & Wang, J. (2003). Statistical methods for survival data analysis (Vol. 476). John Wiley & Sons.

Matthews, D. E., Farewell, V.T., 1988, Using and Understanding Medical Statistics, Switzerland

Milley, A. (2000). Healthcare and data mining. Health Management Technology, 21(8), 44-45.

Nabiyev, V. V. (2012). Yapay zeka: insan-bilgisayar etkileşimi. Basım yeri: Seçkin Yayıncılık.

Nasejje, J. B., Mwambi, H., Dheda, K., & Lesosky, M. (2017). A comparison of the conditional inference survival forest model to random survival forests based on a simulation study as well as on two applications with time-to-event data. BMC medical research methodology, 17(1), 1-17.

Nelson, W. B. (2003). Applied life data analysis (Vol. 521). John Wiley & Sons.

Obenshain, M. K. (2004). Application of data mining techniques to healthcare data. Infection Control & Hospital Epidemiology, 25(8), 690-695.

Özdamar, K. (2015). SPSS ile Biyoistatistik (10. bs.). Ankara: Nisan Kitapevi.

Özekes, S. (2003). Veri madenciliği modelleri ve uygulama alanları.

Pagano, M., & Gauvreau, K. (2018). Principles of biostatistics. CRC Press.

Popa, R. (Ed.). (2012). Genetic algorithms in applications. BoD–Books on Demand.

Qiu, X., Gao, J., Yang, J., Hu, J., Hu, W., Kong, L., & Lu, J. J. (2020). A Comparison study of machine learning (random survival forest) and classic statistic (cox proportional hazards) for predicting progression in high-grade glioma after proton and carbon ion radiotherapy. Frontiers in Oncology, 10, 2311.

Rahman, S. A., Walker, R. C., Crosby, T., Maynard, N., Cromwell, D. A., & Underwood, T. J. (2021). O14: RANDOM FOREST MODELS FOR PREDICTING SURVIVAL AFTER OESOPHAGECTOMY. British Journal of Surgery, 108(Supplement_1), znab117-014.

Ramkumar, G. D., Swami, A. N., & America, H. (1998). Clustering data without distance functions. IEEE Data Eng. Bull., 21(1), 9-14.

Rodrıguez, G. (2010). Parametric survival models. Princeton University, Rapport technique, Princeton.

Sayad, S., Data Mining. Erişim Tarihi: 12. 05. 2021. https://www.saedsayad.com/data_mining.htm (2021).

Schoenfeld, D. (1982). Partial residuals for the proportional hazards regression model. Biometrika, 69(1), 239-241.

Scrucca, L. (2013). GA: a package for genetic algorithms in R. Journal of Statistical Software, 53(4), 1-37.

Segal, M. R. (1988). Regression trees for censored data. Biometrics, 35-47.

Shah, S. C., & Kusiak, A. (2004). Data mining and genetic algorithm based gene/SNP selection. Artificial intelligence in medicine, 31(3), 183-196.

Sitar-tăut, A., Zdrenghea, D., Pop, D., & Sitar-tăut, D. (2009). Using machine learning algorithms in cardiovascular disease risk evaluation. Age, 1(4), 4.

Tüzüntürk, S. (2010). Veri madenciliği ve istatistik. Uludağ Üniversitesi İktisadi ve İdari Bilimler Fakültesi Dergisi, 29(1), 65-90.

Weathers, Brandon, et al. Comparision of Survival Curves Between Cox Proportional Hazards, Random Forests, and Conditional Inference Forests in Survival Analysis. 2017.

Whetten, A. B., Stevens, J. R., & Cann, D. (2021). The implementation of random survival forests in conflict management data: An examination of power sharing and third party mediation in post-conflict countries. PloS one, 16(5)

Yabacı, A. (2017). Sağkalım verilerinde kullanılan ağaç tabanlı yöntemlerin karşılaştırılması (Master's thesis, Uludağ Üniversitesi).

Yadav, S. K., & Pal, S. (2012). Data mining: A prediction for performance improvement of engineering students using classification. arXiv preprint arXiv:1203.3832.

Yayla, M. E. (2013). Yaşam Analizleri Ve Cox Regresyon Modeli.

