T.C.
HACETTEPE ÜNİVERSİTESİ
SAĞLIK BİLİMLERİ ENSTİTÜSÜ

BORUTA VE ELASTİK AĞ ALGORİTMALARININ GEN SEÇİM PERFORMANSLARININ RNA DİZİLEME VERİ SETLERİ ÜZERİNDE KARŞILAŞTIRILMASI: BİR MONTE CARLO BENZETİM ÇALIŞMASI

Özgür SAMAN

Biyoistatistik Programı

YÜKSEK LİSANS TEZİ

ANKARA

2023

T.C.
HACETTEPE ÜNİVERSİTESİ
SAĞLIK BİLİMLERİ ENSTİTÜSÜ

BORUTA VE ELASTİK AĞ ALGORİTMALARININ GEN SEÇİM PERFORMANSLARININ RNA DİZİLEME VERİ SETLERİ ÜZERİNDE KARŞILAŞTIRILMASI: BİR MONTE CARLO BENZETİM ÇALIŞMASI

Özgür SAMAN

Biyoistatistik Programı

YÜKSEK LİSANS TEZİ

TEZ DANIŞMANI

Doç. Dr. Osman DAĞ

ANKARA

2023

BORUTA VE ELESTİK AĞ ALGORİTMALARININ GEN SEÇİM PERFORMANSLARININ RNA DİZİLEME VERİ SETLERİ ÜZERİNDE KARŞILAŞTIRILMASI: BİR MONTE CARLO BENZETİM ÇALIŞMASI

Özgür SAMAN

Danışman: Doç. Dr. Osman DAĞ

Bu tez çalışması 18.8.2023 tarihinde jürimiz tarafından "Biyoistatistik Programı" nda yüksek lisans tezi olarak kabul edilmiştir.

Jüri Başkanı:	Dr. Öğr. Üyesi Sevilay KARAHAN *(Hacettepe Üniversitesi)*
Tez Danışmanı:	Doç. Dr. Osman DAĞ *(Hacettepe Üniversitesi)*
Üye:	Dr. Öğr. Üyesi Erdoğan ASAR *(Sağlık Bilimleri Üniversitesi)*

Bu tez Hacettepe Üniversitesi Lisansüstü Eğitim-Öğretim ve Sınav Yönetmeliğinin ilgili maddeleri uyarınca yukarıdaki jüri tarafından uygun bulunmuştur.

24 Ağustos 2023

Prof. Dr. Müge YEMİŞCİ ÖZKAN

Enstitü Müdürü

YAYIMLAMA VE FİKRİ MÜLKİYET HAKLARI BEYANI

Enstitü tarafından onaylanan lisansüstü tezimin/raporumun tamamını veya herhangi bir kısmını, basılı (kağıt) ve elektronik formatta arşivleme ve aşağıda verilen koşullarla kullanıma açma iznini Hacettepe Üniversitesine verdiğimi bildiririm. Bu izinle Üniversiteye verilen kullanım hakları dışındaki tüm fikri mülkiyet haklarım bende kalacak, tezimin tamamının ya da bir bölümünün gelecekteki çalışmalarda (makale, kitap, lisans ve patent vb.) kullanım hakları bana ait olacaktır.

Tezin kendi orijinal çalışmam olduğunu, başkalarının haklarını ihlal etmediğimi ve tezimin tek yetkili sahibi olduğumu beyan ve taahhüt ederim. Tezimde yer alan telif hakkı bulunan ve sahiplerinden yazılı izin alınarak kullanılması zorunlu metinlerin yazılı izin alınarak kullandığımı ve istenildiğinde suretlerini Üniversiteye teslim etmeyi taahhüt ederim.

Yükseköğretim Kurulu tarafından yayınlanan ***"Lisansüstü Tezlerin Elektronik Ortamda Toplanması, Düzenlenmesi ve Erişime Açılmasına İlişkin Yönerge"*** kapsamında tezim aşağıda belirtilen koşullar haricince YÖK Ulusal Tez Merkezi / H.Ü. Kütüphaneleri Açık Erişim Sisteminde erişime açılır.

- Enstitü / Fakülte yönetim kurulu kararı ile tezimin erişime açılması mezuniyet tarihimden itibaren 2 yıl ertelenmiştir. (1)
- Enstitü / Fakülte yönetim kurulunun gerekçeli kararı ile tezimin erişime açılması mezuniyet tarihimden itibaren ... ay ertelenmiştir. (2)
- Tezimle ilgili gizlilik kararı verilmiştir. (3)

24/08/2023
Özgür SAMAN

i

---

[1] *"Lisansüstü Tezlerin Elektronik Ortamda Toplanması, Düzenlenmesi ve Erişime Açılmasına İlişkin Yönerge"*

(1) *Madde 6. 1. Lisansüstü tezle ilgili patent başvurusu yapılması veya patent alma sürecinin devam etmesi durumunda, tez* ***danışmanının*** *önerisi ve* ***enstitü anabilim dalının*** *uygun görüşü üzerine* ***enstitü*** *veya* ***fakülte yönetim kurulu*** *iki yıl süre ile tezin erişime açılmasının ertelenmesine karar verebilir.*

(2) *Madde 6. 2. Yeni teknik, materyal ve metotların kullanıldığı, henüz makaleye dönüşmemiş veya patent gibi yöntemlerle korunmamış ve internetten paylaşılması durumunda 3. şahıslara veya kurumlara haksız kazanç imkanı oluşturabilecek bilgi ve bulguları içeren tezler hakkında tez* ***danışmanının*** *önerisi ve* ***enstitü anabilim dalının*** *uygun görüşü üzerine* ***enstitü*** *veya* ***fakülte yönetim kurulunun*** *gerekçeli kararı ile altı ayı aşmamak üzere tezin erişime açılması engellenebilir.*

(3) *Madde 7. 1. Ulusal çıkarları veya güvenliği ilgilendiren, emniyet, istihbarat, savunma ve güvenlik, sağlık vb. konulara ilişkin lisansüstü tezlerle ilgili gizlilik kararı,* ***tezin yapıldığı kurum*** *tarafından verilir *. Kurum ve kuruluşlarla yapılan işbirliği protokolü çerçevesinde hazırlanan lisansüstü tezlere ilişkin gizlilik kararı ise,* ***ilgili kurum ve kuruluşun önerisi*** *ile* ***enstitü*** *veya* ***fakültenin*** *uygun görüşü üzerine* ***üniversite yönetim kurulu*** *tarafından verilir. Gizlilik kararı verilen tezler Yükseköğretim Kuruluna bildirilir.*
Madde 7.2. Gizlilik kararı verilen tezler gizlilik süresince enstitü veya fakülte tarafından gizlilik kuralları çerçevesinde muhafaza edilir, gizlilik kararının kaldırılması halinde Tez Otomasyon Sistemine yüklenir

