DOKUZ EYLÜL ÜNİVERSİTESİ

FEN BİLİMLERİ ENSTİTÜSÜ

MAKİNE ÖĞRENİMİ YÖNTEMLERİ İLE SAĞKALIM ANALİZİ

Akın Çağatay ÇALIŞKAN

Eylül, 2024

İZMİR

MAKİNE ÖĞRENİMİ YÖNTEMLERİ İLE SAĞKALIM ANALİZİ

Dokuz Eylül Üniversitesi Fen Bilimleri Enstitüsü
Yüksek Lisans Tezi
İstatistik, Anabilim Dalı Veri Bilimi Programı

Akın Çağatay ÇALIŞKAN

Eylül, 2024

İZMİR

YÜKSEK LİSANS TEZİ SINAV SONUÇ FORMU

AKIN ÇAĞATAY ÇALIŞKAN tarafından **DR. ÖĞR. ÜYESİ AYŞE ÖVGÜ KINAY** yönetiminde hazırlanan **"MAKİNE ÖĞRENİMİ YÖNTEMLERİ İLE SAĞKALIM ANALİZİ"** başlıklı tez tarafımızdan okunmuş, kapsamı ve niteliği açısından bir Yüksek Lisans tezi olarak kabul edilmiştir.

Dr.Öğr.Üyesi Ayşe Övgü KINAY

Danışman

Prof. Dr. Selma GÜRLER

Jüri Üyesi

Prof. Dr. Gözde ULUTAGAY

Jüri Üyesi

Prof. Dr. Okan FISTIKOĞLU

Müdür

Fen Bilimleri Enstitüsü

TEŞEKKÜR

Bu çalışmanın geliştirilmesinde, yüksek lisans süresinin yanı sıra lisans hayatım sürecinde de bana her zaman destek olan, bilgilerini daima benimle paylaşan ve öğrenim hayatımın yanı sıra her zaman bana destek olan ve bitmeyen sabrıyla daha da ileriye gitmemi sağlayan çok değerli danışmanım Dr.Öğr.Üyesi Ayşe Övgü KINAY'a, hayatımın son on yılı boyunca her koşulda beni destekleyen ve motive eden sevgili dostum Müge OLUÇOĞLU'na ve ilk adımımdan itibaren maddi ve manevi her koşulda her zaman yanımda olan ve başarılarımın kaynağı aileme sonsuz teşekkürlerimi sunarım.

Akın Çağatay ÇALIŞKAN

MAKİNE ÖĞRENİMİ YÖNTEMLERİ İLE SAĞKALIM ANALİZİ

ÖZ

Sağkalım analizi veya olay-zaman analizi, bir dizi gözlemi ele alan ve bu gözlemlerle ilgi duyulan olayın meydana gelmesi için geçen süreyi tahmin etmeye çalışan istatistiksel yöntemlerdir. Sağkalım analiz yöntemleri genel olarak istatistiksel yöntemler ve makine öğrenimine dayalı yöntemler olmak üzere iki kategoriye ayrılmaktadır. Her iki yöntemin de hedefi hayatta kalma süresini tahmin etmek ve hayatta kalma süresi boyunca sağkalım olasılığını tahmin etmektir. Son yıllarda makine öğrenimi yöntemlerinin geleneksel yöntemler ile birleştirilmesi ve optimizasyonundaki gelişmeler, makine öğrenimi yöntemlerinin popülerliğini artırmaktadır. Mekanik sistemlerdeki başarısızlıklar ve özellikle tıp alanında biyolojik organizmalarda sağkalım analizlerinde yaygın olarak kullanılmaktadır.

Bu çalışmada üç farklı veri seti makine öğrenimi yöntemlerinden biri olan Rastgele Sağkalım Ormanı (RSO) (Random Survival Forest) algoritması, Ekstra Sağkalım Ağaçları (ESA) (Extra Survival Trees) ve Gradyanla Güçlendirilmiş Modeller (Gradient Boosting Models) ile karşılaştırılmıştır. Kullanılan üç farklı veri setinden birinci veri setinde 859 kayıt, 27 değişken bulunurken, ikinci veri seti 9037 kayıt ile 28 değişkene sahiptir. Üçüncü veri setinde ise 14294 kayıt ile 3 farklı değişken bulunmaktadır. Kullanılan üç farklı veri setiyle, farklı örneklem boyutlarının ve farklı değişken sayılarının her bir algoritma üzerindeki etkileri incelenirken, tahminleme başarıları da ortaya koyulmuştur. Python programlama dili kullanılarak gerçekleştirilen analizler sonucunda bu veri seti için RSO algoritması, ESA algoritmasıyla yaklaşık sonuçlar elde ederken, GGM algoritmasından daha iyi bir performans göstermiştir.

Anahtar kelimeler: Sağkalım analizi, rassal sağkalım ormanı, ekstra sağkalım ağaçları, gradyanla güçlendirilmiş modeller, makine öğrenimi

SURVIVAL ANALYSIS WITH MACHINE LEARNING METHODS

ABSTRACT

Survival analysis or event-time analysis are statistical methods that take a series of observations and try to estimate the time it takes for the event of interest to occur. Survival analysis methods are generally categorized into two groups: statistical methods and machine learning-based methods. The goal of both methods is to estimate the survival time and the probability of survival during the survival time. In recent years, the combination and optimization of machine learning methods with traditional approaches have led to an increase in the popularity of machine learning techniques. It is widely used in failures in mechanical systems and survival analyses in biological organisms, especially in the medical field.

In this study, three different datasets were compared using Random Survival Forest (RSF), Extra Survival Trees (EST), and Gradient Boosting Models (GBM), which are among the machine learning methods. The first dataset contains 859 records and 27 variables, while the second dataset consists of 9037 records and 28 variables. The third dataset includes 14294 records with 3 different variables. The study examined the impact of different sample sizes and variable counts on each algorithm and evaluated their prediction accuracies using these three different datasets. The analysis, conducted using the Python programming language, showed that the RSF algorithm achieved similar results to the EST algorithm and performed better than the GBM algorithm for this dataset.

Keywords: Survival analysis, random survival forest, extra survival trees, gradient boosting models, machine learning

İÇİNDEKİLER

Sayfa

YÜKSEK LİSANS TEZİ SINAV SONUÇ FORMU ii
TEŞEKKÜR iii
ÖZ iv
ABSTRACT v
İÇİNDEKİLER vi
ŞEKİLLER LİSTESİ viii
TABLOLAR LİSTESİ x
KISALTMALAR xi

BÖLÜM BİR - GİRİŞ 1

1.1 Sağkalım Analizi 1
1.2 Sansürlü Veri 3

BÖLÜM İKİ - NOTASYONLAR VE FONKSİYONLAR 5

2.1 Notasyonlar 5
2.2 Fonksiyonlar ve Performans Metriği 5
2.2.1 Sağkalım Fonksiyonu 5
2.2.2 Tehlike Fonksiyonu 7
2.2.3 Performans Metrikleri 9

BÖLÜM ÜÇ - SAĞKALIM ANALİZİ İSTATİSTİKSEL YÖNTEMLERİ . 12

3.1 Sağkalım Analizi Yöntemlerinin Taksonomisi 12
3.2 Geleneksel İstatistiksel Yöntemler 13
3.2.1 Parametrik Olmayan Yöntemler 15
3.2.2 Yarı Parametrik Yöntemler 16

BÖLÜM DÖRT - MAKİNE ÖĞRENMESİ YÖNTEMLERİ 17

4.1 Sağkalım Ağaçları 18
4.2 Gelişmiş Makine Öğrenimi Yöntemleri 19

4.2.1 Rassal Orman (Random Forest) 19
4.2.2 Rassal Sağkalım Ormanı (Random Survival Forest) 20
4.2.3 Ekstra Sağkalım Ağaçları (Extremly Randomized (Extra) Survival Trees) 22
4.2.4 Gradyanla Güçlendirilmiş Modeller (Gradient Boosted Models) 22

BÖLÜM BEŞ - ANALİZ 26

5.1 Uygulama 1 – Meme Kanseri Hastaları Sağkalım Analizi 26
5.1.1 Sağkalım Olasılıkları ve Tehlike Oranları 29
5.1.2 C-İndeks Sonuçları ve Çalışma Süreleri 32
5.2 Uygulama 2 – Hastanede Tedavi Gören Hastalar İçin Sağkalım Analizi 33
5.2.1 Sağkalım Olasılıkları ve Tehlike Oranları 36
5.2.2 C-indeks Değerleri ve Çalışma Süreleri 40
5.3 Uygulama 3 – Prostat Kanseri Hastaları İçin Sağkalım Analizi 41
5.3.1 Sağkalım Olasılıkları ve Tehlike Oranları 42
5.3.2 C-indeks Değerleri ve Çalışma Süreleri 46

BÖLÜM ALTI SONUÇ 48

KAYNAKLAR 50

ŞEKİLLER LİSTESİ

Sayfa

Şekil 1.1 Sansürlü verilere ait bir örnek (Wang vd., 2019)..... 4

Şekil 2.1 Teorik sağkalım fonksiyonu (Kleinbaum ve Klein, 2012) 6

Şekil 2.2 Pratikte sağkalım fonksiyonu (Kleinbaum ve Klein, 2012)..... 6

Şekil 2.3 Tehlike fonksiyonları grafiği (Kleinbaum ve Klein, 2012) 8

Şekil 3.1 İstatistiksel yöntemlerin taksonomisi..... 13

Şekil 4.1 Makine öğrenimi yöntemleri taksonomisi 17

Şekil 4.2 Rassal sağkalım ormanı algoritması akış şeması 21

Şekil 4.3 GGM yöntemi akış şeması..... 25

Şekil 5.1 Uygulama 1'de kullanılan veri setine ait kayıtların yaş dağılımı 27

Şekil 5.2 Uygulama 1'de kullanılan veri setine ait kanser evrelerinin dağılımı 27

Şekil 5.3 Uygulama 1 RSO için hastaların sağkalım olasılıkları grafiği 30

Şekil 5.4 Uygulama 1 RSO için hastaların kümülatif tehlike oranları grafiği 30

Şekil 5.5 Uygulama 1 ESA için hastaların sağkalım olasılıkları grafiği..... 31

Şekil 5.6 Uygulama 1 ESA için hastaların tehlike oranları grafiği..... 31

Şekil 5.7 Uygulama 1 GGM için hastaların sağkalım olasılıkları grafiği..... 32

Şekil 5.8 Uygulama 1 GGM için hastaların kümülatif tehlike oranları grafiği 32

Şekil 5.9 Uygulama 2'de kullanılan veri setinde yer alan hastaların yaş dağılımları. 34

Şekil 5.10 Uygulama 2'de kullanılan veri setinde yer alan hastaların hastalık sınıfları dağılımları 34

Şekil 5.11 Uygulama 2 RSO yöntemi için hastaların sağkalım olasılıkları grafiği ... 37

Şekil 5.12 Uygulama 2 RSO için hastaların tehlike oranları grafiği..... 38

Şekil 5.13 Uygulama 2 ESA için hastaların sağkalım olasılıkları grafiği..... 38

Şekil 5.14 Uygulama 2 ESA için hastaların tehlike oranları grafiği..... 39

Şekil 5.15 Uygulama 2 GGM için hastaların sağkalım olasılıkları grafiği..... 39

Şekil 5.16 Uygulama 2 GGM için hastaların tehlike oranları grafiği..... 40

Şekil 5.17 Uygulama 3'te kullanılan veri setine ait yaş gruplarının dağılımı 41

Şekil 5.18 Uygulama 3'te kullanılan veri setine ait kanser evrelerinin dağılımı..... 42

Şekil 5.19 Uygulama 3 RSO için hastaların sağkalım olasılıkları grafiği 44

Şekil 5.20 Uygulama 3 RSO için hastaların tehlike oranları grafiği..... 44

Şekil 5.21 Uygulama 3 ESA için hastaların sağkalım olasılıkları grafiği..... 45

Şekil 5.22 Uygulama 3 ESA için hastaların tehlike oranları grafiği..... 45

Şekil 5.23 Uygulama 3 GGM için hastaların sağkalım olasılıkları grafiği..... 46

Şekil 5.24 Uygulama 3 GGM için hastaların tehlike oranları grafiği..... 46

TABLOLAR LİSTESİ

Sayfa

Tablo 3.1 Sağkalım analizine yönelik farklı istatistiksel yöntem türlerinin özeti (Wang vd, 2019) 14

Tablo 5.1 Uygulama 1'de kullanılan veri setine ait değişkenlerin bilgi tablosu..... 28

Tablo 5.2 Uygulama 1 için yöntemlerin C-indeks değerleri tablosu 33

Tablo 5.3 Uygulama 2'de kullanılan veri setine ait değişkenlerin bilgi tablosu..... 35

Tablo 5.4 Uygulama 2 için yöntemlerin C-indeks değerleri tablosu 41

Tablo 5.5 Uygulama 3'te kullanılan veri setine ait değişkenlerin bilgi tablosu..... 42

Tablo 5.6 Uygulama 3 için yöntemlerin C-indeks değerleri tablosu 47

KISALTMALAR

ESA	: Ekstra Sağkalım Ağaçları
GGM	: Gradyanla Güçlendirilmiş Modeller
KM	: Kaplan Meier
RSO	: Rassal Sağkalım Ormanı
YT	: Yaşam Tablosu

BÖLÜM BİR
GİRİŞ

1.1 Sağkalım Analizi

Sağkalım analizi, ilgilenilen sonuç değişkeninin bir olay meydana gelene kadar geçen süre olduğu, özellikle tıp alanında yaygın bir şekilde kullanılan istatistiksel yöntemlerdir. Olay-zaman analizi olarak da bilinen bu yöntemlerde ilgilenilen sonuç sadece hastanın ölümüne kadar geçen süre değildir. Bir kırık iyileşene kadar geçen zaman, bir hastalığa dair belirli bir semptom görülene kadar geçen zaman, ilacın etki etme süresi vb. olabileceği gibi tıp alanı dışında mekanik bir parçanın bozulma tahmini, sigorta poliçesinin iptal edilme zamanı veya ekonomide işsizlik süresi tahmini gibi çok farklı uygulamalarda da sağkalım analizi yöntemleri kullanılmaktadır.

Sağkalım analizinin günümüze kadar olan gelişimi 1662 yılında John Graunt tarafından geliştirilen yaşam tablosu (life table) yöntemiyle 17. yüzyıla kadar uzanmaktadır. Yüzyıllar boyunca, yalnızca ölüm oranlarının araştırmasıyla ilişkili olan sağkalım analizi son birkaç yılda gerçekleşen gelişmelerle birlikte biyomedikal alandaki araştırmaların yanı sıra kriminoloji, sosyoloji, pazarlama ve sağlık sigortası gibi birçok alana yayılmıştır (Liberato Camilleri, 2019).

Sağkalım analizi ölüm, uyuşturucu bağımlılığı ya da elektronik cihazların arızalanması gibi zaman içinde gözlemlenen olaylarla ilgili verilerin analizinde kritik bir rol oynamaktadır. Önemli risk faktörlerinin yanı sıra hayatta kalma olasılığı, tehlike oranlarının tahmini gibi bilgiler sağlık alanında çalışan cerrahlar, psikologlar ve mekanik alanda çalışkan mühendisler için önemli bilgiler sağlamaktadır (Liberato Camilleri, 2019).

17. yüzyılda John Graunt'un yaşam tablolarıyla gelişmeye başlayan sağkalım analizi 1958 yılında Kaplan ve Meier'in hayatta kalma olasılıkları ve tehlike oranlarını tahmin etme konusunda yaptığı çalışmalar ve katkılar sağkalım analizi alanında çığır açan bir gelişme oldu. Bu çalışmaları Nelson-Aalen kümülatif tehlike oranı tahmincisi takip etti. Bu çalışmaların yanı sıra 1972 yılında ortaya atılan Cox orantılı tehlike modeli sağkalım analizine önemli katkı sağlamıştır. (Liberato Camilleri, 2019).