Zhang, X., Tang, F., Ji, J., Han, W., & Lu, P. (2019). Risk Prediction of Dyslipidemia for Chinese Han Adults Using Random Forest Survival Model. Clinical epidemiology, 11, 1047–1055. https://doi.org/10.2147/CLEP.S223694

Zhao, C. M., & Luan, J. (2006). Data Mining: Going beyond Traditional Statistics. New Directions for Institutional Research, 131, 7-16.

Zhong, N., & Zhou, L. (Eds.). (2003). Methodologies for Knowledge Discovery and Data Mining: Third Pacific-Asia Conference, PAKDD'99, Beijing, China, April 26-28, 1999, Proceedings. Springer.

Zhu, R. (2013). Tree-based methods for survival analysis and high-dimensional data (Doctoral dissertation, The University of North Carolina at Chapel Hill).

Ekler

Ek-1 R kodları

```
library(ggplot2)
library(survminer)
library(survival)
library(gbm)
library(caret)
library(pROC)
library(tree)
library(ISLR)
library(vip)
library(e1071)
library(rminer)
library(tidyverse)
library(reshape2)
library(ggfortify)
library(rpart)
library(skimr)
library(kableExtra)
library(patchwork)
library(directlabels)
library(randomForest)
library(randomForestSRC)
library(pec)
library(prodlim)
library(ranger)
BreastCancer <- read.csv("C:/Users/info/Desktop/tez/Breastcancer_data.csv")
names(BreastCancer)<-
        c('zaman','Yas','olay','MenarsYas','OrtEmzirmeYılı','DoğumKontrolHapıKul
        l','TümörTanı','NeoadjuvanKull')
```

```
#tanımlayıcı istatistik değerleri
summary(BreastCancer)
write.csv(summary(BreastCancer),"output.csv")
#frekans tabloları
library(epiDisplay)
tab1(BreastCancer$DoğumKontrolHapıKull, sort.group = "decreasing", cum.percent
        = TRUE)
tab1(BreastCancer$TümörTanı, sort.group = "decreasing", cum.percent = TRUE)
tab1(BreastCancer$NeoadjuvanKull, sort.group = "decreasing", cum.percent =
        TRUE)
tab1(BreastCancer$olay, sort.group = "decreasing", cum.percent = TRUE)

data.frame(zaman = BreastCancer$zaman, Yas = BreastCancer$Yas,
        olay=BreastCancer$olay,MenarsYas = BreastCancer$MenarsYas,
      OrtEmzirmeYili = BreastCancer$OrtEmzirmeYili, DogumKontrolHapiKull =
        BreastCancer$DogumKontrolHapiKull,
      TümörTani = BreastCancer$TümörTani, NeoadjuvanKull =
        BreastCancer$NeoadjuvanKull)

BreastCancer$DogumKontrolHapiKull <-
        factor(BreastCancer$DogumKontrolHapiKull,
                        levels = c("1", "2"),
                        labels = c("Evet", "Hayır"))
BreastCancer$TümörTani <- factor(BreastCancer$TümörTani,
                levels = c("1", "2"),
                labels = c("Evet", "Hayır"))
BreastCancer$NeoadjuvanKull <- factor(BreastCancer$NeoadjuvanKull,
                    levels = c("1", "2"),
```

```
                                    labels = c("Evet", "Hayır"))
#kaplan Meier
surv_object <- Surv(time = BreastCancer$zaman, event = BreastCancer$olay)
kaplan.meier<-survfit(surv_object~1,BreastCancer)
kaplan.meier
summary(kaplan.meier)
plot(kaplan.meier, col = c("Red"), xlab = "Gün", ylab =
     "Sağkalım Oranı", main =
     "Kaplan Meier Eğrisi", mark.time = TRUE)
legend(x = 500, y = 1, lty = 1:2, cex = .95, bty = "n")
log.rank1 <- survfit(surv_object~ TümörTani, data = BreastCancer)
summary(log.rank1)
print(log.rank1)