****** Tez* ***danışmanının*** *önerisi ve* ***enstitü anabilim dalının*** *uygun görüşü üzerine* ***enstitü*** *veya* ***fakülte yönetim kurulu tarafından karar verilir.***

ETİK BEYAN

Bu çalışmadaki bütün bilgi ve belgeleri akademik kurallar çerçevesinde elde ettiğimi, görsel, işitsel ve yazılı tüm bilgi ve sonuçları bilimsel ahlak kurallarına uygun olarak sunduğumu, kullandığım verilerde herhangi bir tahrifat yapmadığımı, yararlandığım kaynaklara bilimsel normlara uygun olarak atıfta bulunduğumu, tezimin kaynak gösterilen durumlar dışında özgün olduğunu, Doç. Dr. Osman DAĞ danışmanlığında tarafımdan üretildiğini ve Hacettepe Üniversitesi Sağlık Bilimleri Enstitüsü Tez Yazım Yönergesine göre yazıldığını beyan ederim.

Özgür SAMAN

TEŞEKKÜR

Yüksek lisans eğitimim sırasında ve tez dönemimin her aşamasında bilgi ve deneyimleriyle bana her zaman yol gösteren, emeklerini esirgemeyen, beni sürekli motive eden, öğrencisi olmaktan her zaman gurur duyacağım danışman hocam Doç. Dr. Sayın Osman DAĞ'a en içten dileklerimle saygılarımı sunuyor, teşekkürü bir borç biliyorum.

Bu tezin hazırlandığı tüm süreçte gerek analizlerin yapılması, gerekse tez yazım sürecinde değerli bilgilerini benimle paylaşan Sayın Merve KAŞIKCI'ya çok teşekkür ediyorum.

Ayrıca, Yüksek lisans eğitimim boyunca bizlere özverili bir eğitim sunan Biyoistatistik Anabilim Dalında bulunan tüm hocalarıma teşekkür ederim.

Son olarak manevi desteğini hiçbir zaman eksik etmeyen ve bu süreci de tamamlamamı sabırla bekleyen sevgili eşime sonsuz teşekkürlerimi sunuyorum.

ÖZET

Saman, Ö., Boruta ve Elastik Ağ Algoritmalarının Gen Seçim Performanslarının RNA Dizileme Veri Setleri Üzerinde Karşılaştırılması: Bir Monte Carlo Benzetim Çalışması, Hacettepe Üniversitesi Sağlık Bilimleri Enstitüsü Biyoistatistik Programı Yüksek Lisans Tezi, Ankara, 2023. Bu tez çalışmasında; Kanser Genom Atlası Programından elde edilen farklı kanser türlerine ait sekiz gerçek RNA dizileme veri seti kullanılarak ilgili kanser hastalığına önemli derecede etki eden genlerin seçilmesinde, Boruta Algoritması ve Elastik Ağ Genelleştirilmiş Doğrusal Modeller İçerisinde Determan'ın Algoritması uygulanmıştır. Boruta Algoritmasının kendi içerisinde bulunan, sınıflandırma yöntemlerine ve önemlilik ölçütlerine göre farklılaşan yöntemlerden yedi tanesine ait algoritmalar, ayrı ayrı olarak uygulanmış ve elde edilen sonuçlar sadece Elastik Ağ Algoritması ile değil aynı zamanda birbirleri ile de kıyaslanmıştır. Söz konusu RNA dizileme veri setlerinin her birinde, sınıf dağılımları dengesiz olan iki sınıflı bir yanıt değişkeni bulunmaktadır. Gerçek veri setlerine dayalı olarak bir Monte Carlo benzetim çalışması yapılmıştır. Özellik seçiminde kullanılan gen setleri; filtreleme, normalleştirme, dönüşüm ve tek değişkenli analiz olmak üzere dört aşamada ön işleme adımları uygulanarak elde edilmiştir. Böylece, çeşitli filtrelerle ön işleme adımları tamamlanan veri setlerine özellik seçimi yöntemleri uygulanarak model performansları incelenmiştir. Özellik seçimi yöntemlerinin performansları Pozitif Kestirim Değeri, Duyarlılık, F1 Ölçüsü gibi ölçüler kullanılarak karşılaştırılmıştır. Çalışmada kullanılan veri setlerinin tümünde Elastik Ağ Algoritması, Pozitif Kestirim Değeri açısından öne çıkmıştır. Boruta Algoritmasının Ekstra Ağaçlar (Extra Trees - Extremely Randomized Trees) ve Random Ferns tabanlı yöntemleri, Duyarlılık Oranları açısından Elastik Ağ Algoritmasından daha iyi performans göstermiştir.

Anahtar Kelimeler: Boruta, Elastik Ağ, özellik seçimi, RNA dizileme, benzetim, kanser.

ABSTRACT

Saman, Ö., Comparison of Gene Selection Performances of Boruta and Elastic Net Algorithms on RNA Sequencing Data sets: A Monte Carlo Simulation Study, Hacettepe University Graduate School of Health Sciences Master Thesis in Biostatistics, Ankara, 2023. In this thesis, Boruta Algorithm and Determan's Algorithm in Elastic Net Generalized Linear Models were applied to select the genes that have a significant effect on the related cancer disease using eight real RNA sequencing data set belonging to different cancer types obtained from the Cancer Genome Atlas Program. Algorithms belonging to seven of the methods within the Boruta Algorithm, which differ according to different classification methods and various importance criteria, were applied separately and the results obtained were compared not only with the Elastic Network Algorithm but also with each other. Each of these RNA sequencing data sets contains a two-class response variable with unbalanced class distributions. A Monte Carlo simulation study was performed based on the real data sets. The gene sets used for feature selection were obtained by applying preprocessing steps in four stages: filtering, normalization, transformation and univariate analysis. Thus, feature selection methods were applied to the data sets that were preprocessed with various filters and model performances were analyzed. The performances of the feature selection methods were compared using measures such as Precision, Recall and F1 Measure. For all of the data sets used in the study, Elastic Net Algorithm stood out in terms of Precision. Boruta Algorithm based on Extra Trees (Extremely Randomized Trees) and Random Ferns outperformed the Elastic Net Algorithm in terms of Recall.

Key Words: Boruta, Elastic Net, feature selection, RNA-seq, simulation, cancer.