Son yıllarda sağkalım analizinde gelişmeler özellikle makine öğrenimi ve derin öğrenme alanında olmuştur. Özellikle büyük veriler ve sansürlü verilerle daha iyi başa çıkabilen bu yöntemler sağkalım analizine önemli katkılar sağlamıştır. Literatürde sağkalım analizi alanında yapılan birçok çalışma yer almaktadır. Wang ve arkadaşları (2019) sağkalım analizi hakkında makine öğrenimi yöntemlerinin yanı sıra geleneksel istatiksel yöntemler hakkında da birçok bilgiye yer verilmiş bir çalışma yayınlamışlardır. Bunun yanı sıra Spooner ve arkadaşları yüksek boyutlu veri setleri üzerinde makine öğrenimi yöntemlerinin performanslarını ölçmüşlerdir. 251 değişkene sahip MAS (Memory and Ageing Study) ve 185 farklı değişkeni bulunan ADNI (Alzheimer's Disease Neuroimaging Initiative) veri setleri ile yapılan çalışmada makine öğrenimi alanında yaygın olarak kullanılan RSO (Rassal Sağkalım Ormanı) yöntemi için sırasıyla 0.797 ve 0.924 doğruluk oranlarıyla başarılı sonuçlara ulaşmışlardır. Augusto Di Castelnuovo vd. (2020) 3894 COVID-19 hastasının kardiyovasküler bir riske sahip olmaları üzerine yaptığı çalışmanın yanı sıra, tiroid kanseri hastalarının makine öğrenimi ile sağkalım analizi (Alhashmi vd., 2024), kolorektal kanser hastalarının sağkalım analizi (Buk Cardoso vd., 2023) ,klinik verilerine dayalı akciğer kanseri sağkalım analizi üzerine Fatimah Abdulazim Altuhaifa ve arkadaşlarının (2023) yaptığı çalışma gibi sağkalım analizi alanında literatürde sağlık alanında makine öğrenimi yöntemlerinin kullanıldığı ve incelendiği çalışmalar yer almaktadır. Bunun yanı sıra sağlık alanı dışında yer alan sağkalım analizi ile makine arıza tahmini (Dimitris Papathanasiou vd., 2023) gibi çalışmalar da vardır. Diego Vallarino'nun (2024) yayınlamış olduğu çalışmada Amerika kıtasında yer alan 24 ülkenin yer aldığı 28 yıllık MVI (Multidimensional Vulnerability Index) veri setindeki bilgilerden faydalanarak 120 aylık dönemde işsizliğin %10 artması için geçen süreyi tahminlemek için farklı makine öğrenimi ve derin öğrenme yöntemlerinin tahmin becerilerini incelemiştir. Ambra Macis'in (2023) doktora çalışmasında ise NBA oyuncularının performanslarının değerlendirmesini sağkalım analizi ile yapmıştır. Bu çalışma sağkalım analizinin sağlık alanı dışındaki yer alan çalışmaları arasında ilginç ve iyi bir çalışma örneğidir.

Bu tez çalışmasında da çeşitli makine öğrenimi yöntemleri incelenmiş ve sağlık alanında elde edilmiş 3 farklı veri seti üzerinde 3 farklı makine öğrenimi yöntemi kullanılmıştır. 3 farklı veri seti kullanılmasının sebebi benzer değişken sayısına sahip

farklı boyuttaki veri setlerinin performanslarının karşılaştırılması ve daha az değişkene sahip yüksek boyutlu veri setinin yöntemler üzerindeki çalışma performanslarının karşılaştırılması amaçlanmıştır.

İkinci bölümde sağkalım analizi notasyonları ve önemli iki fonksiyonu olsan sağkalım fonksiyonu ile tehlike fonksiyonu açıklanmıştır.

Üçüncü bölümde sağkalım analizinde kullanılan geleneksel istatistiksel yöntemlerden bahsedilmiştir.

Dördüncü bölümde ise sağkalım analizinde kullanılan makine öğrenimi yöntemleri açıklanmıştır.

Beşinci uygulamada ise sağkalım analizinde kullanılan 3 makine öğrenimi yönteminin 3 farklı uygulama ile testleri yapılıp sonuçları karşılaştırılmıştır.

Altıncı bölümde ise bu tezde yapılan çalışmaların sonuçları yer almaktadır.

Son olarak tezin son kısmında, bu çalışmada yer alan akademik çalışmalar listelenmiştir.

1.2 Sansürlü Veri

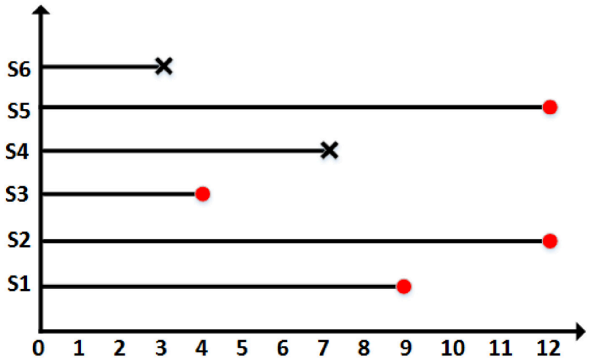
Birçok sağkalım analizinde sansürlü (censored) veri olarak adlandırılan önemli bir problem yer almaktadır. Sansürlü veri (censored), sağkalım süresi hakkında bazı bilgilere sahip olduğumuz anca sağkalım süresini tam olarak bilmediğimiz zaman ortaya çıkmaktadır (Kleinbaum ve Klein, 2012).

Sansürün ortaya çıkmasının genellikle üç nedeni vardır:

- Çalışma süresi boyunca kişinin olayı yaşamaması
- Çalışma süresi boyunca kişi takip dışı kalması
- Çalışması sırasında kişi ilgilenin olaydan farklı bir nedenden dolayı ölmesi ya da ilaç yan etkisi gibi nedenlerden dolayı çalışmadan çekilmesi

Sansür üç temel kategoriden oluşmaktadır. Bunlar sağdan sansürlü, soldan sansürlü ve aralıklı sansürdür.

Sağdan sansür, gerçek hayatta kalma süresi (T), gözlemlenen hayatta kalma süresine (t) eşit veya bundan daha fazla olmasıdır ($T \geq t$). Sağkalım süreci üzerinde çalışılan nedenden dolayı değil de farklı bir nedenle sonlanmışsa bu veri sağdan sansürlü veridir. Soldan sansür ise gerçek hayatta kalma süresi (T), gözlemlenen hayatta kalma süresine (t) eşit veya daha az olmasıdır ($T \leq t$). Son olarak aralıklı sansür: olay gerçek hayatta kalma süresi bilinen bir zaman aralığı dahilindedir (Kleinbaum ve Klein, 2012).



Şekil 1.1 Sansürlü verilere ait bir örnek (Wang vd., 2019)

Şekil 1.1'de sansürlü verilere ait bir örneğin grafiği yer almaktadır. 6 farklı hastanın 12 aylık süreçteki takiplerine ait bu kayıtlar şöyledir:

- X ile işaretlenmiş olan 4. ve 6. gözlemler beklenen olayı yaşamıştır.
- 1, 2, 3 ve 5. kayıtlar12 aylık süreçte gerçekleşmediği için sansürlü veridir.

Daha spesifik olmak gerekirse kayıt 2 ve 5 çalışma süresi boyunca beklenen olay gerçekleşmediği için sansürlenirken 1 ve 3 numaralı kayıtlar çalışmadan ayrılma ya da farklı bir nedenden dolayı çalışma dışı kaldıkları için sansürlü verilerdir.

BÖLÜM İKİ
NOTASYONLAR VE FONKSİYONLAR

2.1 Notasyonlar

Sağkalım fonksiyonlarında da kullanılan büyük T, bir kişinin hayatta kalma süresine ilişkin rastgele değişkendir. T zamanı gösteren bir değişken olduğu için negatif olmayan herhangi bir sayı olabilir. Yani T sıfıra eşit ya da sıfırdan büyük herhangi bir sayıdır. Bir diğer notasyonumuz küçük t ise T rastgele değişkeni için ilgilenilen herhangi bir spesifik değerdir. Örneğin, bir kişinin kanser tedavisi sonrasında 5 yıldan fazla hayatta kalıp kalmadığı incelenmek isteniyorsa t değeri 5'e eşittir. Son değişken ise başarısızlığı veya sansürü $(0,1)$ gösteren küçük d değişkendir. Olay çalışma süresinde gerçeklemesi durumunda başarısızlık $d = 1$ olarak gösterilirken, hayatta kalma süresi çalışma süresi sonunda sansürlenirse $d = 0$ olarak gösterilmektedir (Kleinbaum ve Klein, 2012).

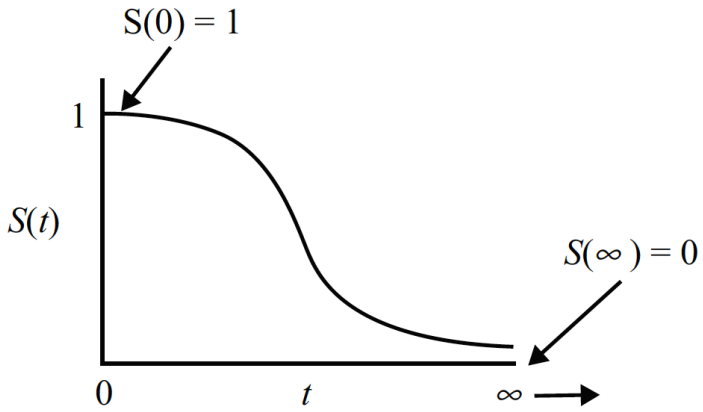
2.2 Fonksiyonlar ve Performans Metriği

2.2.1 Sağkalım Fonksiyonu

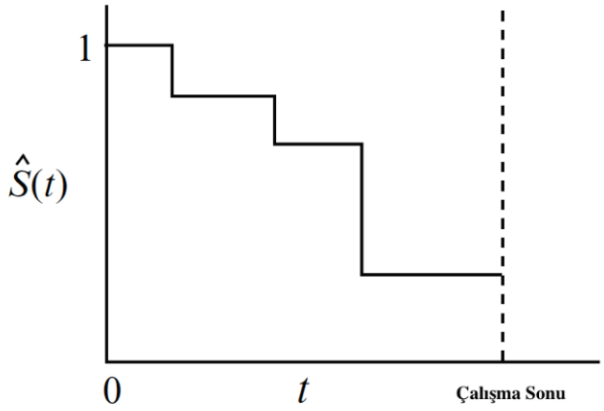
Sağkalım fonksiyonu $S(t)$, rastgele değişken T'nin belirtilen t süresini aşmasıdır. İncelenen olaydaki kişinin belirli t süresinden daha üzün süre hayatta kalma olasılığını verir (Kleinbaum ve Klein, 2012).

$$S(t) = P(T > t) \tag{2.1}$$

Sağkalım analizinin temel fonksiyonu sağkalım fonksiyonudur. Çünkü elde edilecek farklı t değerlerinin sağkalım olasılıkları, sağkalım verileri hakkında önemli özet bilgiler elde edilmesine olanak sağlar. Olay başlangıcında $(t = 0)$, hiçbir denek olayı yaşamayacağından dolayı $S(0) = 1$'dir. $S(\infty) = 0$ olacaktır, çünkü çalışma süresinin sonsuz olduğunda çalışmada yer alan deneklerin hepsi beklenen olayı yaşayacaktır.



Şekil 2.1 Teorik sağkalım fonksiyonu (Kleinbaum ve Klein, 2012)



Şekil 2.2 Pratikte sağkalım fonksiyonu (Kleinbaum ve Klein, 2012)

Teorik olarak, $t(0)$ 'dan sonsuza kadar gittiği için, hayatta kalma fonksiyonu düzgün bir eğri grafiğidir. Pratikte, gerçek veriler kullanılırken, genellikle bu düzgün eğriler yerine Şekil 2.1'de görüldüğü gibi basamak fonksiyonlarına sahip olan grafikler elde edilir. Dikkat edilmesi gereken bir diğer unsur ise çalışma süresi hiçbir zaman sonsuz uzunlukta olmayacağından ve gözlemlenen herkesin olayı yaşamaması mümkündür. Bu nedenle Şekil 2.2'de yer alan grafikte görüldüğü gibi tahmini sağkalım fonksiyonu, çalışmanın sonunda sıfıra kadar inmeyebilir (Kleinbaum ve Klein, 2012).

2.2.2 Tehlike Fonksiyonu

Tehlike fonksiyonu $h(t)$, bireyin t zamanına kadar hayatta kaldığı göz önüne alındığında, olayın meydana gelmesi için birim zaman başına anlık potansiyeli verir. Sağkalım fonksiyonu başarıya odaklanırken, tehlike fonksiyonu başarısızlığa, yani meydana gelmesi beklenen olaya odaklanır. Dolayısıyla tehlike fonksiyonu, sağkalım fonksiyonun verdiği bilginin tam tersini verdiği düşünülebilir. Tehlike fonksiyonunda koşullu olasılık, sağkalım süresi t'den büyük büyük veya ona eşit olması koşuluyla, bir kişinin hayatta kalma süresi T'nin t ile $t + Dt$ zaman aralığında yer alma olasılığını verir. Tehlike fonksiyonu ayrıca koşullu başarısızlık oranı olarak da adlandırılır (Kleinbaum ve Klein, 2012).

Matematiksel olarak tehlike fonksiyonu aşağıdaki gibi tanımlanır:

$$h(t) = \lim_{\Delta t \to 0} \frac{\Pr(t \leq T < t + \Delta t | T \geq t)}{\Delta t} = \lim_{\Delta t \to 0} \frac{F(t+\Delta t) - F(t)}{\Delta t . S(t)} = \frac{f(t)}{S(t)} \qquad (2.2)$$

Ayrıca $f(t) = -\frac{d}{dt} S(t)$ şeklinde ifade edilebileceğinden, tehlike fonksiyonu şu şekilde gösterilir:

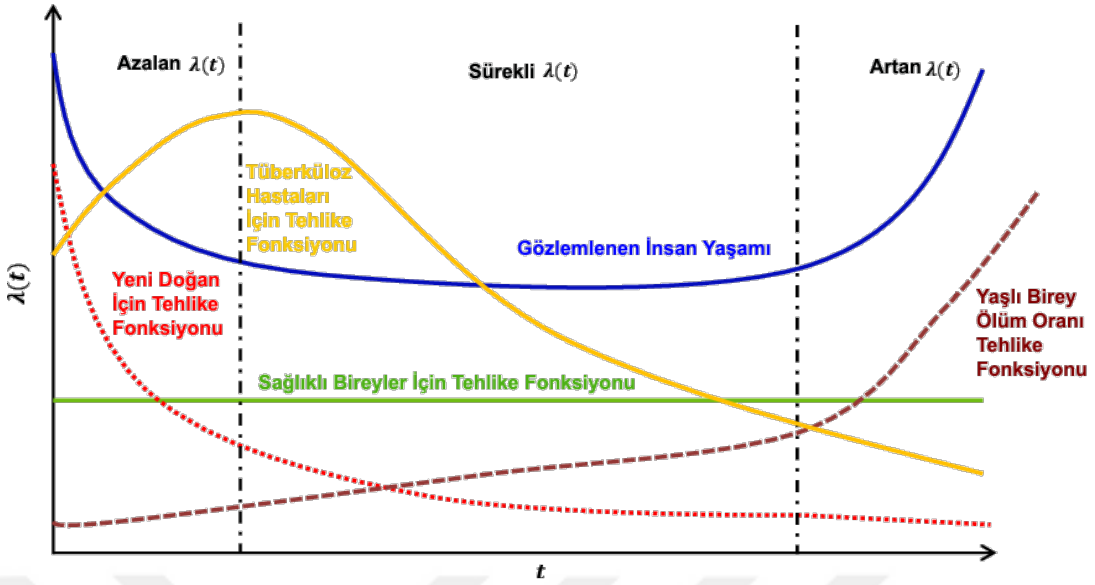
$$h(t) = \frac{f(t)}{S(t)} = -\frac{d}{d(t)} S(t) . \frac{1}{S(t)} = -\frac{d}{dt} [\ln S(t)] \qquad (2.3)$$

Böylece sağkalım fonksiyonu aşağıdaki gibi yeniden yazılabilir:

$$S(t) = \exp\big(-H(t)\big), \qquad (2.4)$$

Burada $H(t) = \int_0^t h(u) du$, kümülatif tehlike fonksiyonunu temsil eder (Lee ve Wang, 2003).

$S(t)$'ye benzer şekilde $h(t)$ de negatif olmayan bir fonksiyondur. $S(t)$ zamanla azalırken tehlike oranı zamanla değiştiği için tehlike fonksiyonları Şekil 2.3'te olduğu gibi çeşitli şekillere sahip olabilir. Herhangi bir yerde başlayıp zaman içinde yükselip azalabilir (Wang vd., 2019).



Şekil 2.3 Tehlike fonksiyonları grafiği (R. Rois, 2017)

Şekil 2.3'te tehlike fonksiyonuna ait bir grafik örneği yer almaktadır. Bu grafikte 5 farklı kayıta ait t zamanındaki tehlike oranları görülmektedir.

Grafikte yer alan kayıtlar için:

- Yeşil renk ile gösterilmiş kayıt 18-40 yaş aralığındaki sağlıklı bir bireylerin t zamandaki tehlike oranını göstermektedir. Sağlıklı bir bireyin tehlike oranı sabittir.
- Mavi renk ile gösterilmiş kayıt ise bir insanın yaşamıdır. Bu kayıtın tehlike oranı yeni doğan ve yaşam süresinin son evrelerinde yüksektir. Çünkü bu dönemlerde ölüm riski daha fazladır. Diğer evrelerde ise daha stabil bir tehlike oranına sahiptir.
- Kırmızı ile gösterilmiş kayıt ise yeni doğana ait tehlike fonksiyonudur. Yeni doğanlarda ölüm riski yüksek olduğu için tehlike oranı da yüksektir. t zamanında bu oran giderek azalmaktadır.
- Kahverengi ile gösterilmiş olan kayıt ise kanser hastası yaşlı bir bireyin t zamandaki tehlike oranıdır. Bireyin kanser hastası ve yaşlı olmasından dolayı ilerleyen sürede tehlike oranı artmaktadır.

- Sarı renk ile gösterilmiş kayıt ise tüberküloz hastasına ait bir tehlike oranı fonksiyondur. Hastanın tedavi başında tehlike oranı yüksekken tedavi ile tehlike oranı da düşmektedir.