ggsurvplot(log.rank1, data = BreastCancer, pval = TRUE, xlab="zaman,gün",
           ylab="Sağkalım Eğrisi",
        legend.title="Tümör Tanı", legend.labs=c("Erken Evre","Geç Evre"))

log.rank2 <- survfit(surv_object~ DogumKontrolHapiKull, data = BreastCancer)
summary(log.rank2)
ggsurvplot(log.rank2, data = BreastCancer, pval = TRUE, xlab="zaman,gün",
           ylab="Sağkalım Eğrisi",
        legend.title="Doğum Kontrol Hapı Kullanımı", legend.labs=c("Evet","Hayır"))

log.rank3 <- survfit(Surv(zaman,olay)~ NeoadjuvanKull, data = BreastCancer)
summary(log.rank3)
ggsurvplot(log.rank3, data = BreastCancer, pval = TRUE, xlab="zaman,gün",
           ylab="Sağkalım Eğrisi",
        legend.title="Neoadjuvan Uygulanımı", legend.labs=c("Evet","Hayır"))
#Cox regresyon
Cox.Model<-coxph(Surv(zaman,olay)~.,data=BreastCancer,x = TRUE)
Cox.Model
```

```
ggforest(Cox.Model,data=BreastCancer)
summary(Cox.Model)
sum.surv <- summary(Cox.Model)
c_indexCox <- sum.surv$concordance
c_indexCox
#Kaplan Meier ve Cox regresyon karşılaştırma#

plot(kaplan.meier, conf.int = F, col = "black", main = "Model Uyumlarının
        Karşılaştırılması",
    xlab = "Zaman,gün", ylab = "Sağkalım oranı")

lines(survfit(Cox.Model, conf.int = F), col = "red")
legend(x = 700, y = 1, legend = c("Kaplan-Meier", "Cox-PH"), lty = 1,
      col = c("#238b45", "red"),
      cex = 1, bty = "n")
#randomforest
set.seed(800)
train <- sample(nrow(BreastCancer), 0.7*nrow(BreastCancer), replace = FALSE)
TrainSet <- BreastCancer[train,]
ValidSet <- BreastCancer[-train,]
summary(TrainSet)
summary(ValidSet)
fitform1 <-
        Surv(zaman,olay)~Yas+MenarsYas+OrtEmzirmeYili+DogumKontrolHapi
        Kull+TümörTani+NeoadjuvanKull
set.seed(123)
fit<-rfsrc(fitform1, data = TrainSet, ntree = 80,splitrule = "logrank",importance =
        TRUE)
plot(fit)
get.cindex(time = TrainSet$zaman, censoring = TrainSet$olay, predicted =
        fit$predicted.oob)
```

```
plot.survival.rfsrc(fit,plots.one.page = FALSE,cens.model = "rfsrc")
#Prediction error curve
extends <- function(...) TRUE
library("doMC")
library("pec")
library("survival")
library(Rcpp)
registerDoMC()
set.seed(0692)
fitpec1 <- pec(list("CPH" = Cox.Model, "RSF" = fit), data = BreastCancer,
          formula = fitform1, splitMethod = "cv10", B = 6,
          keep.index = TRUE, keep.matrix = TRUE)
plot(fitpec1, what = "crossvalErr", xlim = c(0, 1022), legend = F)
legend(x = 890, y = 0.30, legend = c("KM", "CPH", "RSF"), lty = 1,
     col = c("black", "red", "green"), bty = "n", cex = 1, lwd = 2)
title("Comparison of Prediction Error Curves", line = 1, cex = 6)
#C-Index
startTime <- Sys.time()
set.seed(0692)
ApparrentCindex1 <- cindex(list("Cox" = Cox.Model, "RSF" = fit),
                  formula = fitform1, data = BreastCancer,
                  eval.times = seq(1, 1022, 1))
endTime <- Sys.time()
(totalRunTime <- endTime - startTime)
plot(ApparrentCindex1, legend = F, xlim=c(0,1000),ylim=c(0,1.0),col = c("red",
          "green"))
legend(x = 800, y = 1, legend = c("CPH", "RSF"), lty = 1, col = c("red", "green"), bty
          = "n", cex = 1, lwd = 2)
title("Comparison of Concordance", line = 1, cex = 6)
```