İÇİNDEKİLER

SİMGELER VE KISALTMALAR

BA Dengeli Doğruluk

COAD Kolon Adenokarsinomu

CRAN Kapsamlı R Arşiv Ağı

DEG Diferansiyel olarak ifade edilen gen

EA Elastik Ağ

KICH Böbrek Kromofobisi

KIRC Böbrek Renal Şeffaf Hücreli Karsinom

KCV K-katlı Çapraz Doğrulama

LIHC Karaciğer Hepatoselüler Karsinomu

LUAD Akciğer Adenokarsinomu

LUSC Akciğer Yassı Hücreli Karsinomu

MCC Matthews Korelasyon Katsayısı

MZSA Gölge nitelikler arasında maksimum Z puanı

PPV Pozitif Kestirim Değeri

PRAD Prostat Adenokarsinomu

TCGA Kanser Genom Atlası

THCA Tiroid Karsinomu

ŞEKİLLER

TABLOLAR

1. GİRİŞ

Özellik sayısı bakımından yüksek boyutlu veriler, günümüzde makine öğrenimi problemlerinde giderek daha yaygın hale gelmektedir. Bu yüksek hacimli verilerden faydalı bilgilerin elde edilmesi için gürültünün (noise) veya gereksiz verilerin azaltılması amacıyla istatistiksel tekniklerin kullanılması gerekmektedir. Bunun nedeni, bir modelin eğitilmesi için genellikle her bir özelliğin kullanılmasına gerek olmamasıdır. Bir model, yalnızca ilişkisiz ve gereksiz olmayan özellikler kullanılarak geliştirilebilmektedir. Bu noktada özellik seçiminin önemi oldukça büyüktür. Özellik seçimi, sadece modelin daha hızlı eğitilmesine yardımcı olmakla kalmamakta aynı zamanda modelin karmaşıklığını azaltmakta ve sonuçların yorumlanmasını kolaylaştırmaktadır (1).

Özellik seçimi, makine öğrenimi yöntemlerinin uygulamalarında önemli bir husustur, çünkü korelasyonlu değişkenlerden ve istenmeyen gürültülerden arındırılmış tahmin modelleri oluşturmaya yardımcı olmaktadır. Pratik modeller oluşturulması için modern veri setleri, genellikle çok fazla değişkenle tanımlanmaktadır. Aşırı büyük özellik setleriyle uğraşmanın çeşitli dezavantajları bulunmaktadır. Bunlardan birisi teknik sebeplere dayanmakta olup büyük özellik setleriyle uğraşılması, algoritmaları yavaşlatmakta ve çok fazla kaynak gerektirmektedir. Oldukça önemli olan bir diğer dezavantaj ise, birçok makine öğrenimi algoritmasının değişken sayısı optimumdan önemli ölçüde fazla olduğunda doğrulukta bir düşüş sergilemesidir (2, 3).

Diferansiyel olarak ifade edilen genlerin (DEG) tespit edilmesi, hastalıkların tanı ve tedavileri için oldukça önemli bir süreçtir. Bu çalışmada, DEG'lerin seçilmesi için Boruta ve Elastik Ağ Algoritmalarının kapasiteleri araştırılmıştır. Bu amaçla önce Kanser Genom Atlası (The Cancer Genome Atlas - TCGA) programından sekiz gerçek veri seti elde edilmiştir. İkinci olarak, söz konusu veri setlerinin; gen sayısı, gözlem sayısı, ortalama vektörler ve sınıf oranları gibi özellikleri bulunmuştur. Gerçek veri setlerine dayalı olarak bir Monte Carlo benzetim çalışması yapılmıştır. Oldukça dengesiz sınıf dağılımlarına sahip bahse konu veri setleri kullanılarak iki grup arasında

önemli ölçüde farklı 10 genin üretildiği toplam 19.947 gen ile 8 senaryo üzerinde çalışılmıştır.

Bu çalışmanın amacı, sekiz gerçek RNA dizileme veri setine dayalı olarak tasarlanan kapsamlı bir Monte Carlo benzetim çalışması ile Boruta ve Elastik Ağ Algoritmalarının gen seçim performanslarını değerlendirmektir. Çalışmada kullanılan veri setlerinde yer alan genlerin tamamının özellik seçimine dâhil edilmesi hem işlem sürelerini uzatmakta, hem de model performanslarını düşürebilmektedir. Bu nedenle, ön işleme adımları uygulanarak boyut azaltma işlemi gerçekleştirilmiştir. Filtreleme, normalleştirme, dönüşüm ve tek değişkenli analiz yöntemleri kullanılarak dört aşamada ön işleme yapılmıştır. Her bir veri setinde bulunan ve 19.947 olan gen sayısı ön işleme aşamalarının ardından, sekiz farklı modelde kullanılacağı ve 1.000 tekrarlı benzetim çalışması yapılacağı için 200'e düşürülmüş ve bu 200 gen kullanılarak modellerin özellik seçimi performansları kıyaslanmıştır.

Çalışma 5 temel bölümden oluşmaktadır. Bölüm 2'de özellik seçiminde kullanılan yöntemlerden ve performansları karşılaştırmada kullanılan ölçülerden bahsedilmiştir. Bölüm 3'te çalışmada kullanılan RNA dizileme veri setleri hakkında bilgi verilmiştir. Ayrıca, veri setlerinin elde edilme sürecine, benzetim çalışmasına ve özellik seçimine uygun hale getirilme aşamalarına da yer verilmiştir. Bölüm 4'te özellik seçimi analizlerinin sonuçları, her bir veri setine ve özellik seçimi yöntemlerine göre tablo ve grafiklerle özetlenmiştir. Bölüm 5'te özellik seçimi yöntemlerinin performansları karşılaştırılmış ve alanyazında yer alan benzer çalışmaların sonuçlarından bahsedilmiştir. Son bölümde, bölüm 6'da, çalışmada ulaşılan genel sonuçlara yer verilmiştir.

2. GENEL BİLGİLER

Büyük ölçekli gen verilerinde ilgili hastalık açısından önemli olan genlerin tanımlanması, hastalık mekanizmasının anlaşılmasında kritik öneme sahiptir. Büyük miktarda veriyi işleyebildikleri için önemli genler üzerinde yapılan araştırmalarda çoğunlukla makine öğrenimi algoritmaları uygulanmaktadır. Makine öğrenimi algoritmaları, önemli genlerin keşfedilmesinde geleneksel istatistiksel yöntemlerden (faktör analizi, kümeleme analizi vb.) daha etkili olabilmektedir. Çünkü;

- Genler arasındaki ilişkileri dikkate almaktadırlar,
- Büyük miktarda veriyi işleyebilmektedirler,
- Herhangi bir dağılım varsayımı bulunmamaktadır (4).

2.1. Özellik Seçimi

Özellik seçimi, bir makine öğrenimi işlem hattının belirleyici bir parçasıdır. Özellik seçimi sürecinde çok muhafazakar davranılarak önemsiz özellikler çıkarılmadan devam edilmesi gereksiz gürültünün ortaya çıkması anlamına gelmektedir. Diğer taraftan, bu süreçte çok agresif davranılarak önemli özellikler çıkarılarak devam edilmesi yararlı bilgilerin atılması anlamına gelmektedir. Özellikler hakkında önemli kararlar vermek, tahmin modelinin başarısını sağlamak için kritik öneme sahiptir (5).