2.2.3 Performans Metrikleri

Sağkalım verilerinde sansürlü verilerin bulunmasından dolayı, ortalama kare hatası (mean squared error) ya da R^2 gibi değerlendirme ölçütleri tahmin performansını ölçmek için yeterli değildir. (Heagerty ve Zheng, 2005). Bunun yerine performans ölçümlerinin sağkalım analizi için özelleştirilmiş değerlendirme ölçütleri kullanılarak değerlendirilmesi gerekir. (Wang vd., 2019)

Sağkalım analizinde kullanılan performans metrikleri şunlardır:

- Concordance Index (C-Index)
- Brier Puanı (Brier Score)
- Ortalama Mutlak Hata (Mean Absolute Error (MAE))
- ROC ve AUC
- Discrimination Index (D-Index)
- Restricted Mean Survival Time (RMST)
- Harrell's D Statistic
- Kalibrasyon Eğrileri
- Log-rank testi

Yapılan tahminin doğruluğu vurgulanmak istendiğinde C-İndeksi önde gelen ölçütlerden biri olurken, tahminin kalitesi ve güvenliği ölçülmek istendiğinde ise Brier puanı yöntemi kullanılmaktadır.

Sağkalım analizi alanında yapılan çalışmalarda C-İndeks yöntemi yaygın olarak kullanılmaktadır. Bu tez çalışmasında da performans ölçümlerinde C-İndeks yöntemi kullanılmıştır.

Harrell ve arkadaşları tarafından 1982 yılında geliştirilen C-İndeks metriği özellikle sağkalım analizi alanında modellerin performanslarını değerlendirmek için yaygın olarak kullanılmaktadır. Bu metrik iki örneğin tahmin edilen risk sıralamasının gerçek sonuçlarla ne kadar uyumlu olduğunu ölçer. Sağkalım analizi için önemli bir metrik

olmasının sebebi ise sansürlü verilerin doğrudan değerlendirilememesi dolayısıyla modelin başarı oranı sıralama doğruluğu üzerinden ölçülmesidir. Sansürlü veri sayısı artıkça doğruluk oranı da artmaktadır.

$$C = \frac{Uyumlu\ Çiftlerin\ Sayısı + 0.5\ \times\ Bağlı\ Çiftlerin\ Sayısı}{Karşılaştırılabilir\ Çiftlerin\ Toplam\ Sayısı} \tag{2.5}$$

Uyumlu Çiftler, i ve j gözlemleri arasında:

- $T_i > T_j$ (gerçek sürelerinde) olduğunda,
- $R_i > R_j$ (risk skorları) ise çift uyumludur.
- Risk sıralaması doğru bir şekilde yapıldıysa bu çift uyumludur.

Bağlı Çiftler:

- Eğer $T_i = T_j$ veya $R_i = R_j$ ise bu çiftler bağlıdır.
- Bağlı çiftlerin ağırlığı yarım (0.5) olarak alınmaktadır.

Karşılaştırılabilir Çiftler:

- Sansürlenmiş verilerde göz önünde bulundurularak, bir gözlemin diğeriyle sıralanmasının mümkün olduğu çiftler karşılaştırılabilir çiftlerdir.
- T_i sansürlü bir veri ise ve $T_i > T_j$ hakkında bir bilgi yoksa bu çift karşılaştırılamaz.

C-indeks metriğinden elde edilen değer aralığı $C \in [0.5,1]$'dir.

- 0.5 değeri rastgele tahmin anlamına gelir. Yani modelin performansı kötüdür ve hiçbir ayrım gücü yoktur.
- 0.5-0.7 değer aralığında ise modelin performansı zayıftır. Risk farkları tahmin edilebilse de bu farklar belirgin değildir.
- 0.7-0.8 model performansı orta düzeydedir. Riskler makul düzeyde ayrılabilir.
- 0.8-0.9 model performansı iyidir. Model sıralama tahminlerinde yüksek bir doğruluğa sahiptir.
- 0.9-1.0 model performansı çok iyi ya da mükemmeldir. Tüm risk grupları doğru şekilde ayrılmıştır ve sıralama tahminlerinin neredeyse tamamı doğrudur.

- 1.0 değeri mükemmel model anlamına gelirken, pratikte genellikle rastlanılmayan 0.0 değerinde modelin tahminleri ters yöndedir ve yüksek riskte olması gereken gözlemleri düşük riskte olarak sıralar.

BÖLÜM ÜÇ
SAĞKALIM ANALİZİ İSTATİSTİKSEL YÖNTEMLERİ

3.1 Sağkalım Analizi Yöntemlerinin Taksonomisi

Sağkalım analizi yöntemlerini genel olarak istatistiksel yöntemler ve makine öğrenimine dayalı yöntemler olmak üzere iki ana kategoriye ayrılır. İstatistiksel yöntemler ve makine öğrenimine dayalı yöntemler ortak hedefleri paylaşırlar. Her iki yönteminde hedefi hayatta kalma süresini tahmin etmek ve hayatta kalma süresi boyunca sağkalım olasılığını tahmin etmektir. İstatistiksel yöntemler sağkalım eğrilerini tahminlerken, sağkalım olay zamanlarını ve parametre tahminin istatistiksel özelliklerini karakterize etmeye odaklanır. Makine öğrenimi yöntemleri ise makine öğrenimi yöntemlerine geleneksel sağkalım yaklaşımını dahil ederek olayın belirli bir zamanda gerçekleşmesini tahmin edilmesine odaklanır. İstatistiksel yöntemler genellikle düşük boyutlu problemlere uygulanırken, makine öğrenmesine dayalı yöntemler genellikle yüksek boyutlu problemlere uygulanmaktadır. Ayrıca sağkalım analizi için makine öğrenmesine dayalı yöntemler, sağkalım problemlerini hem istatistikler yöntemlerle hem de makine öğrenmesi yöntemleriyle birleştirerek ortak değişkenler ile sağkalım süreleri arasındaki ilişkileri farklı şekillerde öğrenmek için makine öğrenmesi ve makine öğrenmesi optimizasyonundaki son gelişmelerden yararlanarak daha etkili algoritmalar sunar.

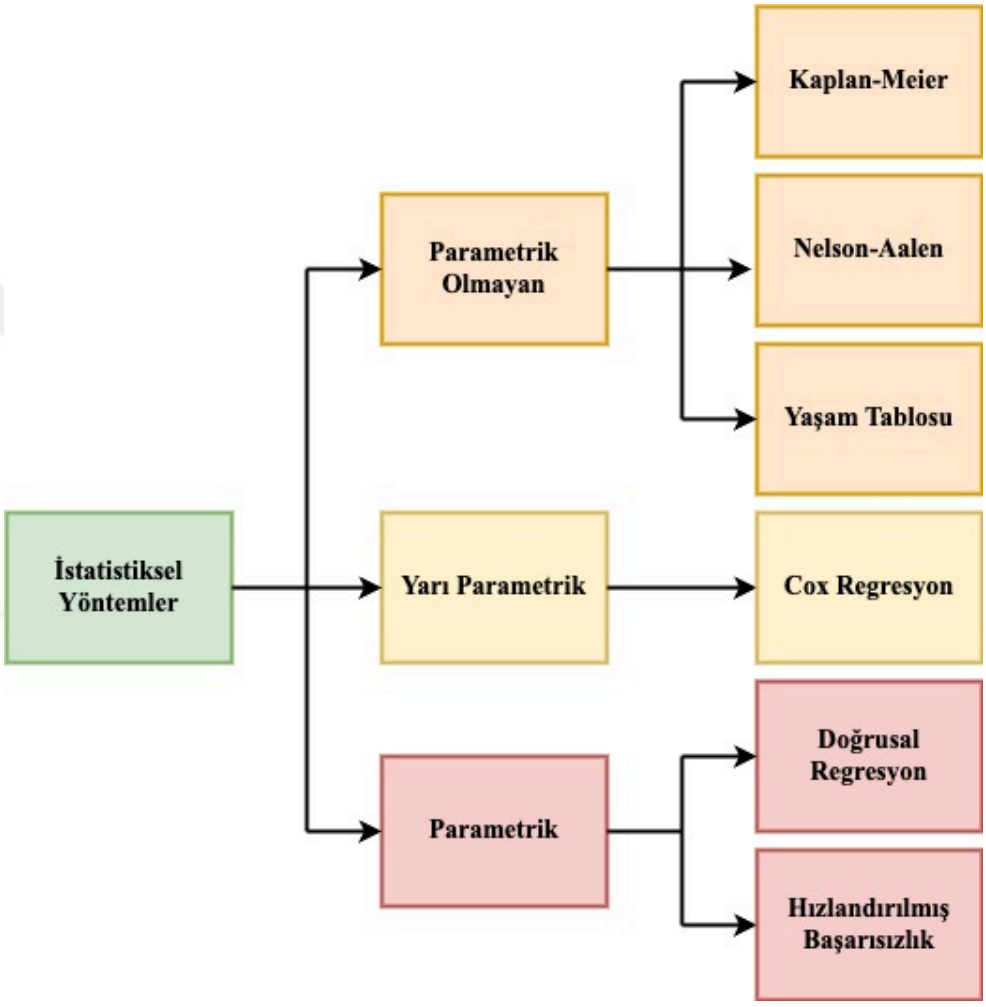
Modelde kullanılan parametrelerin varsayımlarına ve kullanımlarına göre, geleneksel istatistiksel yöntemleri ana üç kategoride inceleyebiliriz:

- Parametrik olmayan istatiksel yöntemler
- Yarı-Parametrik istatiksel yöntemler
- Parametrik istatiksel yöntemler

Son yıllarda daha popüler hale gelen sağkalım ağaçları (survival trees), bayesian yöntemleri, sinir ağları ve destek vektör makineleri gibi makine öğrenimi algoritmaları ayrı bir dal altında yer almaktadır. Gelişmiş makine öğrenimi yöntemlerinden olan topluluk öğrenimi (ensemble learning), aktif öğrenme (active learning), transfer öğrenimi (transfer learning) ve çok görevli öğrenme (multi-task learning) yöntemleri de dahildir. Wang vd., 2019).

3.2 Geleneksel İstatistiksel Yöntemler

Bu bölümde sağkalım fonksiyonu ve tehlike fonksiyonlarını tahmin etmek için kullanılan parametrik olmayan, yarı parametrik ve parametrik yöntemler olmak üzere üç geleneksel istatistiksel yöntemler tanıtılmaktadır. Bu yöntemlerin Şekil 3.1'de görülebilirken, yöntemlerin genel karşılaştırılması Tablo 3.1'de yer almaktadır.



Şekil 3.1 İstatistiksel yöntemlerin taksonomisi

Parametrik olmayan yöntemler, olay zamanı içi temel bir dağılım olmadığında veya orantılı tehlike varsayımı olmadığında daha etkilidir. Parametrik olmayan yöntemlerde, Kaplan-Meier yöntemi, Nelson-Aalen tahmincisi veya Yaşam Tablosu yöntemi kullanılarak hayatta kalma fonksiyonun ampirik bir tahmini elde edilir.

Yarı parametrik yöntemlerde, Cox modeli sağkalım çalışmalarında sıklıkla tercih edilen bir regresyon analizidir. Orantısal tehlike ve parametrelerin tahminlenmesinde kısmi olabilirlik fonksiyonunda yararlandığı için diğer yöntemlere göre büyük farklılık göstermektedir. Cox regresyonu, parametrik bir regresyon modeline dayansa bile sonucun dağılımı hakkında bilgi sahibi olunmadığı için yarı parametrik yöntemler arasında yer almaktadır. Literatürde Cezalandırılmış Cox Modelleri (Zhang ve Lu, 2007), CoxBoost algoritması (Binder ve Schumacher, 2008) ve Zamana Bağlı Cox (TD-COX) modeli (Kleinbaum ve Klein, 2006) gibi temel Cox modelinin çeşitli varyantları önerilmiştir.

Parametrik yöntemler, ilgilenilen olaya kadar geçen süre belirli parametreler açısından belirtilebilen belirli bir dağılımı izlediğinden tahmin yapmak için daha verimli ve doğrudur. Parametrik modellerle ilgili olaya kadar geçen süreyi tahminlemek daha kolaydır, fakat Cox regresyon analiziyle tahmin yapmak zor hatta imkânsız hale gelir. Parametrik sağkalım yöntemlerinde en çok kullanılan yöntem Doğrusal regresyondur. Bu yöntemin yanı sıra Tobit modeli, Buckley-James regresyon modeli ve cezalandırılmış regresyon modelleri de sağkalım analizinde sıklıkla kullanılan parametrik yöntemler arasında yer almaktadır. Sağkalım sürelerini ortak değişken fonksiyonu olarak modelleyen Hızlandırılmış Arıza Süresi (Accelerated Failure Time) gibi farklı parametrik modellerde (Kelinbaum ve Klein, 2006) yaygın olarak kullanılmaktadır (Wang vd., 2019).

Tablo 3.1 Sağkalım analizine yönelik farklı istatistiksel yöntem türlerinin özeti (Wang vd, 2019)

Tip	**Avantajlar**	**Dezavantajlar**	**Spesifik Yöntemler**
Parametrik Olmayan	Uygun teorik dağılımlar bilinmediğinde daha etkilidir.	Yorumlaması zor; hatalı tahminler verir.	Kaplan-Meier Nelson-Aalen Life-Table
Yarı Parametrik	Sağkalım sürelerinin temel dağılımının bilinmesi gerekli değildir.	Yorumlaması zor; sonucun dağılımı bilinmiyor.	Cox modeli CoxBoost Zamana Bağlı Cox
Parametrik	Sağkalım süreleri belirli bir dağılımı izlediğinde yorumlaması kolay, daha verimli ve doğrudur.	Dağılım varsayımı ihlal edildiğinde, tutarsız olabilir ve optimal olmayan sonuçlar verebilir.	Tobit Buckley-James Cezalandırılmış Regresyon Hızlandırılmış Arıza Süresi

3.2.1 Parametrik Olmayan Yöntemler

Tüm fonksiyon seçenekleri arasında, sağkalım fonksiyonu ya da onun grafiksel gösterimi en yaygın kullanılandır. 1958 yılında Kaplan ve Meier, olayın gerçek gözlem süresini kullanarak sağkalım fonksiyonunu tahmin etmek için Çarpım-Limit (Product Limit) tahmincisi olarak da bilinen Kaplan-Meier eğrisini geliştirdi (Wang vd., 2019).

Kaplan-Meier (KM) tahmincisi gözlem sürelerine ilişkin, verileri zaman aralıklarına bölmeden sağkalım fonksiyonunu hesaplamaktadır. KM tahmincisi hayatta kalma fonksiyonunu tahmin etmek için en yaygın kullanılan yöntemdir.

$T_1 < T_2 < \cdots < T_k, N\ (K \leq N)$ örnek için gözlemlenen farklı sıralı olay zamanlarının bir kümesi olduğunu varsayalım. Bu olay zamanlarına ek olarak, olay zamanları gözlemlenmeyen sansürlü zamanlarda vardır. Belirli bir olay zamanı $T_j (j = 1, 2, \ldots, K)$ için, gözlemlenen olay sayısı $d_j \geq 1$ 'dir ve r_j örnek olayları, olay zamanları veya sansürlenmiş zamanları T_j 'den büyük veya eşit olduğu için "risk altında" kabul edilecektir. Sansürleme nedeniyle r_j 'yi basitçe $r_j - 1$ ve $d_j - 1$ arasındaki fark olarak değerlendiremeyiz. r_j 'yi elde etmenin doğru yolu

$r_j = r_j - 1 - d_j - 1 - c_j - 1'$dir; burada $c_j - 1, T_j - 1$ ile T_j arasındaki zaman diliminde sansürlenen örneklerin sayısını temsil eder. T_j zamanın ilerisinde sağkalımın koşullu olasılığı artık şu şekilde tanımlanabilir:

$$p(T_j) = \frac{r_j - d_j}{r_j} \tag{3.1}$$

Bu koşullu olasılığa dayanarak, $S(t) = P(T > t)$ sağkalım fonksiyonun çarpım-limit tahmincisi aşağıdaki gibidir:

$$S(t) = \prod_{j:T_j<t} p(T_j) = \prod_{j:T_j<t} d(1 - \frac{d_j}{r_j}) \tag{3.2}$$

d_j: T_j anındaki ölüm sayısı

c_j: T_j anındaki sansürlü gözlem sayısı

r_j: Risk altında kabul edilen örnek olayları

Kaplan-Meier formülüne çarpım limit formülü denir. Çünkü hayatta kalma olasılıklarının çarpımı ile belirli bir sağkalım süresi ile sınırlıdır (Kleinbaum ve Klein, 2012).

Ancak, veri setindeki gözlemler belirli bir zaman aralıklarına göre gruplandırılmışsa, gözlem sayısı çok fazlaysa Yaşam Tablosu (YT) analizi daha uygun bir yöntem olacaktır. KM ve YT yöntemlerinin aksine Nelson-Aalen tahmincisi sansürlü veriler için kümülatif tehlike fonksiyonunu bir sayma süreci yaklaşımına dayalı olarak tahmin eder (Wang vd., 2019).