Özellik seçimi, tahmine dayalı modelleme için oldukça önemlidir. Gerek özel ve gerekse kamu kurumları her türlü işlemleri için verileri izlemekte ve çeşitli niteliklere ilişkin bilgi toplamaktadır. Bu durum, tahmine dayalı bir model için çok fazla tahmin ediciye erişim olanağı sunmaktadır. Ancak her özellik belirli bir görevin tahmini için önemli değildir. Bu nedenle, önemli özelliklerin belirlenmesi ve gereksiz özelliklerin ortadan kaldırılması önem arz etmektedir. Aşağıda yer alan maddeler özellik seçiminin neden önemli olduğunu açıklamaktadır.

- Gereksiz bir özelliğin çıkarılması model doğruluğunun artırılmasına yardımcı olmaktadır. Benzer şekilde, ilgili bir özelliğin dâhil edilmesi de model doğruluğu üzerinde olumlu bir etkiye sahiptir.

- Çok fazla özellik aşırı uyuma neden olabilmekte, bu da modelin genelleştirilemeyeceği anlamına gelmektedir.
- Çok fazla özellik bulunması, hesaplamanın yavaşlamasına neden olduğundan daha fazla bellek ve donanım gerektirmektedir (6).

Genel olarak üç tür özellik seçme yöntemi bulunmaktadır:

1. **Filtre Yöntemleri**: Filtre yöntemleri genellikle bir ön işleme adımı olarak kullanılmaktadır. Özelliklerin seçimi herhangi bir makine öğrenimi algoritmasından bağımsızdır. Bunun yerine özellikler, sonuç değişkeni ile korelasyonları için çeşitli istatistiksel testlerdeki puanlarına göre seçilmektedir. Bazı yaygın filtre yöntemleri arasında korelasyon metrikleri (Pearson, Spearman), Ki-Kare testi, Anova ve Fisher Skoru bulunmaktadır.
2. **Sarmalayıcı Yöntemler**: Özelliklerin bir alt kümesi kullanılarak bir model eğitmeye çalışılmaktadır. Önceki modelden elde edilen çıkarımlara dayanarak, alt kümeye özellik eklemeye veya çıkarmaya karar verilmektedir. İleri seçim ve geriye doğru eleme, sarmalayıcı yöntemler için verilebilecek örnekler arasında yer almaktadır.
3. **Gömülü Yöntemler**: Kendi yerleşik özellik seçim yöntemlerine sahip algoritmalardır. Lasso regresyonu buna bir örnektir (1).

2.2. Özellik Seçiminde Kullanılan Makine Öğrenimi Algoritmaları

Makine öğrenimi algoritmaları, genler arasındaki ilişkileri dikkate almakta ve gen seçimini zaman açısından verimli bir şekilde gerçekleştirmektedir. Ayrıca, bu algoritmalar özellik seçimi konusunda istenmeyen gürültülerden ve korelasyonlu değişkenlerden arındırılmış tahmin modelleri oluşturmaya yardımcı olmaktadır. Bu çalışmada, özellik seçimi konusunda makine öğrenimi algoritmaları arasında yüksek performans gösteren Boruta ve Elastik Ağ Algoritmaları karşılaştırılmıştır.

2.2.1. Boruta Algoritması

Boruta Algoritması, Rastgele Orman (Random Forest) sınıflandırıcısı etrafında oluşturulmuş bir sarmalayıcı yaklaşımı kullanmaktadır (7). Algoritma, gerçek özelliklerin uygunluğunu rastgele özellikler üzerinden elde edilen bir eşik değeri ile

karşılaştırarak belirleme fikrine dayanmaktadır (8). Bir bilgi sisteminden önemli ve önemsiz özelliklerin tarafsız ve istikrarlı bir şekilde seçilmesini sağlayan yeni bir Rastgele Orman tabanlı özellik seçim yöntemi olan Boruta Algoritması, Kursa ve Rudnicki tarafından 2010 yılında geliştirilmiştir. Yinelemeli yapısı sayesinde, yöntem hem Rastgele Orman önem ölçütünün dalgalanan doğasıyla hem de özellikler arasındaki etkileşimlerle başa çıkabilmektedir (2). Söz konusu yöntem Boruta isimli bir R paketi olarak CRAN'de (CRAN: Comprehensive R Archive Network) mevcut bulunmaktadır.

Boruta ismi Slav mitolojisindeki orman ruhunun adından gelmektedir. Aslında, Boruta tek başına bir algoritma değildir, Rastgele Orman Algoritmasının temeline dayanmaktadır. Algoritmanın çalışma mantığının anlaşılabilmesi için öncelikle Rastgele Orman Algoritması üzerinde durulmuştur. Rastgele Orman Yöntemi, yüksek derecede tahmin doğruluğu elde etmek için önyükleme toplaması (bootstrap aggregation) ve tahmin edicilerin rastgeleleştirilmesini kullanan bir regresyon ağacı tekniğidir. Bu yöntemde, bir sınıflandırma işleminin sonucu modellerden gelen oyların çoğunluğundan oluşmakta, bir regresyon işleminin sonucu ise çeşitli modellerin ortalamasından oluşmaktadır (9, 10).

Bir sınıflandırma görevinde, Rastgele Orman Yönteminin özellik önemini (feature importance) tahmin etme şekli iki aşamalı olarak çalışmaktadır. İlk aşamada, her karar ağacı bir tahmin oluşturmakta ve bu tahmin değerlerini depolamaktadır. İkinci aşamada, belirli özelliklerin değerleri çeşitli eğitim örneklerinde rastgele permütasyona tabi tutulmakta ve tahminlerin sonucu tekrar izlenerek önceki adım tekrarlanmaktadır. Tek bir karar ağacındaki bir özelliğin önemi, orijinal özellikleri kullanan model ile permütasyona uğramış özellikleri kullanan model arasındaki performans farkının eğitim setindeki örnek sayısına bölünmesiyle hesaplanmaktadır. Bir özelliğin önemi, o özellik için tüm ağaçlardaki ölçümlerin ortalamasından oluşmaktadır. Bu prosedür sırasında, her bir özellik için Z Skorları hesaplanmamakta olup Boruta Yöntemi burada devreye girmektedir (9).

Rastgele Orman Yönteminde Z Skoru, ortalama doğruluk kaybının standart sapmasına bölünmesiyle hesaplanmakta ve tüm değişkenler için önem ölçüsü olarak kullanılmaktadır. Ancak bu yöntemde hesaplanan Z Skoru, değişken öneminin bulunması için bir ölçü olarak kullanılamamakta, çünkü bu skor değişken öneminin istatistiksel anlamlılığı ile doğrudan ilişkili değildir. Bu sorunun aşılması için Boruta Paketi hem orijinal hem de rastgele değişkenler üzerinde Rastgele Orman Algoritmasını çalıştırmakta ve tüm özelliklerin önemini hesaplamaktadır (6). Önem ölçüsü, Rastgele Orman Algoritmasının rastgeleliğine bağlı olarak değişmektedir. (Rastgele Orman Yönteminde her defasında farklı ağaç kurulmakta ve farklı sonuçlar elde edilmektedir.) İlaveten, veri setindeki önemli olmayan özelliklerin varlığına da duyarlıdır. Bu sebeple, istatistiksel olarak geçerli sonuçların elde edilmesi için bu işlem yinelemeli bir şekilde devam etmektedir (11).

Ayrıca, bireysel Rastgele Orman çalışmalarında sürekli olarak yüksek önem puanları alan özelliklerin önemli olarak seçildiği açıkça görülmektedir. Öte yandan, bireysel puanlarda oldukça büyük bir değişkenlik gözlemlenebilmektedir. Rastgele bir özelliğin tek yinelemedeki en yüksek puanı, iki önemli özelliğin en yüksek önem puanından daha yüksek olabilmektedir. Bu durum, Boruta Algoritmasının sonuçlarının genellikle tek bir Rastgele Orman çalışmasına dayalı özellik seçimi yöntemleriyle üretilen sonuçlardan daha istikrarlı olduğunu ve bu nedenle birkaç yinelemenin gerekli olduğunu açıkça göstermektedir (2).

Bir veri setine Rastgele Orman modeli uygularken, her yineleme sonucunda iyi performans göstermeyen özellikler özyinelemeli (recursive) olarak çıkarılmaktadır. Yöntem, Rastgele Orman modelinin hatasını en aza indirdiğinden sonunda minimum optimum özellik alt kümesine yol açmaktadır. Bu durum, girdi veri setinin aşırı budanmış bir versiyonunun seçilmesiyle gerçekleşmekte ve böylece bazı ilgili özellikler çıkarılmaktadır. Öte yandan Boruta Algoritması, karar değişkeniyle güçlü ya da zayıf bir şekilde ilgili olan tüm özellikleri bulmaktadır. Bu da onu, hangi insan genlerinin (özelliklerin) belirli bir tıbbi durumla (hedef değişken) bir şekilde bağlantılı olduğunu belirlemekle ilgilenilebilecek biyomedikal uygulamalar için oldukça uygun bir hale getirmektedir (12).

Boruta Algoritmasının çalışması sırasında Rastgele Orman Yönteminin yineleme sayısının sınırlı olması nedeniyle, hesaplama sonucunda özellikler hala onaylanmadığında ya da reddedilmediğinde ve nihayetinde geçici olarak işaretlenen özellikler olduğunda algoritma erken durmaya zorlanabilmektedir. Geçici özelliklerin kalması durumunda, yineleme sayısı sınırının artırılması gerekmektedir. Bununla birlikte, MZSA'ya (Gölge nitelikler arasında maksimum Z puanı) o kadar yakın öneme sahip özellikler olabilir ki Rastgele Orman Algoritmasının gerçekleştirdiği yineleme sayısı, Boruta Algoritmasının istenen güvenle karar vermesi için yeterli olmayabilir. Bazı durumlarda, önemsiz özellikler en önemli gölge özellikten daha yüksek bir Z Skoru elde edebilir. Bu durumda, istatistiksel olarak anlamlı bir karara varmak için birden fazla Rastgele Orman çalışmasına bu nedenle ihtiyaç duyulmaktadır. Bu bağlamda; Boruta Algoritması önemli ve ilgisiz özellikler arasında istatistiksel olarak anlamlı bir ayrım elde etmek için birkaç Rastgele Orman çalışması gerçekleştirmektedir. Tek bir Rastgele Orman çalışmasında elde edilen sıralamanın Boruta Algoritmasından elde edilene oldukça benzer olması beklenmektedir. Bir özelliğin önem ölçüsü, nesneler arasında özellik değerlerinin rastgele permütasyonunun neden olduğu sınıflandırma doğruluğu kaybı olarak elde edilmektedir. Sınıflandırma için belirli bir özelliği kullanan ormandaki tüm ağaçlar için ayrı ayrı hesaplanmaktadır. Ardından doğruluk kaybının ortalaması ve standart sapması hesaplanmaktadır. Alternatif olarak, ortalama kaybın standart sapmaya bölünmesiyle hesaplanan Z Skoru özellik öneminin ölçülmesinde doğrudan kullanılamadığından, herhangi bir özelliğin öneminin anlamlı olup olmadığına yani rastgele dalgalanmalardan kaynaklanabilecek önemden ayırt edilip edilemeyeceğine karar verilebilmesi için bazı dış referanslara ihtiyaç duyulmaktadır. Bu amaçla, bilgi sistemi tasarım gereği rastgele olan özelliklerle genişletilmektedir. Her bir özellik için değerleri, orijinal özelliğin değerlerinin karıştırılmasıyla elde edilen bir “gölge” (shadow) özellik oluşturulmaktadır. Ardından, genişletilmiş bu veri setinin tüm özellikleri kullanılarak bir sınıflandırma yapılmakta ve tüm özelliklerin önemi hesaplanmaktadır (2). Gölge özelliklerin önem kümesi, hangi özelliklerin gerçekten önemli olduğuna karar verebilmek için bir referans ölçüsü olarak kullanılmaktadır.

Boruta Algoritması, R paketi randomForest'te (13) uygulanan Rastgele Orman sınıflandırma algoritması etrafında oluşturulmuş bir sarmalayıcıdır. Rastgele Orman Algoritması nispeten hızlıdır, genellikle parametreler ayarlanmadan çalışabilmekte ve özelliklerin öneminin sayısal birer tahminini vermektedir. Sınıflandırmanın, birden fazla yansız zayıf sınıflandırıcının (karar ağaçlarının) oylanmasıyla gerçekleştirildiği bir topluluk yöntemidir. Kısacası Boruta Algoritması, Rastgele Orman sınıflandırıcısının temelini oluşturan sisteme rastgelelik ekleyerek ve rastgele örneklerden oluşan bir topluluktan sonuçlar toplayarak rastgele dalgalanmaların ve korelasyonların yanıltıcı etkisinin azaltılabileceği fikrine dayanmaktadır. Burada, bu ekstra rastgelelik hangi özelliklerin gerçekten önemli olduğuna dair daha net bir görüş sağlamaktadır (2).