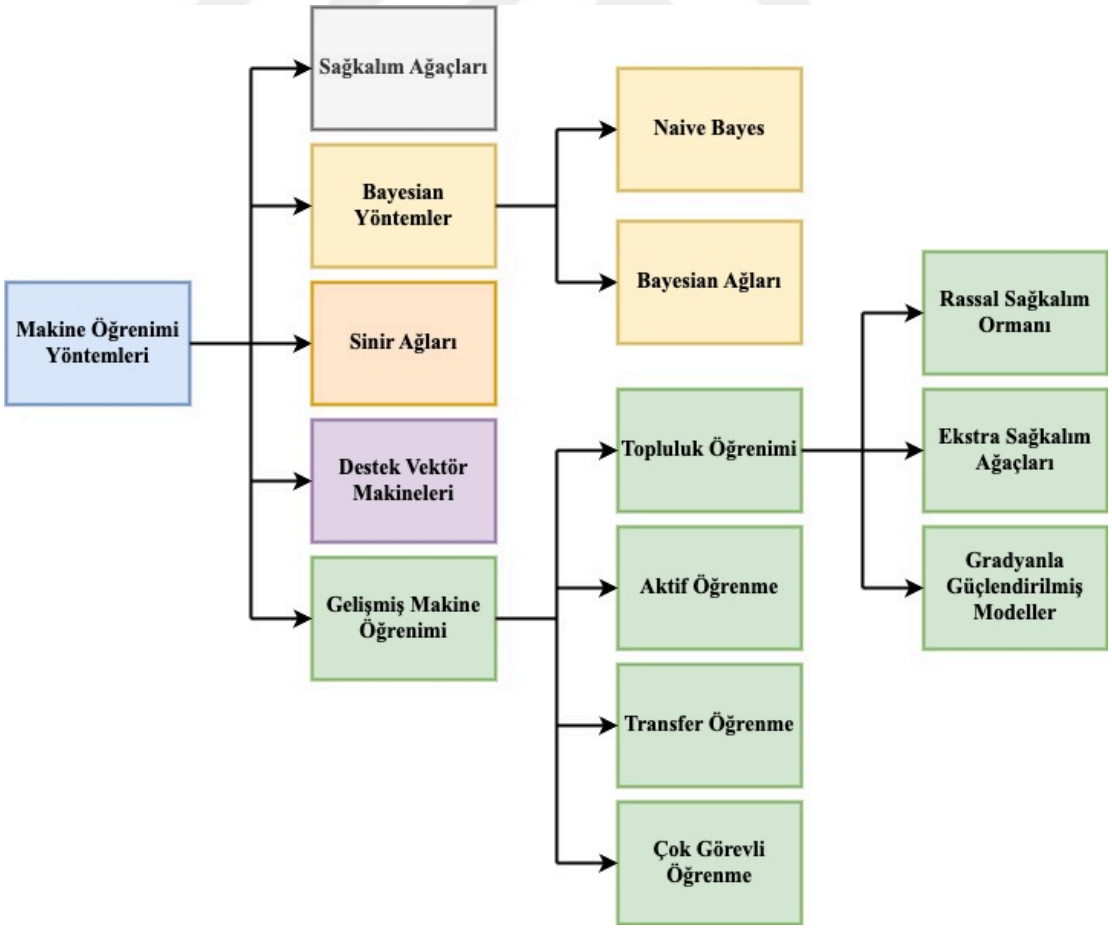
3.2.2 Yarı Parametrik Yöntemler

Parametrik ve parametrik olmayan yöntemlerin bir karışım olan yarı parametrik yöntemler, parametrik yöntemlere göre daha geniş koşullar atlında daha tutarlı sonuçlar verirken, parametrik olmayan yöntemlere göre daha kesin tahminler elde edebilir (Powell, 1994). Cox regresyon analizi bu kategori altında en çok kullanılan sağkalım analizi yöntemidir. Parametrik yöntemlerin aksine ilgilenilen olaya kadar geçen sürenin temel dağılımı hakkında hiçbir bilgiye gerek yoktur, ancak özniteliklerin sonuca üstel bir etkisi olduğu varsayılır (Wang vd., 2019).

BÖLÜM DÖRT
MAKİNE ÖĞRENMESİ YÖNTEMLERİ

Geçtiğimiz birkaç yılda, doğrusal olmayan ilişkilerin modellenmesi ve genel tahminlerde gerçekleşen gelişmelerin, makine öğrenmesi yöntemlerine kazandırdığı birtakım avantajları nedeniyle, birçok çalışma konusunda önemli başarılar elde edilmesini sağladı. Sağkalım analizinde, makine öğreniminin karşılaştığı ana zorluk, sansürlenmiş verilerle ve sağkalım zamanını uygun bir şekilde tahminlemektir.

Makine öğrenimi, uygun boyutlu bir özellik uzayında çok sayıda örnek bulunduğunda efektiftir ancak sağkalım analizinde yer alan problemlerin bazıları için durum böyle değildir (Zupan vd. 2000). Bu bölümde sağkalım analizi için etkili olan makine öğrenmesi yöntemleri incelenecektir. Makine öğrenimi yöntemlerinin taksonomisi Şekil 4.1'de yer almaktadır.



Şekil 4.1 Makine öğrenimi yöntemleri taksonomisi

4.1 Sağkalım Ağaçları

Sağkalım verileri için ağaç yapılarını kullanmaya yönelik ilk girişim 1981 yılında Ciampi vd. (1981) tarafından gerçekleştirilmişse de sağkalım ağaçları ilk olarak Gordon ve Olshen (1985)'in 1985 yılında yayınladıkları bir makalede tartışılmıştır (Gordon ve Olshen 1985).

Sağkalım ağaçları, sınıflandırma ve regresyon ağaçlarının sansürlenmiş verileri işlemesi için özel olarak tasarlanmıştır. Ağaç tabanlı modellerinin basit temeli, veri setinin belirlenen bir bölme kuralına uygun şekilde yinelemeli olarak bölünmesi ve ilgili olaya göre birbirine benzeyen özelliklerin ortak bir düğümde yer almasıdır (Wang vd., 2019).

Sağkalım ağaçları ve standart karar ağaçları arasındaki temel fark, bölme kriterinin seçimidir. Karar ağacı yöntemi, her bir özellik için bir eşik belirleyerek yinelemeli olarak bölümleme gerçekleştirir. Bu bölünme sırasında özellikler arasındaki ilişkileri ve modelde yer alan sansürlü verileri dikkate almaz. (Safavian ve Landgrebe, 1991). Sağkalım ağaçları için bölme kriterleri iki kategoriye ayrılabilir; yani, düğüm içi homojenliği en aza indiren kriterler ve düğümler arası heterojenliği maksimuma çıkaranlar. Bu yaklaşımlardan ilki düğüm içindeki homojenlik kriterini kullanarak kayıp fonksiyonunu en aza indirir. Gordon ve Olshen tahmin edilen dağılım fonksiyonları arasındaki homojenliği ve Hellinger mesafelerini Wassertein metriğini kullanarak ölçmüştür. İkinci tür bölme kriteri için Ciampi ve arkadaşları (1986) düğümler arası heterojenlik için log-rank test istatistiklerini kullanmıştır. Bir sağkalım ağacının standart bir karar ağacına göre sahip olduğu temel avantaj, ağaç yapısını kullanarak sansürlenmiş verileri işleme yeteneğidir (Wang vd., 2019).

Sağkalım ağacı oluşturmanın bir diğer önemli yönü de son ağacın seçimidir. En uygun ağaç yapısının seçilmesi için geriye doğru veya ileriye doğru seçim gibi prosedürler izlenebilir (Bou-Hamad vd., 2011). Bununla birlikte, bir ağaç topluluğu nihai ağaç seçimi sorunun ortadan kaldırabilir ve tek bir ağaç ile mümkün olandan daha iyi bir performans sağlayabilir (Wang vd., 2019).

4.2 Gelişmiş Makine Öğrenimi Yöntemleri

Son birkaç yılda, sansürlenmiş verilerle başa çıkabilmek ve bunlara dayalı tahminler yapmak için bir dizi gelişmiş makine öğrenimi yöntemleri geliştirildi. Bu yöntemler, şimdiye kadar geliştirilen yöntemlere kıyasla sağkalım verileri için benzersiz avantajlar sunar (Wang vd., 2019).

Sağkalım analizi için öne çıkan yöntemler topluluk öğrenimi alanında ortaya çalışmalardır. Topluluk öğrenimi, birbirinden farklı sınıflandırıcılardan oluşan bir grup oluşturur ve bu sınıflandırıcılardan gelen tahminlerin sonuçlarını oylama ya da ortalamasını alarak sınıflandırma yapar. Birden fazla tahmini birleştirerek daha iyi bir tahmin sonucu elde etmeye çalışır. Özellikle yetersiz veri olduğunda topluluk öğrenimi toplulukların başlangıç noktalarını değiştirerek daha iyi topluluklar oluşturarak daha iyi bir yaklaşım elde edilmesini mümkün kılar. Özellikle torbalama (bagging) (Breiman 1996) ve rassal orman (random forest) (Breiman 2001) algoritmaları topluluk öğrenimi dendiğinde öne çıkan iki algoritmadır. Torbalama ve rassal orman topluluk öğrenimi modelleri sağkalım analizine başarılı bir şekilde uyarlanmıştır.

4.2.1 Rassal Orman (Random Forest)

Sıkça kullanılan topluluk öğrenimi yöntemlerinden biri olan rassal orman algoritması, ağaç tabanlı modelleri kullanarak tahminler elde etmek için geliştirilmiştir. Karmaşık bir sorunu çözmek ya da modelin performansını artırmak için kullanılan rassal orman yöntemi birden fazla ağacın birleşmesinden oluşmaktadır. Birden fazla ağaç kullanılması modelin tahmin gücünü artırırken karar ağaçlarında yaşanan aşırı öğrenme sorununda önüne geçiyor. Bir diğer topluluk öğrenimi olan torbalama yöntemiyle benzer olan rassal ormanın temel farkı değişkenlerin seçiminde rastgele bir alt küme kullanmasıdır. Torbalama yönteminde değişkenlerin hepsi kullanılırken rassal ormanlarda rastgele bir alt küme kullanılmaktadır. Rassal orman yönteminde her ağaç için rastgele seçilen değişkenler ağaçlar arasındaki ilişkiyi azaltır ve modelin tahmin performansını geliştirir.

4.2.2 Rassal Sağkalım Ormanı (Random Survival Forest)

Rassal Sağkalım Ormanı (Ishwaran, 2008), tahminleme yapmak için sağkalım ağaçlarını kullanarak Breiman'ın rassal orman yöntemini genişletti ve geliştirdi (Wang vd. 2019). Rassal sağkalım ormanı, regresyon analizi ve sınıflandırma yöntemlerine daha çok odaklanmış olan rassal orman yönteminin sansürlü verilere uygun hale getirilmiş bir genişletmesidir. Tehlike fonksiyonlarının orantısal kısıtlamasından kaçınmak için kullanılan bu yöntem kümülatif tehlike fonksiyonu için bir topluluk tahmini oluşturur. Elde edilen bu ağaçlar performans tahminlerini önemli ölçüde artırmaktadır. Rassal sağkalım ormanında bölme kriteri log-rank testi kullanılır.

Rassal sağkalım ormanı algoritmasının uygulanması sırasında aşağıdaki adımlar takip edilir. Algoritmaya ait akış şeması Şekil 4.2'de yer almaktadır.

Adım 1. Veri setinden bootstrap yöntemi ile rastgele örneklemler seçilir. Seçilen örneklemlerin %37'si OOB (out-of-bag) verisi olarak ayrılır.

Adım 2. Oluşturan örneklemlerin her biri için, rastgele seçilmiş özelliklerle bir sağkalım ağacı oluşturulur ve çocuk düğümler arasındaki fark maksimize edilene kadar aday değişkenlerle bölünme devam eder.

Adım 3. Bölünme önceden belirlenen benzersiz ölüm sayısından fazla ya da kısıta ulaşana kadar devam eder. Daha sonra elde edilen ağaçların kümülatif tehlike fonksiyonlar hesaplanır ve bunların ortalaması elde edilir.

Adım 4. Son adımda ise OOB (out-ouf-bag) verisi kullanılarak kümülatif tahmin hatası hesaplanır.



Şekil 4.2 Rassal sağkalım ormanı algoritması akış şeması

4.2.3 Ekstra Sağkalım Ağaçları (Extremly Randomized (Extra) Survival Trees)

Ekstra sağkalım ağaçları algoritması topluluk öğrenimi yöntemi olan ekstra ağaçlar (extra trees) algoritmasına dayanır. Rassal sağkalım orman algoritmasına benzeyen Ekstra sağkalım ağaçları algoritmasının en temel farkı rassal sağkalım ormanı her bir ağaç için veri setinden oluşturulan alt örneklemleri kullanırken, ekstra sağkalım ağaçları yöntemi tüm veri setini kullanır. Ekstra sağkalım ağaçlarında oluşturulan karar ağaçlarında bölünme noktaları ve her bir düğümde kullanılacak özellik rastgele bir şekilde seçilir. Her düğümde en iyi bölünme noktasını seçmek yerine rastgele aday bölünme noktaları seçilir ve bunlar arasından en iyisi seçilir. Ekstra sağkalım ağaçları yönteminde her yaprak düğümde, sağkalım fonksiyonunu tahmin edilir. Sağkalım fonksiyonun tahminlemesinde genellikle Kaplan-Meier tahmincisi kullanılır.

Ekstra sağkalım ağaçları yöntemi rastgele bölünme noktaları seçtiği için modeli daha hızlı eğitebilir ve rastgelelik ile sağlanan ağaçlar arasındaki çeşitlilik aşırı öğrenme riskini de azaltmaktadır (Geurts vd., 2006).

4.2.4 Gradyanla Güçlendirilmiş Modeller (Gradient Boosted Models)

Grandyanla güçlendirilmiş modeller ya da gradient boosted models (GBM) makine öğreniminde kullanılan topluluk öğrenimi yöntemlerinden biridir. Bu yöntemin amacı zayıf öğreniciler bir araya getirerek güçlü bir öğrenici oluşturmaktır. GGM yöntemi sağkalım analizinin en önemli iki parçası olayların meydana gelme süreleri ve sansürlü verileri göz önünde bulundurarak sağkalım fonksiyonlarını tahmin etmeye çalışır. Sağkalım analizi için uyarlanan bu yöntem güçlü ve esnek bir yaklaşım sunmaktadır. Özellikle karmaşık veri setleri üzerinde etkili olan yöntemlerden biridir. GGM, tek bir model yerine birden fazla kayıp fonksiyonunu optimize etmek için kullanılan esnek bir yapıya sahiptir. Temel öğreniciler aynı zamanda zayıf öğreniciler adlandırılmaktadır. Bunun sebebi ise bu temel öğrencilerin rastgele tahminlerden biraz daha iyi sonuçlar elde eden basit modeller olmasıdır. RSO yöntemine benzerlik gösteren GGM yönteminin temel farkı oluşturduğu bu güçlendirilmiş modeldir. RSO yöntemi birbirinden bağımsız ağaçların tahminlerin ortalamasını kullanırken, GGM yönteminde güçlendirilmiş model açgözlü (greedy) bir şekilde sıralı olarak inşa edilir.

GGM yöntemi 3 temel bileşenden oluşmaktadır. Bu bileşenler:

Zayıf Öğreniciler: Genellikle karar ağaçlarından oluşan bu zayıf öğreniciler yinelemeli olarak eklenir ve bir önceki modelin hatalarını düzeltmeye çalışır.

Kayıp (Loss) Fonksiyonu: Kayıp fonkyonu olayların meydana gelme süreleri ve sansürlü verilerini dikkate alarak modellerin gerçek sağkalım süreliyle ne kadar eşleştiğini hesaplar. Genellikle Cox orantısal tehlike (Cox propotional hazards) modelinin kısmi olasılık (Cox's partial likelihood) fonksiyonu kullanılmaktadır. Amaç log-kısmi olabilirlik fonksiyonunu maksimize etmektir fakat geleneksel lineer model, katkısal model $f(x)$ ile değiştirilir.

$$\underset{f}{\operatorname{argmin}} \sum_{i=1}^{n} \delta_i \left[f(x_i) - \log\left(\sum_{j \in R_i} \exp\left(f(x_j) \right) \right) \right] \tag{4.1}$$

Gradyan Azaltma (Gradient Descent): Her iterasyonda, kayıp fonksiyonun negatif gradyanları hesaplanır ve bu hesaplanan gradyanlar modelin hatalarını düzeltmek için kullanılır. Oluşturulan yeni zayıf öğrenciler bu gradyanları öğrenmek için eğitilir.

Tahminleri eklemeli bir şekilde birleştirerek genel modeli iyileştiren model şu şekildedir:

$$f(x) = \sum_{m=1}^{M} \beta_m g(x; \theta_m), \tag{4.2}$$

$M > 0$, temel öğrencilerin sayısını ifade ederken, $\beta \in R$ ağırlık terimidir. g fonksiyonu, bir temel öğreniciyi tanımlar ve θ vektörü ile parametreleştirilir. Bağımsız temel öğrenciler parametrelerinin yapılandırılmasındaki farklılıklar nedeniyle birbirinden farklıdır, bu farklılar θm ile belirtilmektedir.