Diğer taraftan, Boruta Algoritması istatistiksel olarak temellendirilmiş ve kullanıcı tarafından herhangi bir özel girdi olmadan bile son derece iyi çalışan bir özellik seçim algoritmasıdır. Bu algoritmada özellikler kendi aralarında rekabet etmemekte, bunun yerine rastgele bir versiyonu ile rekabet etmektedirler. Uygulamada özellik matrisinden (X) başlayarak, her bir özellik rastgele karıştırılarak başka bir veri çerçevesi oluşturulmaktadır. Bu karıştırılmış özellikler, gölge özellikler olarak adlandırılmaktadır. Bu noktada, gölge veri çerçevesi orijinal veri çerçevesine eklenerek X'in iki katı sütuna sahip yeni bir veri çerçevesi elde edilmektedir. Daha sonra, her bir orijinal özelliğin önemi bir eşikle karşılaştırılmaktadır. Eşik, gölge özellikler arasında kaydedilen en yüksek özellik önemi olarak tanımlanmaktadır. Bir özelliğin öneminin söz konusu eşik değerinden daha yüksek olması halinde, bu durum "isabet" olarak adlandırılmaktadır. Diğer bir ifadeyle, bir özelliğin yalnızca en iyi rastgele özellikten daha iyisini yapabilmesi durumunda yararlı olduğu fikrine dayanmaktadır. Boruta Algoritmasında red ve kabul alanları arasında kesin bir eşik yoktur. Bunun yerine, üç farklı alan bulunmaktadır:

- **Kabul alanı**: Burada bulunan özellikler tahmin edici olarak kabul edilmekte ve bu nedenle tutulmaktadırlar.
- **Ret alanı**: Buraya gelen özellikler gürültü olarak kabul edilmekte ve bu yüzden atılmaktadırlar.

- **Kararsızlık alanı**: Boruta Algoritması bu alandaki özellikler konusunda kararsızdır; Boruta Algoritmasının kabul etmeyi ya da reddetmeyi başaramadığı özelliklerdir ve seçim veri bilimcisine kalmaktadır. Bu özellikler kullanım durumuna bağlı olarak red ya da kabul edilmektedir (5).

Bu çalışmada yapılan analizler, istatistiksel hesaplama ve grafikler için yazılım ortamı sağlayan R programlama dili kullanılarak yapılmıştır. R'de kullanılan Boruta Paketi, ayarlanması gereken çok fazla parametre olmadığı için kullanımı kolay bir pakettir. Boruta Algoritması ile önemli özelliklerin kontrol edilmesi için kullanılacak veri setinin eksik değer içermemesi gerekmektedir. Ayrıca bu algoritma, herhangi bir sınıflandırma ya da regresyon probleminde de kullanılabilmektedir (12).

R yazılımında özellik seçimi için birçok paket bulunmaktadır. Özellik seçiminde Boruta Paketi'nin tercih edilmesinin sebepleri aşağıda açıklanmaktadır.

- Hem sınıflandırma hem de regresyon problemleri için oldukça iyi sonuçlar vermektedir.
- Özellik seçimi için oldukça popüler bir yöntem olan Rastgele Orman Yönteminin özellik önem ölçüsüne dayanılarak geliştirilmiştir.
- Rastgele Orman önem ölçütünün dalgalanan doğası ile başa çıkabilmektedir.
- Özellikler arasındaki etkileşimleri ele alabilmektedir.
- Sonuç değişkeniyle ilgili değişkenleri dikkate alan tüm ilgili değişken seçimi yöntemini izlemektedir (6).

Ayrıca, Boruta Algoritması herhangi bir sınıflandırma yöntemiyle çalışabilmekte ve varsayılan olarak Rastgele Orman sınıflandırma yöntemini kullanmaktadır. Algoritma, orijinal özelliklerin önemini rastgele elde ettiği bir eşik değeri ile karşılaştırmakta ve ilgisiz özellikleri aşamalı bir şekilde ortadan kaldırmaktadır. Boruta Algoritması, gerek temel aldığı sınıflama yöntemlerine gerekse kullandığı önem kaynağına göre farklılaşan ve getImp fonksiyonu ile çağrılarak kullanılan birçok farklı yönteme sahiptir. Bu yöntemlerden aşağıda listelenen yedi tanesi çalışmamızda kullanılmıştır (14).

1. **ExtraZ Yöntemi:** Rastgeleliğin artırıldığı bu yöntemde normalleştirilmiş (Z-skoru ile standartlaştırılmış) permütasyon önemi kullanılmaktadır.
2. **ExtraRaw Yöntemi:** Ham permütasyon önemini kullanmaktadır.
3. **ExtraGini Yöntemi:** Gini safsızlık önemini üretmektedir.
4. **LegacyRfZ Yöntemi:** Normalleştirilmiş (Z-skoru ile standartlaştırılmış) permütasyon önemini ve Random Forest Algoritmasını kullanmaktadır.
5. **RfZ Yöntemi:** Normalleştirilmiş (Z-skoru ile standartlaştırılmış) permütasyon önemini kullanmaktadır.
6. **Ferns Yöntemi:** Rastgele Ferns önem hesaplamasında öncelikle derinlik parametresinin değeri optimize edilmekte ve önemin yakınsaması için gereken topluluktaki ferns sayısı Rastgele Orman durumundaki ağaç sayısından daha yüksek olmaktadır. Bu nedenle, Ferns Yöntemi Rastgele Orman yönteminden daha hızlı bir kullanıma sahiptir.
7. **Xgboost Yöntemi:** Uygulamada, bu işlevsellik XgBoost Yöntemine dayanarak Boruta Algoritmasını minimum optimal yönteme dönüştürmektedir (14).

Boruta Algoritmasının işleyişi aşağıda sıralı bir şekilde açıklanmaktadır.

- Gölge özellikler üretmek için tüm özelliklerin karıştırılmış kopyaları oluşturularak bilgi sistemine rastgelelik eklenmektedir (orijinal kümedeki özelliklerin sayısı 5'ten az olsa bile bilgi sistemi her zaman en az 5 gölge özellik ile genişletilmektedir).
- Genişletilmiş bilgi sistemi üzerinde bir Rastgele Orman sınıflandırıcısı çalıştırılmakta ve hesaplanan Z Skorları toplanmaktadır.
- Gerçek bir özelliğin gölge özelliklerin en iyisinden daha yüksek bir öneme sahip olup olmadığını kontrol etmek için gölge özellikler arasında maksimum Z Skoru (MZSA) bulunmakta ve ardından MZSA'dan daha iyi puan alan her özelliğe bir isabet atanmaktadır.
- Önemi belirlenmemiş her bir özellik için MZSA ile iki taraflı bir eşitlik testi gerçekleştirilmektedir.
- MZSA'dan önemli ölçüde daha düşük öneme sahip olan özellikler 'önemsiz' olarak kabul edilmekte ve bilgi sisteminden kalıcı olarak çıkarılmaktadır.

- MZSA'dan önemli ölçüde daha yüksek öneme sahip olan özellikler 'önemli' olarak kabul edilmektedir.
- Tüm gölge özellikler kaldırılmaktadır.

Tüm özellikler için önem ataması yapılana veya algoritma önceden belirlenen Rastgele Orman çalıştırma sınırına ulaşana kadar yukarıdaki prosedür tekrarlanmaktadır (2).

2.2.2. Elastik Ağ Algoritması

2015 yılında Determan en önemli özelliklerin tanımlanması, sınıflandırılması ve omik veri kümelerinin analiz edilmesi için bir algoritma önermiştir. Algoritma, OmicsMarkeR R paketinde (15) yer almaktadır. Bu çalışmada, Determan'ın Elastik Ağ Genelleştirilmiş Doğrusal Modeller İle Optimal Gen Seçim Algoritması kullanılmaktadır (4).

Elastik Ağ (EA), güçlü bir şekilde ilişkili tahmin edicilerin birlikte modele girme veya modelden çıkma eğiliminde olduğu bir gruplandırma etkisini teşvik etmektedir. Bu yöntem, özellikle tahmin edicilerin sayısının (p) gözlem sayısından (n) çok daha büyük olduğu durumlarda daha verimli çalışmaktadır (16).

EA, düzenli hale getirme sürecindeki ceza terimlerinin ayarlanmasında esnek olan ve zaman verimliliği yüksek yöntemlerden bir tanesidir. Çok katmanlı özellik seçim uygulamalarında, EA düzenli hale getirme sürecindeki ceza terimlerini ayarlama esnekliği ve zaman verimliliği nedeniyle eğitim modeli olarak seçilebilmektedir. EA kullanmanın avantajları arasında, özellik ağırlıklarında seyrekliği sağlayabilmesi ve aynı zamanda yüksek korelasyonlu özelliklerin gruplama etkisini desteklemesi yer almaktadır (17).

Diğer taraftan, cezalandırılmış lojistik regresyon (penalized logistic regression) modelleri, bağımsız değişkenler arasındaki yüksek korelasyon sorununun üstesinden gelmek için önerilmektedir. Cezalandırılmış lojistik regresyon modelleri Ridge, Lasso, Elastik Ağ (Ridge ve Lasso karışımı) lojistik regresyon modellerini içermektedir. Bu modellerde temel hedef, korelasyonlu tahmin edicilerin katsayılarını küçültmektir.

Karma parametre (mixing parameter) 0'a sabitlendiğinde, model "Ridge Lojistik Regresyon" olarak, karma parametre 1'e sabitlendiğinde ise model "Lasso Lojistik Regresyon" olarak adlandırılmaktadır. İlaveten, karma parametre 0 ile 1 arasında ise model "Elastik Ağ Lojistik Regresyonu" olarak adlandırılmaktadır. Bu çalışmada, karma parametre 0,5 olarak sabitlenmektedir (18).

Genel olarak, lojistik regresyon dâhil olmak üzere doğrusal modeller için özellik seçimi algoritmaları bir dizi kritere göre bazı özellik ağırlıklarını veya parametre tahminlerini sıfıra ayarlayarak açık bir şekilde özellik seçimi gerçekleştirebilmektedir. Öte yandan, bir algoritma bunun yerine özellik ağırlıklarının sıfıra doğru küçültüldüğü ancak asla tam olarak sıfır yapılmadığı (Ridge Cezası) veya en azından bazı ağırlıkların tam olarak sıfır yapılmasına izin veren bir küçültme işlemi yoluyla değişken seçimi (Lasso) gerçekleştirebilmektedir. Elastik Ağ, Ridge ve Lasso Cezaları arasında bir uzlaşmayı temsil etmekte olup Elastik Ağ Cezası basitçe bu iki cezanın doğrusal bir kombinasyonu olarak yazılmaktadır. Lasso ceza terimi, ortaya çıkan modelin parametre tahminlerinde seyrekliği teşvik etmek için hareket ederken, Ridge terimi bir gruplama etkisi uygulayan korelasyonlu özelliklerin parametre tahminlerinin ortalamasını almak için hareket etmektedir. Dolayısıyla, EA hem küçültme hem de otomatik özellik seçimi gerçekleştirmektedir. Kullanıcının tercihlerine ve problemin özelliklerine bağlı olarak Elastik Ağ Cezası, Lasso veya Ridge Cezalarına daha fazla ağırlık verecek şekilde ayarlanabilmektedir (19).

Lasso Yöntemine benzer şekilde, EA aynı anda otomatik değişken seçimi yapmakta ve birbiriyle ilişkili değişken gruplarını seçebilmektedir. Simülasyon çalışmaları ve gerçek veri örnekleri, Elastik Ağ Yönteminin tahmin doğruluğu açısından genellikle Lasso Yönteminden daha iyi performans gösterdiğini ortaya koymaktadır (16). Diğer taraftan, veri sayısı arttıkça Elastik Ağ tahmincisinin sadece tahmin için değil, aynı zamanda özellik (değişken) seçimi için de tutarlılığı artmaktadır (20).

2.3. Model Performans Ölçüleri

Boruta Algoritmasının kendi içerisinde bulunan, sınıflandırma yöntemlerine ve önemlilik ölçütlerine göre farklılaşan yöntemlerden yedi tanesine ait algoritmalar, RNA dizileme veri setlerine ayrı ayrı olarak uygulanmış ve elde edilen sonuçlar sadece Elastik Ağ Algoritması ile değil aynı zamanda birbirleri ile de kıyaslanmıştır. Bu çalışmada, Boruta Yöntemleri ile Elastik Ağ Algoritmasının gen seçim performansları karşılaştırılmıştır.

Özellik seçimi yöntemlerinin performansları karşılaştırılmak istenildiğinde dikkate alınabilecek belirli ölçüler bulunmaktadır. Bu çalışmada kullanılan modellerin gen seçim performansları, Tablo 2.1.'de yer alan karışıklık matrisi (confusion matrix) kullanılarak elde edilmiştir. Karışıklık matrisleri bir sınıflandırma problemindeki tahmin sonuçlarının özetini bir tablo formatında sunmakta ve model performansları bu tabloda yer alan sayılar üzerinden hesaplanmaktadır. Model performansları; özellikle Pozitif Kestirim Değeri, Duyarlılık ve F1 Ölçüsü sonuçları üzerinden detaylı bir şekilde değerlendirilmiştir. Ayrıca; Seçicilik (Specificity), Doğruluk (Accuracy), Kappa, Matthews Korelasyon Katsayısı (MCC), Dengeli Doğruluk (Balanced Accuracy) ve Youden İndeksi sonuçları hakkında da kısa bir değerlendirme yapılmıştır.