Şekil 4.3'te akış şeması bulunan GBM algoritmasının adımları şu şekildir:

Adım 1. Başlangıç modeli $h_0(t)$ olarak belirlenir. Bu model, veri setinin ortalama sağkalım süresini ya da başlangıç tehlike fonksiyonunu tahmin edebilir.

Adım 2. Her bir iterasyonda kısmı olabilirlik fonksiyonun negatif gradyanları hesaplanır. Hesaplanan bu gradyanlar modelin hatalarını temsil etmektedir.

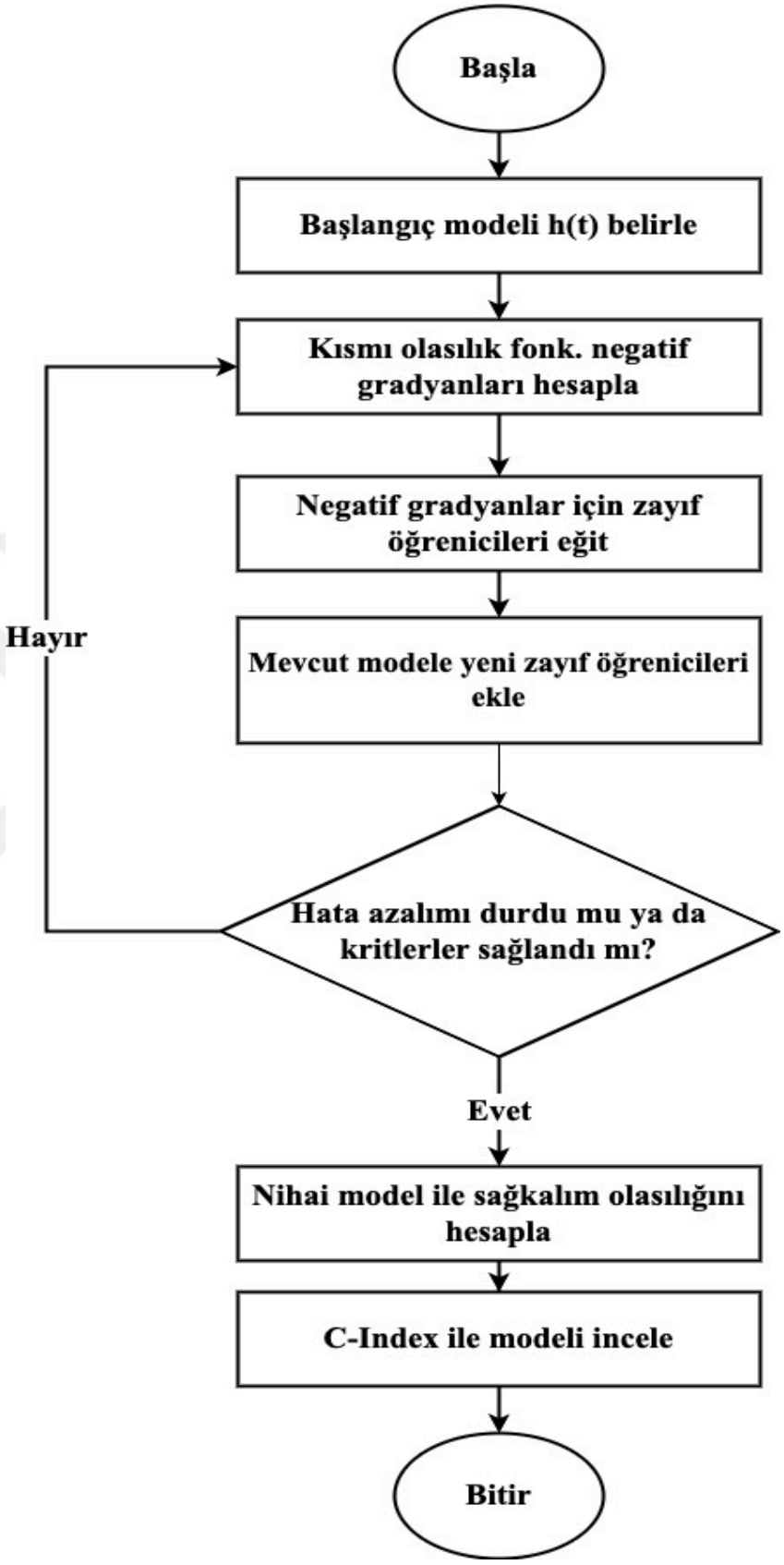
Adım 3. Hesaplanan negatif gradyanları öğrenmek için yeni zayıf öğreniciler eğitilir.

Adım 4. Mevcut modelin tahminlerine yeni zayıf öğrenici eklenir (Formülde yer alan γ öğrenme katsayıdır.):

$$h_m(t) = h_{m-1}(t) + \gamma h_m(x) \tag{4.3}$$

Adım 5. Modelin hata azaltımı durana kadar ya da iterasyon sayısı belirtilen koşula kadar 2-4 adımları tekrarlanır.

Adım 6. Belirlenen koşullara ulaşıldıktan sonra nihai model tüm zayıf öğrencilerin eklenmesiyle oluşur ve bu model sağkalım fonksiyonun tahmin edilmesinde kullanılır.



Şekil 4.3 GGM yöntemi akış şeması

BÖLÜM BEŞ
ANALİZ

Tezin bu bölümünde birbirinden farklı seçilen 3 veri seti makine öğrenimi yöntemleri olan rassal sağkalım ormanı, ekstra sağkalım ormanı ve gradyanla güçlendirilmiş yöntemler ile çalıştırılmıştır. Kullanılan bu üç veri seti birbirinden farklı boyutlara ve farklı değişken sayıda değişkene sahiptir. Bu üç veri seti %75 eğitim verisi ve %25 test verisi olarak kullanılmak üzere bölünmüştür. Her bir algoritma için eğitim verisi üzerinden modeller eğitilmiştir. 100 defa çalıştırılan algoritmaların eğitim verileri ve test verilerinin ortalama C-indeks değerleri, sağkalım olasılıkları, algoritmaların çalışma süreleri elde edilmiş olup yorumlanmış ve karşılaştırılmıştır.

Uygulamalar Python yazılım dilinde geliştirilmiş olup, yöntemler için scikit-survival kütüphanesinden yararlanılmıştır. Uygulamaların testleri macOs 14.6.1 sürümünde, Apple M1 Pro işlemci ve 16 GB RAM hafızasına sahip olan bir bilgisayar ile gerçekleştirilmiştir. Scikit-survival kütüphanesi, scikit-learn üzerine inşa edilmiş bir kütüphanedir. Bu kütüphane Python yazılım diliyle birlikte sağkalım analizi çalışmalarında kullanılmasına olanak tanır. Bu çalışmada kullanılan üç algoritmanın testleri scikit-survival kütüphanesinde yer alan modüller aracılığıyla sağlanmıştır. Modellerin eğitilmesinde kullanılan parametreler şu şekilde seçilmiştir:

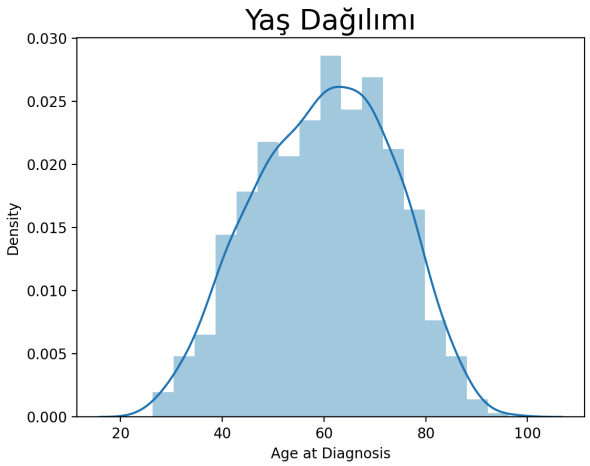
- **n_estimators (ağaç sayısı):** 250
- **min_samples_split (bölünmek için gereken minimum örneklem):** 3
- **min_samples_leaf (yaprakta bulunması gereken min. örnek sayısı):** 5

Bu parametreler nihai sonuçların elde edilmesinden önce yapılan ön çalışmalar sırasında elde edilmiştir. Çeşitli parametre denemelerinin sonunda sonuçların değişmediği ya da kötü sonuçların elde edildiği parametreler elenmiş ve optimum sonuçların elde edildiği parametre değerleri seçilmiştir.

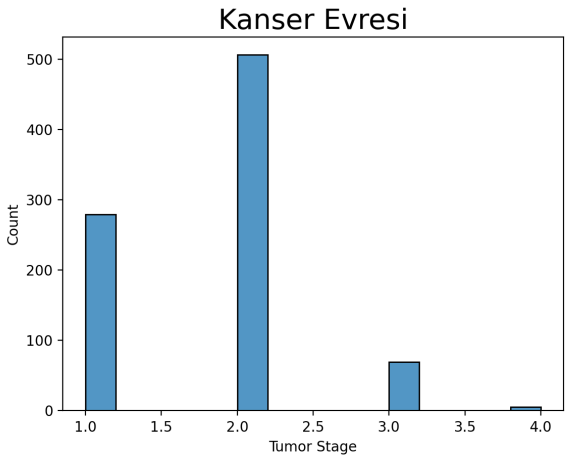
5.1 Uygulama 1 – Meme Kanseri Hastaları Sağkalım Analizi

Birinci uygulamada kullanılan veri setinde meme kanserinin en yaygın türü olan duktal invaziv meme kanseri hastası olan 859 kadın hastaya ait veriler

bulunmaktadır.859 kayıtın yer aldığı bu veri setinde 27 farklı değişken yer almaktadır. Veri setinde bulunan değişkenler ve bu değişkenlere ait bilgiler Tablo 5.1'de yer almaktadır. Veri setinde bulunan hastaların yaş dağılımları ise Şekil 5.1'deki gibidir. Buna göre hastaların yaşlarının 50-70 yaşları arasında yüksek olduğu görülmektedir. Hastaların kanser evrelerinin dağılımı ise Şekil 5.2'de verilmiştir. Bu bilgiler ışığında yaklaşık 500 hastanın 2. evre kanser hastası olduğu görülmektedir. Bu veri seti her 3 algoritmada test edilmiştir. Bu testler sonucunda eğitim verileri ve test verileri için c-indeks değerliyle birlikte sağkalım olasılıkları hesaplanmıştır.



Şekil 5.1 Uygulama 1'de kullanılan veri setine ait kayıtların yaş dağılımı



Şekil 5.2 Uygulama 1'de kullanılan veri setine ait kanser evrelerinin dağılımı

Tablo 5.1 Uygulama 1'de kullanılan veri setine ait değişkenlerin bilgi tablosu

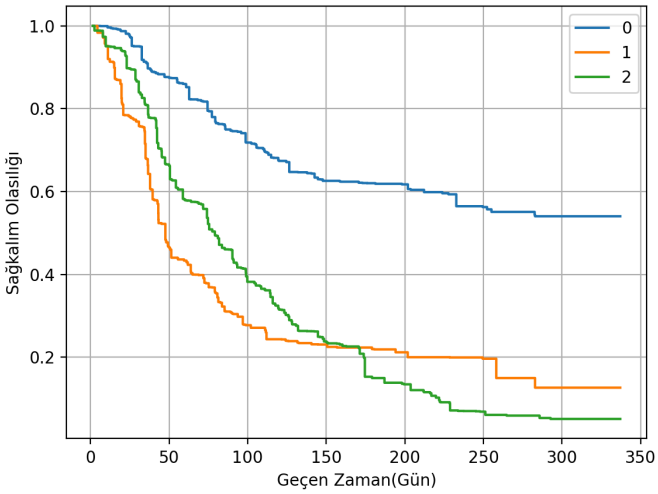
Değişkenler	**Açıklama**	**Değişken Tipi**	**Değerler**
Age at diagnosis	Hastanın ilk kez tanı aldığındaki yaşı	Sürekli	-
Type of breast surgery	Meme ameliyatı türü	Kategorik	Masektomi, meme koruma
Cellularity	Tümör yatağındaki kanser oranı	Kategorik	Yüksek, orta derece
Chemothreapy	Hasta kemoterapi aldı mı?	Kategorik	Evet, hayır
Pam50 + claudin-low subtype	Claudin alt türü	Kategorik	LumA, LumB, Normal, Basal, Claudin-Low, Her2
Chorot	Çalışma grubu	Sürekli	-
Er status measured by ihc	Östrojen reseptörü durumu	Kategorik	Pozitif, negatif
Er status	Östrojen reseptörü durumu	Kategorik	Pozitif, negatif
Neoplasm histologic grade	Kanser hücresi çekirdeğinin normal hücreye benzerliği	Sürekli	-
Her2 status measured by snp6	Epidermal büyüme faktörü	Kategorik	Doğal, kazanç, kayıp
Her2 status	Epidermal büyüme faktörü	Kategorik	Pozitif, negatif
Tumor other histologic subtype	Diğer histolojik alt tipler	Kategorik	Ductal/NST, Lobular, Medullary, Mucinous, Mixed
Hormone therapy	Horman tedavisi alıyor mu?	Kategorik	Evet, hayır
Inferred menopausal state	Menapoz durumu	Kategorik	Önce, sonra
Integrative cluster	Kanser alt küme tipleri	Kategorik	1-10
Primary tumor laterality	Birincil tümör lateralitesi	Kategorik	Sol, sağ
Lymph nodes examined positive	Kanser içeren lenf düğümü sayısı	Sürekli	-
Mutation count	Mutasyon sayısı	Sürekli	-
Nottingham prognostic index	Ameliyat edilebilir kanserin sağkalımını öngören prognostik ölçüt	Sürekli	-
Overall survival (months)	Sağkalım süresi	Sürekli	-
Overall survival status	Sağkalım durumu	Kategorik	Hayatta, ölmüş
Pr status	Progestreon reseptörü durumu	Kategorik	Pozitif, Negatif
Radio therapy	Radyoterapi	Kategorik	Evet, hayır
Sex	Cinsiyet	Kategorik	Kadın
3-gene classifier subtype	3 genli sınıflandırı alt tür	Kategorik	ER+/HER-2 düşük, ER+/HER2 yüksek, ER-/HER2-, HER2+
Tumor size	Kanser hücresi boyutu	Sürekli	-
Tumor stage	Kanser evresi	Sürekli	-

5.1.1 Sağkalım Olasılıkları ve Tehlike Oranları

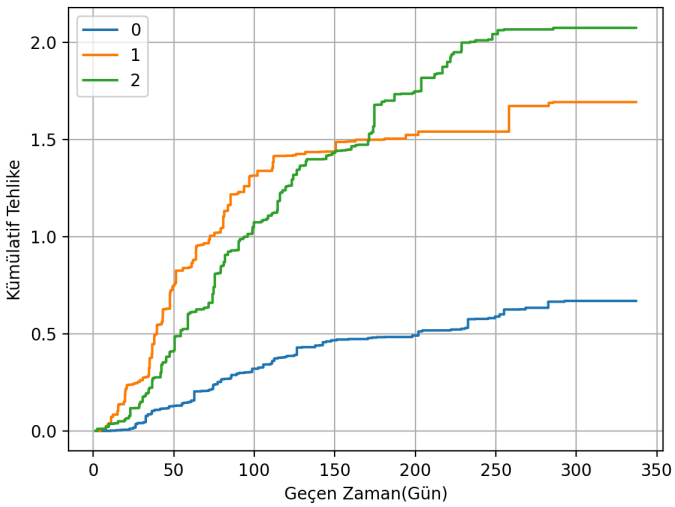
Yapılan uygulamada sonucunda elde edilen sağkalım olasılıkları ve tehlike oranları seçilen 3 hasta üzerinde incelenmiştir. Bu incelemelerin yapılma amacı yaş, tümör boyutu, kanserin evresi gibi değişkenlerin sağkalım olasılığının üzerindeki etkisini de göstermektir. Seçilen hastalardan birinci hastanın yaşı 38, tümör büyüklüğü 19 mm ve kanserin evresi 1. evredir. İkinci hastanın yaşı 56, tümör büyüklüğü 99 mm ve 3. kanser evresinde olan bir hastadır. Seçilen üçüncü hasta ise 90 yaşında 20 mm büyüklüğünde tümörü olan 2. evre kanser hastasıdır. RSO algoritması için hastaların sağkalım olasılıkları Şekil 5.3'te görülürken, kümülatif tehlike oranları ise Şekil 5.4'te yer almaktadır. Şekil 5.5'te ESA algoritmasının sağkalım olasılıkları yer alırken Şekil 5.6'da kümülatif tehlike oranları yer almaktadır. Şekil 5.7'de ve Şekil 5.8'de ise GGM yönteminin üç hasta için sağkalım olasılıkları ve tehlike oranları elde edilmiştir. Bu grafiklerde yer alan çizgilerden mavi olan 38 yaşındaki birinci hasta, turuncu ile gösterilen hasta 56 yaşındaki ikinci hasta ve son olarak grafikte yeşil olarak gösterilen hasta 90 yaşındaki üçüncü hastadır.

RSO ve ESA algoritmalarının sağkalım olasılıkları ve kümülatif oranları her üç hasta içinde hemen hemen aynıdır. Her üç hasta içinde yaklaşık 300 gün sonunda sağkalım olasılıklarında değişiklik gözlemlenmemiştir. 1. Evre ve yaşça daha genç olan birinci hastanın sağkalım olasılığı 0.6'dır. 3. Evre kanser hastası olan orta yaşlı ikinci hastanın ve 2. evre kanser hastası olan 90 yaşındaki üçüncü hastanın sağkalım oranları birbirlerine oldukça yakındır. Bu durum bize yaş ve kanser evresinin sağkalım olasılığı üzerindeki etkisi hakkında bilgi vermektedir.

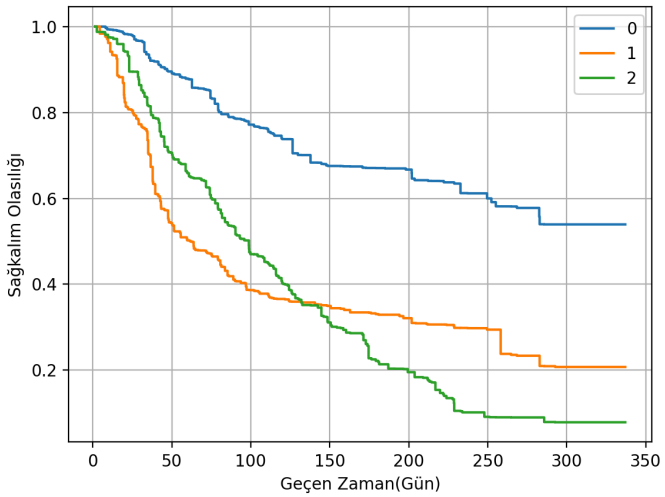
GGM yönteminde ise diğer iki yönteme göre sağkalım oranları çok düşüktür. Şekil 5.7'de görüldüğü üzere yaklaşık 300 gün sonunda ilk hastanın sağkalım olasılığı çok düşük iken diğer iki hastanın sağkalım olasılığı sıfırdır yani bu iki hasta ölüm olayını yaşamaktadır.



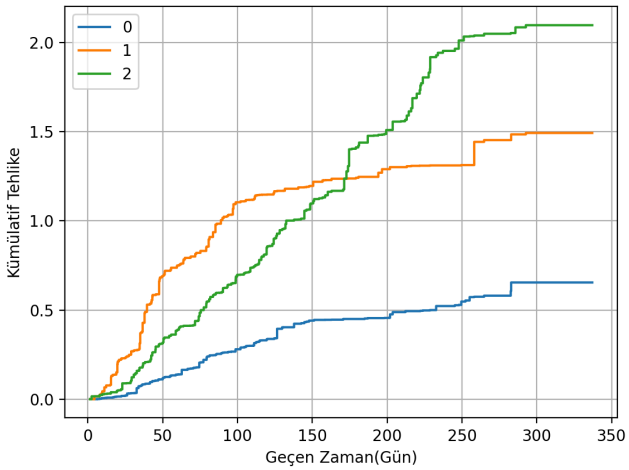
Şekil 5.3 Uygulama 1 RSO için hastaların sağkalım olasılıkları grafiği



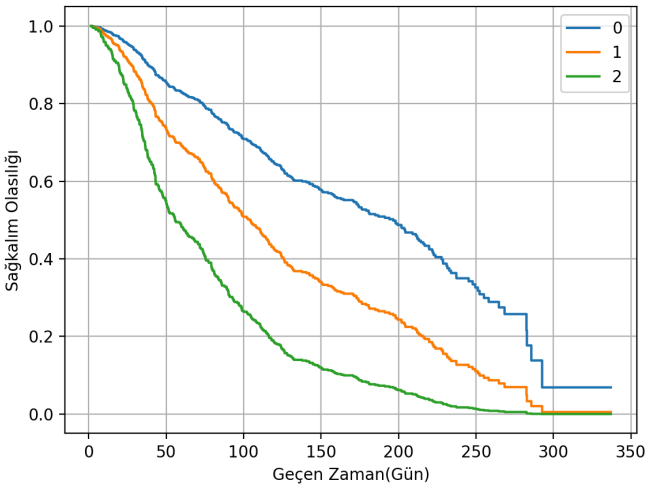
Şekil 5.4 Uygulama 1 RSO için hastaların kümülatif tehlike oranları grafiği



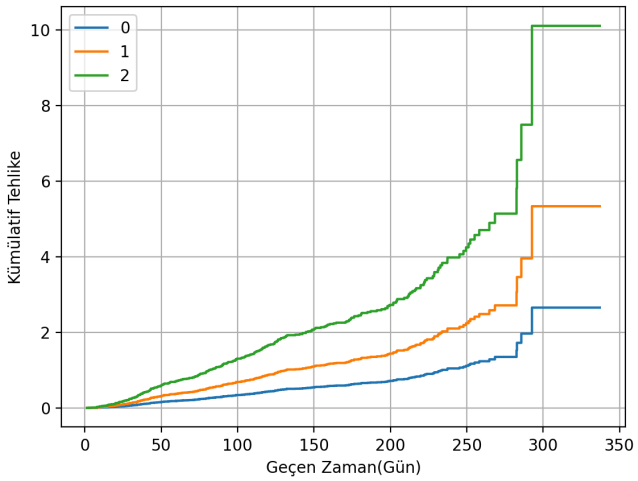
Şekil 5.5 Uygulama 1 ESA için hastaların sağkalım olasılıkları grafiği



Şekil 5.6 Uygulama 1 ESA için hastaların tehlike oranları grafiği



Şekil 5.7 Uygulama 1 GGM için hastaların sağkalım olasılıkları grafiği



Şekil 5.8 Uygulama 1 GGM için hastaların kümülatif tehlike oranları grafiği

5.1.2 C-İndeks Sonuçları ve Çalışma Süreleri

Yöntemlerin C-indeks değerleri Tablo 5.2'de yer almaktadır. RSO ve ESA yöntemleri birbirlerine yakın sonuçlar elde ederken GGM yöntemi diğer iki yönteme göre daha düşük C-indeks değerine sahiptir. Eğitim veri setlerinde en yüksek C-indeks değeri 0.8282 ile RSO yöntemine ait iken en düşük C-indeks değeri 0.6851 ile GGM yöntemine aittir. Test veri setlerinde ise en yüksek C-indeks değeri eğitim veri setinde

olduğu gibi 0.6821 ile RSO yöntemine ait iken en düşük C-indeks değeri ise 0.6268 ile GGM yöntemine aittir.

Kullanılan üç yöntemin çalışma süreleri sırasıyla:

- RSO yöntemi çalışma süresi 50,33 saniye
- ESA yöntemi çalışma süresi 42,21 saniye
- GGM yöntemi çalışma süresi 171,81 saniye

ESA yöntemi en hızlı çalışan yöntem iken GGM yöntemi diğer iki yönteme oranla 4 kat daha yavaş çalışmaktadır.

Tablo 5.2 Uygulama 1 için yöntemlerin C-indeks değerleri tablosu

	Eğitim Veri Seti			**Test Veri Seti**		
Yöntem	**Minimum**	**Maksimum**	**Ortalama**	**Minimum**	**Maksimum**	**Ortalama**
RSO	0.8215	0.8282	0.8243	0.6694	0.6821	0.6760
ESA	0.7796	0.7876	0.7831	0.6639	0.6778	0.6699
GGM	0.6851	0.6903	0.6871	0.6268	0.6343	0.6299

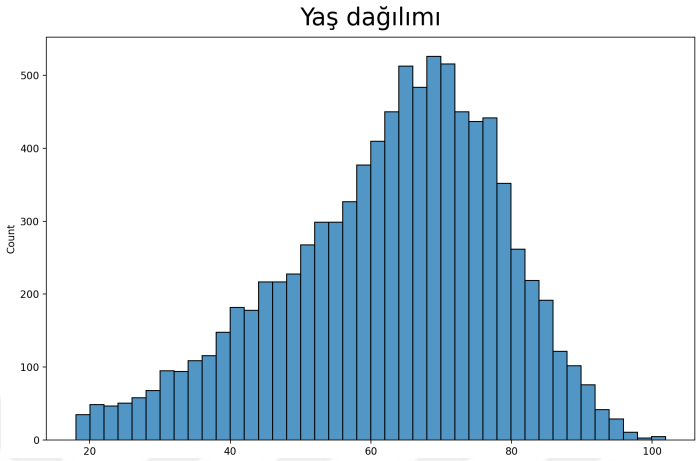
5.2 Uygulama 2 – Hastanede Tedavi Gören Hastalar İçin Sağkalım Analizi

İkinci uygulamada kullanılan veri seti SurvSet paketi içinde yer alan support2 veri setidir. Bu veri setinde ciddi bir hastalık nedeniyle hastaneye kaldırılmış 9037 hastaya ait veriler içermektedir. Veri seti 28 farklı değişkenden oluşmaktadır.

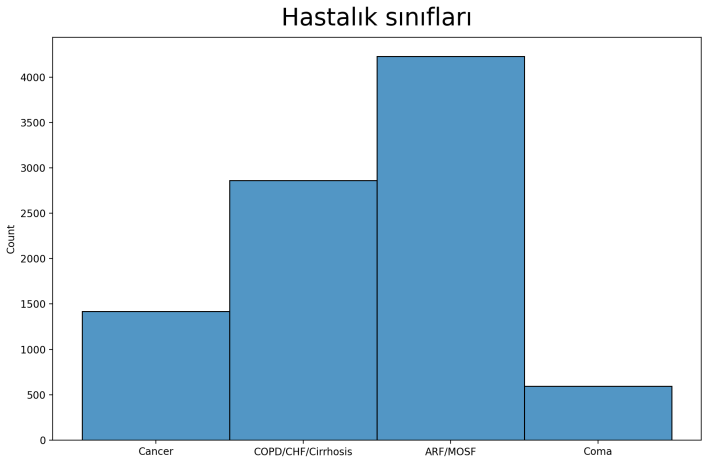
Kanser hastalarının yanı sıra siroz, koma, organ yetmezliği gibi farklı hasta bilgilerinin de yer aldığı bu veri seti sağkalım analizi için oldukça kullanışlı bir veri setidir. Özellikle hastaların hastanede kaldığı gün süreleri ve kan değerleri ve sahip oldukları hastalıklarının yanı sıra sahip olduğu başka hastalık türleri ve hastalık sayıları da bu veri setini sağkalım analizi için kullanışlı hale getiren nedenlerdir. Veri setinde yer alan hastaların büyük bir çoğunluğu 60-80 yaşları arasında yer almaktadır.

Veri seti hakkında verilen bazı bilgilere ait histogramlar elde edilmiş. Hastalara ait yaş dağılımları Şekil 5.9'te görülmektedir. Şekilde 5.9'da da görüldüğü üzere hastaların 60-80 yaş aralığında yer alan hasta sayısı bir hayli fazladır. Ayrıca bu hastaların hastalık sınıflarına ait dağılımı da Şekil 5.10'te görülmektedir. Şekil 5.10'da

görüldüğü üzere böbrek/organ yetmezliğine sahip hastaların daha fazla olduğu görülmektedir.



Şekil 5.9 Uygulama 2'de kullanılan veri setinde yer alan hastaların yaş dağılımları



Şekil 5.10 Uygulama 2'de kullanılan veri setinde yer alan hastaların hastalık sınıfları dağılımları

Veri setine ait değişkenler, değişkenler hakkında açıklamalar, değişken tipleri ve değişkenlerin sahip olduğu değerler Tablo 5.3'te görülmektedir.

Tablo 5.3 Uygulama 2'de kullanılan veri setine ait değişkenlerin bilgi tablosu

Değişkenler	Açıklama	Değişken Tipi	Değerler
Event	Sağkalım durumu	Sürekli	-
Time	Sağkalım süresi	Sürekli	-
Age	Yaş	Sürekli	-
Num_co	Eşlik eden hastalık sayısı	Sürekli	-
Scoma	Glasgow koma skoru	Sürekli	-
Hday	Hastanedeki gün sayısı	Sürekli	-
SPS	Fizyoloji tabanlı puan sistemi	Sürekli	-
Meanbp	Kan basıncı	Sürekli	-
Wblc	Beyaz kan hücresi	Sürekli	-
Hrt	Kalp atış hızı	Sürekli	-
Resp	Solunum hızı	Sürekli	-
Temp	Vücut sıcaklığı	Sürekli	-
Pafi	PaO2/FiO2 oranı,	Sürekli	-
Alb	Albümin değeri	Sürekli	-
Bili	Bilirubin değeri	Sürekli	-
Crea	Kreatinin değeri	Sürekli	-
Sod	Sodyum değeri	Sürekli	-
Ph	Arteriyal kan pH değeri	Sürekli	-
Glucose	Glikoz değeri	Sürekli	-
Bun	Üre değeri	Sürekli	-
Urine	İdrar çıkışı	Sürekli	-
Sex	Cinsiyet	Kategorik	Kadın, Erkek
Dzgroup	Hastalık grubu	Kategorik	
Dzclass	Hastalık sınıfı	Kategorik	
Diabetes	Şeker hastası	Kategorik	Evet, hayır
Dementia	Demans hastası	Kategorik	Evet, hayır
Ca	Hasta kanser hastası mı?	Kategorik	Evet, Hayır, metastatik
Dnr	Öldükten sonra diriltildi	Kategorik	Dnr yapılmadı, Dnr yapıldı

- Dzgroup'da yer alan gruplandırılmış hastalıklar ARF/MOSF w/Sepsis (Akut böbrek yetmezliği/ Çoklu organ yetmezliği kan zehirlenmesi ile birlikte), COPD (KOAH), CHF (kalp yetmezliği), siroz, koma, kolon kanseri, akciğer kanseri ve MOSF w/Malig (Kötü huylu tümör ile organ yetmezliği)'dir.
- Dzclass'da yer alan sınıflandırılmış hastalıklar ARF/MOSF, COPD/CHF/Siroz, koma ve kanserdir.

5.2.1 Sağkalım Olasılıkları ve Tehlike Oranları

Uygulamada eğitilen modellerin sağkalım oranları ve tehlike oranları seçilen 5 hasta üzerinde incelenmiştir. Genç, orta yaş ve yaşlı yaşlarda yer alan bu hastaların hastaneye yatış nedenleri ve sahip oldukları hastalıklarda birbirlerinden farklıdır. Bu durumlar göz önüne alarak veriler incelenmiş ve yorumlanmıştır.

Grafikte mavi renkte gösterilen birinci hasta 18 yaşında organ yetmezliği olan ve herhangi bir hastalığı bulunmayan bir hastadır. İkinci hasta 28 yaşında böbrek yetmezliği olan ve bunun dışında farklı bir hastalığı da olan bir hastadır. İkinci hasta grafikte turuncu olarak görülmektedir. Üçüncü yani grafikte yeşil olarak görülen hasta 52 yaşında komada olan bir hastadır. Bu hastanın diyabeti ve 2 farklı hastalığı da bulunmaktadır. Grafikte kırmızı renk ile gösterilen 70 yaşındaki akciğer kanseri hastasının kanserin yanı sıra 2 farklı hastalığı da bulunmaktadır. Mor renk ile gösterilen son hasta ise 102 yaşında kalp yetmezliği ve 4 farklı hastalığı olan bir hastadır. Bu hastalardan dördüncü hastanın kanseri metastaz yapmış ve farklı organlara da sıçramıştır. İlk hastada da kanser hücresi bulunurken diğer hastalarda kesinlikle kanser görülmemektedir.

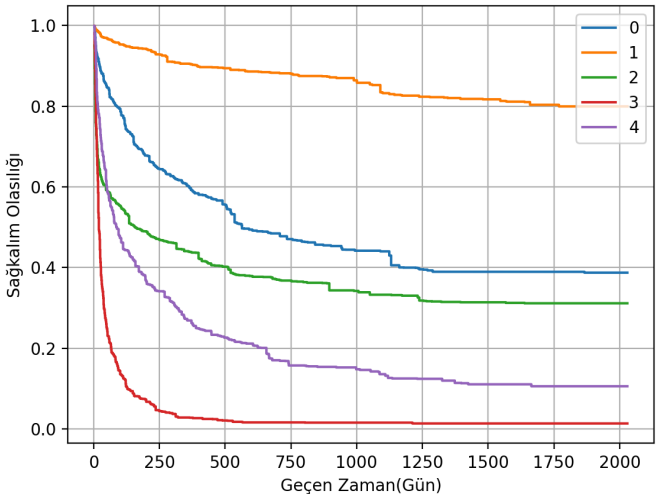
RSO algoritması için sağkalım olasılıkları 5.11'de yer alırken, tehlike oranları grafiği ise Şekil 5.12'de yer almaktadır. ESA algoritmasının sağkalım olasılıkları grafiği Şekil 5.13'te ve tehlike oranları grafiği ise Şekil 5.14'te yer almaktadır. Şekil 5.15'te GGM yönteminin sağkalım olasılıkları grafiği yer alırken şekil 5.16'da ise tehlike oranları grafiği yer almaktadır.

RSO yöntemi için incelediğimizde yaklaşık 500 gün sonrasında sağkalım olasılıklarında değişim olmadığı gözlemlenmektedir. Bu hastalar arasında akciğer

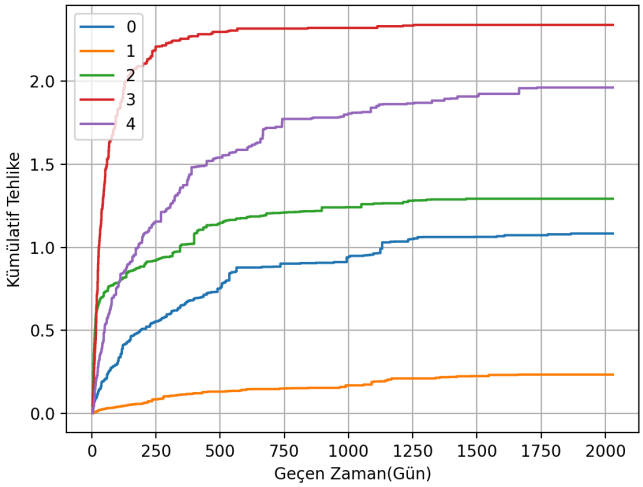
kanseri olan ve kanseri diğer organlara da sıçramış olan hastanın sağkalım oranı en düşük olarak görülmektedir. Böbrek yetmezliği olan ve kanser hastalığı görülmeyen ikinci hastanın sağkalım olasılığı bir hayli yüksektir ve ölüm olayını yaşamamaktadır. Tehlike oranları da her geçen gün giderek artmakta ilk hasta için ilk 250 günde hızla yükselirken daha sonrasında benzer bir oranda ilerlemektedir. Böbrek yetmezliği olan hastanın tehlike oranı çok az artarak belli bir doğrultuda ilerlemektedir. Diğer üç hastanın ise tehlike oranları 750 gün sonrasında yükselişleri yavaşlamaktadır.

ESA algoritması, RSO algoritmasına benzer sonuçlar elde ederken Bazı hastalar için hesaplarında farklılıklar vardır. Böbrek yetmezliği olan hastanın sağkalım oranı bu yöntemde daha düşük iken kalp yetmezliği olan en yaşlı hastanın sağkalım oranı daha yüksektir. Fakat her iki yöntem içinde beklenen ölüm olayı bu hastalar için gerçekleşmemektedir.

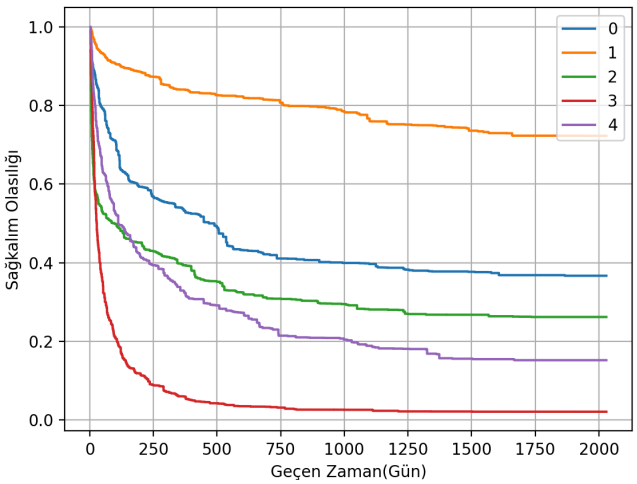
Üçüncü ve son yöntemimiz olan GGM yönteminde ise diğer iki yöntem ile benzer sonuçlar gözlemlenmiştir. Akciğer kanseri olan dördüncü hastanın bu yöntem sonunda sağkalım olasılığı daha düşüktür fakat 2000 günlük süreçte ölüm olayı yaşanmamaktadır.



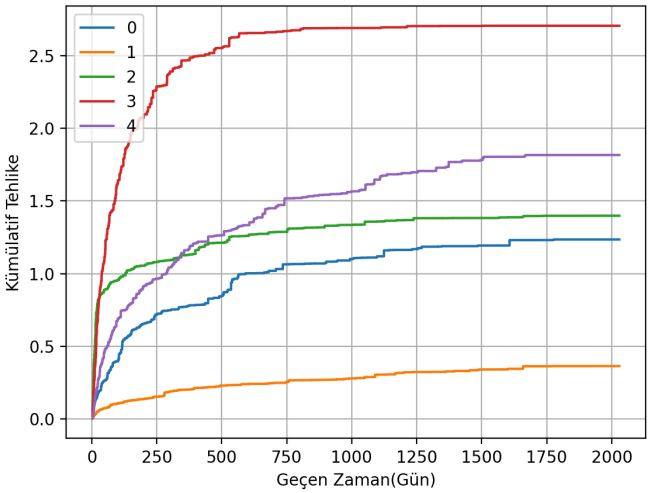
Şekil 5.11 Uygulama 2 RSO yöntemi için hastaların sağkalım olasılıkları grafiği



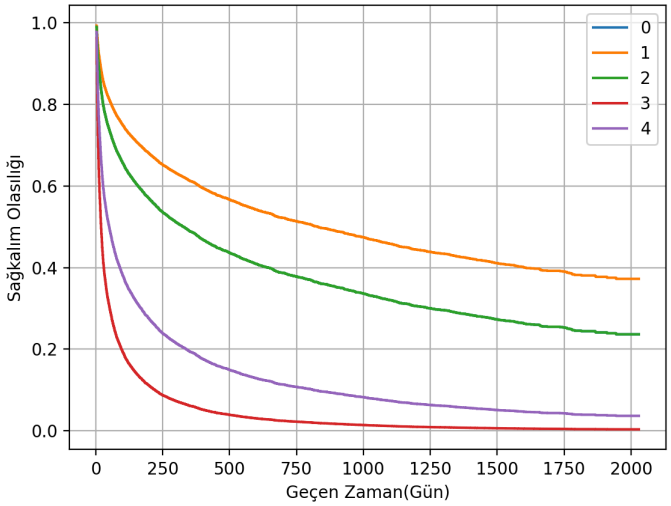
Şekil 5.12 Uygulama 2 RSO için hastaların tehlike oranları grafiği



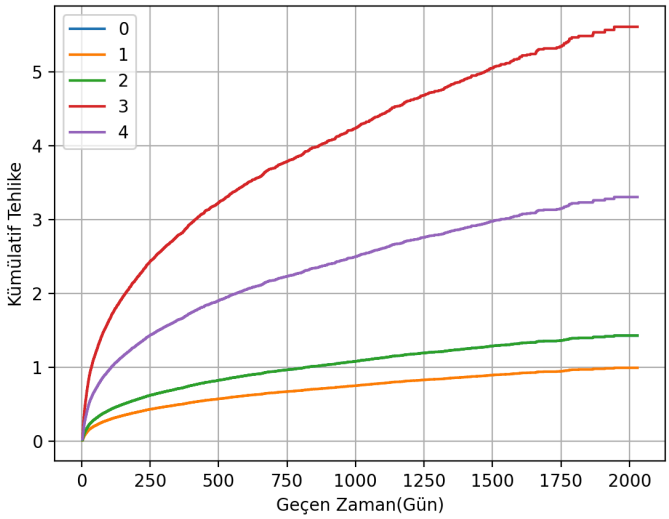
Şekil 5.13 Uygulama 2 ESA için hastaların sağkalım olasılıkları grafiği



Şekil 5.14 Uygulama 2 ESA için hastaların tehlike oranları grafiği



Şekil 5.15 Uygulama 2 GGM için hastaların sağkalım olasılıkları grafiği



Şekil 5.16 Uygulama 2 GGM için hastaların tehlike oranları grafiği

5.2.2 C-indeks Değerleri ve Çalışma Süreleri

Yöntemlerin C-indeks değerleri Tablo 5.4'te yer almaktadır. RSO ve ESA yöntemleri birbirlerine yakın sonuçlar elde ederken GGM yöntemi diğer iki yönteme göre daha düşük C-indeks değerine sahiptir. Eğitim veri setlerinde en yüksek C-indeks değeri 0.8610 ile RSO yöntemine ait iken en düşük C-indeks değeri 0.7413 ile GGM yöntemine aittir. Test veri setlerinde ise en yüksek C-indeks değeri eğitim veri setinde olduğu gibi 0.7655 ile RSO yöntemine ait iken en düşük C-indeks değeri ise 0.7402 ile GGM yöntemine aittir.

Yöntemlerin çalışma süreleri arasında oldukça büyük farklar ortaya çıkmaktadır. ESA yöntemi en hızlı çalışan yöntem olurken, RSO yöntemi bu yönteme oranla 3 kat daha yavaş çalışmaktadır. GGM yöntemi ise yaklaşık 4 saatlik bir çalışma süresi ile diğerler iki yönteme göre çok daha yavaş çalışmaktadır.

Kullanılan üç yöntemin çalışma süreleri sırasıyla:

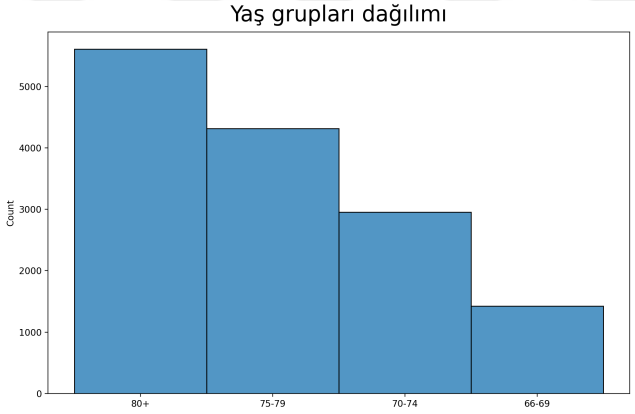
- RSO yöntemi çalışma süresi 2094.43 saniye
- ESA yöntemi çalışma süresi 735.91 saniye
- GGM yöntemi çalışma süresi 14002.25 saniye

Tablo 5.4 Uygulama 2 için yöntemlerin C-indeks değerleri tablosu

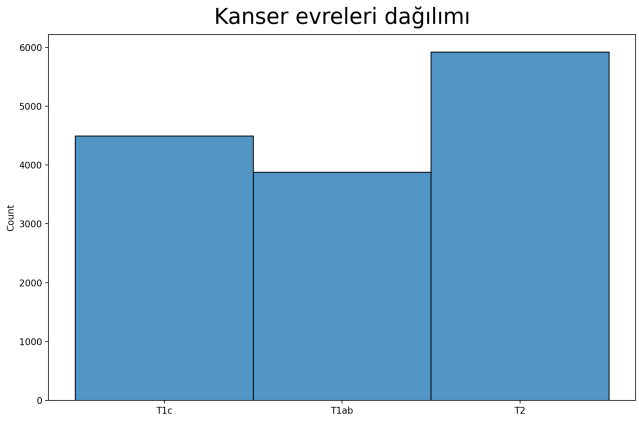
	Eğitim Veri Seti			Test Veri Seti		
Yöntem	**Minimum**	**Maksimum**	**Ortalama**	**Minimum**	**Maksimum**	**Ortalama**
RSO	0.8596	0.8610	0.8603	0.7627	0.7655	0.7640
ESA	0.8083	0.8106	0.8095	0.7584	0.7612	0.7597
GGM	0.7413	0.7502	0.7465	0.7402	0.7492	0.7457

5.3 Uygulama 3 – Prostat Kanseri Hastaları İçin Sağkalım Analizi

Algoritmaların analizi için yapılan üçüncü uygulamada ise SurvSet paketi içerisinde yer alan 3 farklı değişkene sahip 14294 prostat kanseri hastasına ait bir veri setidir. Veri setinde yer alan hastalar 66 yaş ve üzeri hastalardır. Bu hastaların 5000'den fazlası 80 yaş ve üzeri grupta yer almaktadır. Yaş dağılımlarına ait detaylar Şekil 5.17'de görülmektedir. Ayrıca bu hastaların kanser evrelerine ait dağılımda Şekil 5.18'de görülmektedir. Bu veri setine ait değişkenler ve bilgiler Tablo 5.5'te yer almaktadır.



Şekil 5.17 Uygulama 3'te kullanılan veri setine ait yaş gruplarının dağılımı



Şekil 5.18 Uygulama 3'te kullanılan veri setine ait kanser evrelerinin dağılımı

Tablo 5.5 Uygulama 3'te kullanılan veri setine ait değişkenlerin bilgi tablosu

Değişkenler	**Açıklama**	**Değişken Tipi**	**Değerler**
Event	Sağkalım durumu	Sürekli	-
Time	Sağkalım süresi	Sürekli	-
Grade	Kanser yayılma hızı	Kategorik	Normal, Hızlı
Stage	Evre	Kategorik	T1ab, T1c, T2
Agegroup	Yaş grubu	Kategorik	66-69, 70-74, 75-79, 80+

- T1ab, tümör bilgisayarlı tomografi ya da ultrason gibi herhangi bir yöntem ile görüntülenemez. Farklı bir tıbbi durum için yapılan ameliyat sırasında bulunabilir. Çok küçük boyutlu, başlangıç seviyesi bir tümördür.
- T1c, kandaki PSA (prostat içinde bulunan prostat spesifik antijen) seviyesi yüksek olduğu için yapılan biyopsi sırasında tümör bulunur.
- T2, prostatta kanserin olduğu durumdur. Doktor tarafından yapılan fiziksel muayene sırasında ya da çeşitli görüntüleme yöntemleriyle tümör bulunabilir.

5.3.1 Sağkalım Olasılıkları ve Tehlike Oranları

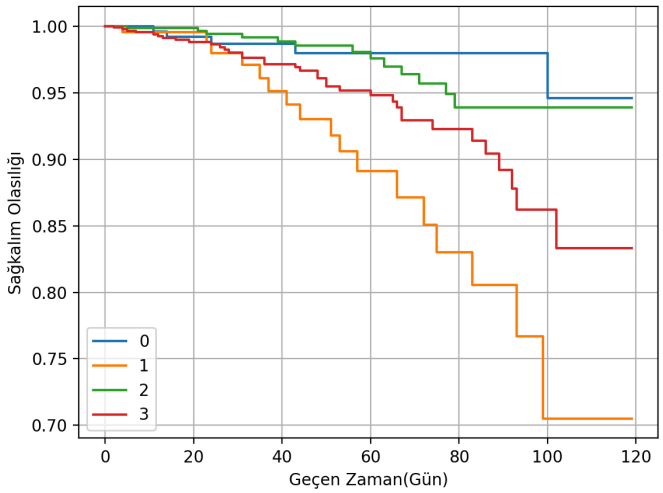
Uygulamanın sağkalım olasılıkları ve tehlike fonksiyonları 4 farklı yaş grubundan seçilen 4 farklı hastadan veriler üzerinden incelenmiştir. Seçilen bu hastalardan ilk hasta 66-69 yaş grubuna ait T1c kanser evresine sahip bir hastadır. İkinci hasta ise 70-74 grubu içinde T2 kanser evresinde olan bir hastadır. 74-79 yaş grubu aralığında olan

üçüncü hastanın kanser evresi ise T1ab'dir. Son hastamız ise 80+ yaş grubunda yer alan hastamız ise T1c kanser evresinde olan bir hastadır. İkinci hastanın kanser yayılma hızı hızlıyken diğer üç hastanın kanser yayılma hızları normaldir.

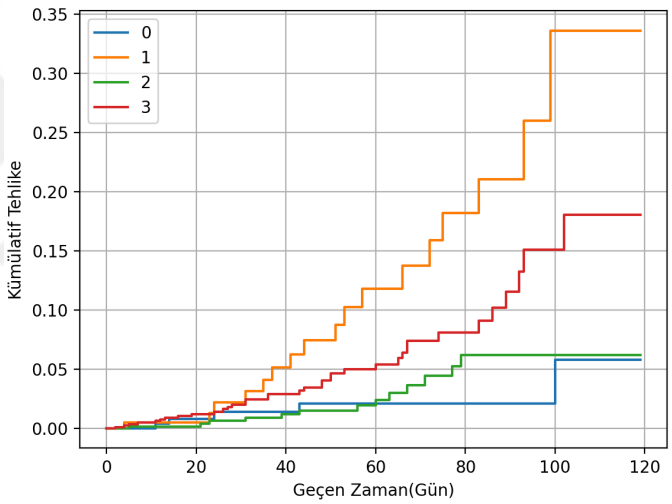
RSO algoritması için sağkalım olasılıkları grafiği Şekil 5.19'da yer alırken, tehlike oranları grafiği Şekil 5.20'de yer almaktadır. ESA algoritmasının sağkalım olasılıkları grafiği 5.21'de gösterilirken, tehlike oranları grafiği ise Şekil 5.22'de gösterilmiştir. Son olarak GGM yöntemine ait sağkalım olasılıkları grafiği 5.23'te ve tehlike oranları grafiği 5.24'te bulunmaktadır.

RSO ve ESA algoritmaları hastalar için yakın sağkalım oranları hesaplamış olup hastaların sağkalım olasılıkları 0.70'ten fazla olduğu gözlemlenmiştir. En düşük sağkalım olasılığına sahip olan hasta ikinci hastadır. Diğer hastalara göre kanseri daha hızlı yayılan ve daha ileri evre kanser hastası olan bu hastanın daha düşük sağkalım oranına sahip olması kanser yayılma hızının ve kanser evresinin sağkalım olasılığı üzerindeki etkisini göstermektedir. Aynı yayılma hızına ve benzer kanser evresine sahip olan hastalar arasında ise en düşük sağkalım olasılığına sahip olan hasta dördüncü yani yaşça en büyük hastadır. Bu da bize benzer koşullarda hasta yaşının sağkalım olasılığı üzerindeki etkisini göstermektedir. Hastalar beklenen ölüm olayını yaşamamıştır. Kanser hastalığının başlangıç evresinde olan hastaların ölüm olayını yaşamamış olması normaldir. Fakat sağkalım oranlarının bir hayli yüksek olmasının nedeni veri setinde yer alan hasta kayıtlarının olay zamanlarının kısa olması ve değişken sayılarının az olmasından kaynaklıdır.

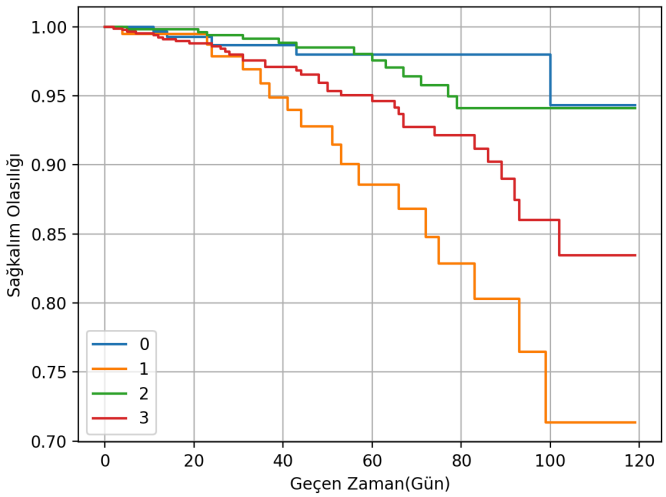
GGM yöntemi ise verimli bir şekilde çalışmamıştır. Diğer iki algoritmayla benzer sonuçlar veren bu yöntem iki hasta için herhangi bir olasılık ve tehlike oranı hesaplayamamıştır. Bunun sebebi ise veri seti her ne kadar büyük bir veri seti olsa da değişken sayısının az olması algoritma performansına doğrudan etki etmiştir.



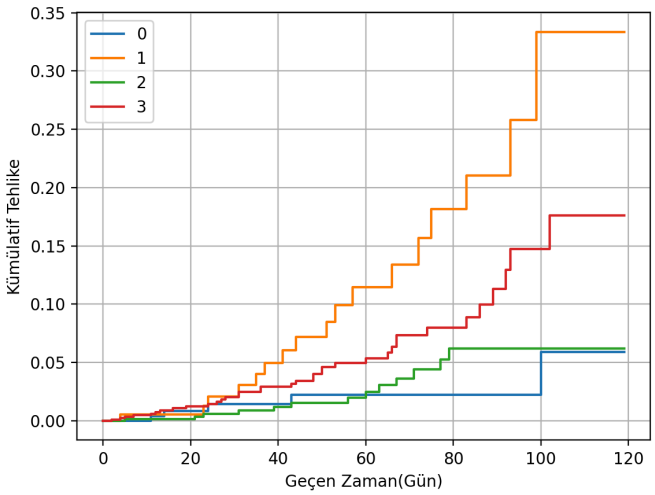
Şekil 5.19 Uygulama 3 RSO için hastaların sağkalım olasılıkları grafiği



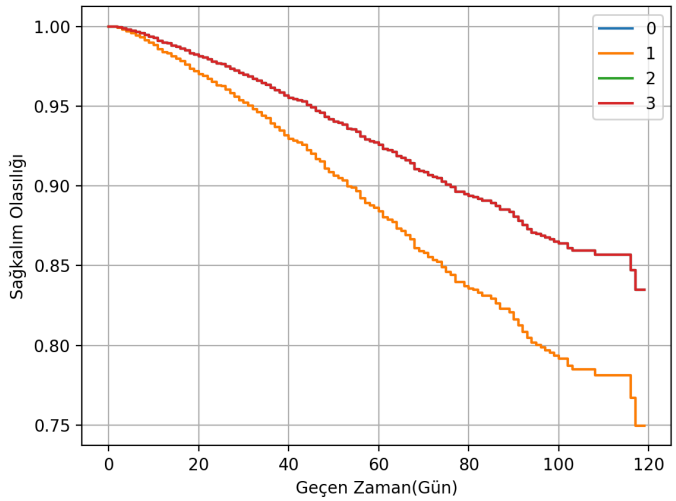
Şekil 5.20 Uygulama 3 RSO için hastaların tehlike oranları grafiği



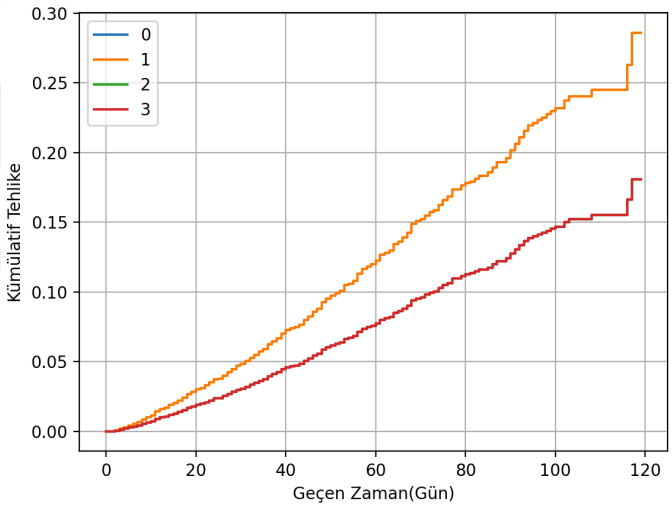
Şekil 5.21 Uygulama 3 ESA için hastaların sağkalım olasılıkları grafiği



Şekil 5.22 Uygulama 3 ESA için hastaların tehlike oranları grafiği



Şekil 5.23 Uygulama 3 GGM için hastaların sağkalım olasılıkları grafiği



Şekil 5.24 Uygulama 3 GGM için hastaların tehlike oranları grafiği

5.3.2 C-indeks Değerleri ve Çalışma Süreleri

Yöntemlerin C-indeks değerleri Tablo 5.6'da yer almaktadır. RSO ve ESA yöntemleri birebir aynı sonuçlar elde ederken GGM yöntemi diğer iki yönteme göre daha düşük C-indeks değerine sahiptir. Eğitim veri setlerinde en yüksek C-indeks değeri 0.7732 ile RSO ve ESA yöntemine ait iken en düşük C-indeks değeri 0.6920 ile GGM yöntemine aittir. Test veri setlerinde ise en yüksek C-indeks değeri eğitim veri setinde olduğu gibi 0.7518 ile ESA yöntemine ait iken en düşük C-indeks değeri ise 0.6793 ile GGM yöntemine aittir.

Yöntemlerin çalışma süreleri incelendiğinde ise ESA ve RSO yöntemlerinin çalışma hızları birbirine çok yakındır. ESA yöntemi, RSO yönteminden 10 saniye daha hızlı çalışarak en hızlı çalışan yöntemken GGM yönteminin çalışma süresi bir hayli fazladır. GGM yöntemi bu veri seti için yaklaşık 7 saat sonunda hesaplamaları tamamlamıştır.

Kullanılan üç yöntemin çalışma süreleri sırasıyla:

- RSO yöntemi çalışma süresi 86.45 saniye
- ESA yöntemi çalışma süresi 77.18 saniye
- GGM yöntemi çalışma süresi 26240.14 saniye

Tablo 5.6 Uygulama 3 için yöntemlerin C-indeks değerleri tablosu

	Eğitim Veri Seti			**Test Veri Seti**		
Yöntem	**Minimum**	**Maksimum**	**Ortalama**	**Minimum**	**Maksimum**	**Ortalama**
RSO	0.7712	0.7732	0.7719	0.7482	0.7517	0.7504
ESA	0.7709	0.7772	0.7718	0.7482	0.7518	0.7504
GGM	0.6920	0.6920	0.6920	0.6793	0.6793	0.6793

BÖLÜM ALTI
SONUÇ

Bu tez çalışmasında sağkalım analizi çalışmalarında yaygın olarak kullanılan rassal sağkalım ormanı yöntemiyle birlikte ekstra sağkalım ormanı ve gradyanla güçlendirilmiş modeller yöntemleri incelenmiştir. Farklı veri setleri ve farklı veri seti boyutlarıyla birlikte bu yöntemlerin performansları incelenmiştir.

Bölüm 5'te yer alan üç farklı uygulamadan iki uygulamada veri setlerinde benzer sayıda değişken bulunurken veri setlerinin boyutları farklıdır. Üçüncü uygulamada ise az değişkene sahip olan fakat diğer iki veri setinden daha fazla veri seti kullanılarak performansın ölçülmesi amaçlanmıştır.

Rassal sağkalım ormanı ilk ve ikinci uygulamada ekstra sağkalım ağaçları benzer sonuçlar elde etmektedir. Yine de rassal sağkalım ormanları biraz daha iyi sonuçlar elde etmektedir. Gradyanla güçlendirilmiş modeller yöntemi ise bu iki algoritmaya göre daha düşük sonuçlar vermektedir.

Üçüncü uygulamada ise rassal sağkalım ormanı ve ekstra sağkalım ağaçları hemen hemen aynı sonuçları vermiştir. Gradyanla güçlendirilmiş modeller ise iki algoritmaya göre daha kötü c-indeks değerine sahiptir.

Kullanılan üç yönteminde çalışma süreleri karşılaştırıldığında küçük veri setine sahip olan ilk uygulamada ekstra sağkalım ağaçları ve rassal sağkalım ormanları birbirlerine yakın bir sürede çalışmalarını tamamlarken gradyanla güçlendirilmiş modeller yaklaşık üç kat daha yavaş çalışmaktadır.

Daha büyük veri setine sahip fakat ilk uygulamayla benzer sayıda değişkeni olan ikinci uygulamada ise veri setinin büyüklüğünün çalışma süresine olan etkisi açık bir şekilde görülmektedir. İlk uygulamaya göre daha uzun çalışma sahip olan ikinci uygulamada en iyi performansı ekstra sağkalım ağaçları göstermektedir. Rassal sağkalım ormanı yöntemi yaklaşık 3 kat daha yavaş çalışırken, gradyanla güçlendirilmiş modeller yaklaşık 20 kat daha yavaş çalışmaktadır.

En büyük veri setine sahip olan üçüncü uygulamada daha az değişken olduğu için rassal sağkalım ormanı ve ekstra sağkalım ağaçları yöntemleri ikinci uygulamaya göre

daha hızlı çalışmaktadır. Buradan ağaç tabanlı yöntemlerin çalışma performansında veri setinin büyüklüğünün yanı sıra değişken sayısındaki farklılığında önemi olduğu ortaya çıkmaktadır. Gradyanla güçlendirilmiş modeller yönteminde ise değişken sayısının çalışma süresinde etkisi olmadığı görülmüştür. Tüm uygulamalar arasında en uzun çalışma süresine sahip olan yöntem gradyanla güçlendirilmiş modeller yöntemi olmuştur.

Çalışma sürelerinde seçilen iterasyon sayısı, oluşturulan minimum örneklem sayısı, oluşturulacak minimum yaprak sayısı gibi parametrelerinde etkisi vardır. İterasyon sayısı artıkça çalışma süreleri de artmaktadır. Uygun parametrelerin seçimi sırasında seçilen denenen bazı iterasyonlarda programların çalışmalarının çok yavaşladığı görülmüştür. Yüksek boyutlu veri setlerinde uygulamalar yarıda kesilebilir ve çalışma sonuçları elde edilemeyebilir.

Sağkalım analizinde makine öğrenimleri alanında gelişmeler devam etmektedir. Özellikle topluluk öğrenmesi yöntemlerinden olan ağaç tabanlı yöntemler sağkalım analizi için iyi sonuçlar vermektedir. Sağkalım analizi için geliştirilen rassal sağkalım ormanı en iyi sonuçları verirken çalışma performansı açısından ekstra sağkalım ağaçları yöntemi iyi bir alternatiftir. Gelecek çalışmalarda büyük veri setleri üzerinde ekstra sağkalım ağaçlarının kullanılması daha performanslar elde edilmesini sağlayabilir ve bunun üzerine yeni araştırmalar yapılabilir.

KAYNAKLAR

Alhashmi, S. M., Polash, M. S. I., Haque, A., Rabbe, F., Hossen, S., Faruqui, N., ... & Abubacker, N. F. (2024). Survival Analysis of Thyroid Cancer Patients Using Machine Learning Algorithms. IEEE Access.

Altuhaifa, F. A., Win, K. T., & Su, G. (2023). Predicting lung cancer survival based on clinical data using machine learning: A review. Computers in Biology and Medicine, 107338.

Blaž Zupan, Janez Demšar, MichaelW. Kattan, Robert J. Beck, ve Ivan Bratko. (2000). Machine learning for survival analysis: A case study on recurrence of prostate cancer. Artificial Intelligence in Medicine 20(1), 59–75.

Bou-Hamad, I., Larocque, D., & Ben-Ameur, H. (2011). A review of survival trees.

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

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

Buk Cardoso, L., Cunha Parro, V., Verzinhasse Peres, S., Curado, M. P., Fernandes, G. A., Wünsch Filho, V., & Natasha Toporcov, T. (2023). Machine learning for predicting survival of colorectal cancer patients. Scientific reports, 13(1), 8874.

Camilleri, L. (2019). History of survival analysis. The Sunday Times of Malta, pp. 53.

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.

Di Castelnuovo, A., Bonaccio, M., Costanzo, S., Gialluisi, A., Antinori, A., Berselli, N., ... & Covid, T. (2020). Common cardiovascular risk factors and in-hospital mortality in 3,894 patients with COVID-19: survival analysis and machine learning-based findings from the multicentre Italian CORIST Study. Nutrition, Metabolism and Cardiovascular Diseases, 30(11), 1899-1913.

Dietterich, T. G. (2000). Ensemble methods in machine learning. In International workshop on multiple classifier systems (pp. 1-15). Berlin, Heidelberg: Springer Berlin Heidelberg.

Erik, D. (2022). SurvSet: An open-source time-to-event dataset repository. arXiv preprint arXiv:2203.03094.

Friedman, J. H. (2001). Greedy function approximation: a gradient boosting machine. Annals of statistics, 1189-1232.

Friedman, J. H. (2002). Stochastic gradient boosting. Computational statistics & data analysis, 38(4), 367-378.

Geurts, P., Ernst, D., & Wehenkel, L. (2006). Extremely randomized trees. Machine learning, 63, 3-42.

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

Harrell, F.E., Califf, R.M., Pryor, D.B., Lee, K.L., Rosati, R.A. (1996). Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors, Statistics in Medicine, 15(4), 361-87.

Heagerty, P. J., & Zheng, Y. (2005). Survival model predictive accuracy and ROC curves. *Biometrics*, *61*(1), 92–105.

Ishwaran, H., Kogalur, U. B., Blackstone, E. H., ve Lauer, M. S. (2008). Random survival forests. https://doi.org/10.48550/arXiv.0811.1645

Kaplan, E. L., & Meier, P. (1958). Nonparametric estimation from incomplete observations. Journal of the American statistical association, 53(282), 457-481.

Kar, İ. (2021). Sağkalım analizi için derin öğrenme yaklaşımı [Doktora tezi]. Ankara Üniversitesi.

Kleinbaum, D. G., ve Klein, M. (2012). Survival analysis a self-learning text. Springer.

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

Macis, A. (2023). Statistical models and machine learning for survival data analysis [Doktora tezi]. Università degli Studi di Brescia.

Papathanasiou, D., Demertzis, K., & Tziritas, N. (2023). Machine failure prediction using survival analysis. Future Internet, 15(5), 153.

Powell, J. L. (1994). Estimation of semiparametric models. Handbook of econometrics, 4, 2443-2521.

Pölsterl, S. (2020). scikit-survival: A Library for Time-to-Event Analysis Built on Top of scikit-learn. Journal of Machine Learning Research, 21(212), 1-6.

Rois, R. (2017). Nonparametric Tests for Change Points in Hazard Functions under Random Censorship in Survival Analysis (Doctoral dissertation, Technische Universität Kaiserslautern).

Safavian, S. R., & Landgrebe, D. (1991). A survey of decision tree classifier methodology. IEEE transactions on systems, man, and cybernetics, 21(3), 660-674.

Scikit-Survival, Python Kütüphanesi, erişim: https://scikit-survival.readthedocs.io/en/latest/index.html

Spooner, A., Chen, E., Sowmya, A., Sachdev, P., Kochan, N. A., Trollor, J., & Brodaty, H. (2020). A comparison of machine learning methods for survival analysis of high-dimensional clinical data for dementia prediction. Scientific reports, 10(1), 20410.

Vallarino, D. (2024). Unveiling the Temporal Layers of Economic Transformations: A Survival Analysis on Unemployment Trends in the Americas. Available at SSRN 4731406.

Wang, P., Li, Y., ve Reddy, C. K. (2019). Machine learning for survival analysis: A survey. ACM Computing Surveys (CSUR), 51(6), 1-36.