Tablo 2.1. Özellik seçiminde kullanılan karışıklık matrisi.

Tahmin Edilen Gen Sayıları	**Gerçek Durumdaki Gen Sayıları**		
	Önemli Gen Sayısı	Önemli Olmayan Gen Sayısı	**Toplam**
Önemli Gen Sayısı	Doğru Pozitif (DP)	Yanlış Pozitif (YP)	**DP + YP**
Önemli Olmayan Gen Sayısı	Yanlış Negatif (YN)	Doğru Negatif (DN)	**YN + DN**
Toplam	**DP + YN**	**YP + DN**	**n**

Tablo 2.1.'de yer alan karışıklık matrisi normalde bilinen karışıklık matrislerinden farklı bir yapıya sahiptir. Normalde sıklıkla kullanılan karışıklık

matrisleri, gözlemlerin sınıflanması üzerine oluşturulmaktadır. Ancak burada kişiler değil değişkenler sınıflandırılmaktadır. Diğer bir ifadeyle, Tablo 2.1.'deki karışıklık matrisi kişiler üzerinden hasta-sağlıklı gibi bir sınıflandırma yapılması üzerine değil, değişkenlerin doğru bulunması üzerine oluşturulmaktadır. Buradan elde edilen performans ölçüleri önemli genlerin bulunmasını esas almaktadır.

Pozitif Kestirim Değeri (Precision): Doğru sınıflandırılan pozitif örnek sayısı, pozitif olduğu tahmin edilen toplam örnek sayısına oranlanarak hesaplanmaktadır. Sağlık alanında, Pozitif Kestirim Değeri (Positive Predicted Value - PPV) şeklinde adlandırılmaktadır. Bu performans ölçütü, Eşitlik 2.21. ile elde edilmektedir (21).

$$\text{Pozitif Kestirim Değeri} = \frac{\text{DP}}{\text{DP} + \text{YP}} \tag{2.21.}$$

Duyarlılık (Recall/Sensitivity): Doğru sınıflandırılan pozitif örnek sayısı, toplam pozitif örnek sayısına oranlanarak bulunmaktadır. Sınıf sayısı ikiden fazla olduğu durumda Duyarlılığın hesaplanmasında kullanılan formül, Eşitlik 2.22.'de yer almaktadır (22).

$$\text{Duyarlılık} = \frac{\text{DP}}{\text{DP} + \text{YN}} \tag{2.22.}$$

Seçicilik (Specificity): Gerçekte negatif sınıfta bulunan örnekler arasında tahmin edilen sınıfı negatif olan örneklerin oranını göstermektedir. Sınıflama yöntemi tarafından negatif değerlere sahip örneklerin belirlenmesindeki performansıdır ve Eşitlik 2.23. ile hesaplanmaktadır (23).

$$\text{Seçicilik} = \frac{\text{DN}}{\text{DN} + \text{YP}} \tag{2.23.}$$

F1 Ölçüsü (F1 Measure): Bu performans ölçüsü, Duyarlılık ve Pozitif Kestirim Değerinin harmonik ortalaması hesaplanarak bulunmaktadır. F1 Ölçüsünün sıklıkla tercih edilmesinin nedenlerinden birisi; Duyarlılık ve Pozitif Kestirim Değeri gibi iki önemli performans ölçüsünün birlikte değerlendirilmesidir. Özellikle bağımlı

değişkenin değerlerinin dengesiz bir şekilde dağıldığı durumda bu ölçütün kullanılması önerilmektedir. F1 Ölçüsü, Eşitlik 2.24. ile hesaplanmaktadır (22).

$$F1 = \frac{2}{\left(\frac{1}{\text{Duyarlılık}} + \frac{1}{\text{Kesinlik}}\right)} \tag{2.24.}$$

Doğruluk (Accuracy): Sınıflandırıcının etkinliğinin genel bir ölçüsü olan Doğruluk, doğru sınıflandırılan örnek sayısının toplam örnek sayısına oranını göstermektedir. Doğruluk ölçütünün hesaplanmasında kullanılan formül, Eşitlik 2.25.'de yer almaktadır (22).

$$\text{Doğruluk} = \frac{DP + DN}{n} \tag{2.25.}$$

Kappa: 1960'lı yıllarda Kappa istatistiği (K), Cohen tarafından psikolojik davranış gözlemcilerinin arasındaki uyumun bir göstergesi olarak tanıtılmıştır. Farklı yöntemlerin sınıflama performanslarının karşılaştırılmasında da kullanılan Kappa istatistiğinin, aynı olayı gözlemleyen iki ya da daha çok gözlemci arasındaki uyum seviyesinin ölçülmesinde de kullanılmaktadır. Bu alanda Kappa istatistiği, ilgilenilen yöntem ile ilgili sınıflama modelinin tahmin değerleri ile gerçek değerler arasındaki uyum seviyesinin ölçülmesi için kullanılmaktadır. İki durumlu olaylar için hesaplanan Kappa istatistiği, Eşitlik 2.26. ile elde edilmektedir (24).

$$K = \frac{P_a - P_c}{1 - P_c} \tag{2.26.}$$

P_c şansa bağlı beklenen uyum oranını, P_a ise uyum oranını göstermektedir. Tablo 2.1. yardımıyla P_a ve P_c oranları, Eşitlik 2.27. ve Eşitlik 2.28. ile hesaplanmaktadır.

$$P_a = \frac{DP + DN}{n} \tag{2.27.}$$

$$P_c = \frac{(DP + YP)(DP + YN) + (YN + DN)(YP + DN)}{n^2} \tag{2.28.}$$

P_c'nin hesaplanabilmesi için, ikiden fazla durum olduğunda ilk olarak her durum için şansa bağlı olarak beklenen uyum sıklıkları hesaplanarak toplanmaktadır. Şansa bağlı olarak beklenen uyum sıklığı toplam gözlem sayısına bölünerek P_c hesaplanmaktadır. Kappa katsayısı -1 ile +1 değerleri arasında değişmektedir (22, 25).

Matthews Korelasyon Katsayısı (Matthews Correlation Coefficient - MCC): Dengesiz veri seti sorunundan etkilenmeyen alternatif bir ölçüt olarak gerçek ve tahmin edilen değerler arasındaki Pearson Momentler Çarpımı Korelasyon Katsayısını hesaplamaktadır. MCC, yalnızca ikili tahmin edicinin pozitif ve negatif veri örneklerinin çoğunu doğru bir şekilde tahmin edebilmesi durumunda yüksek bir puan üreten tek ikili sınıflandırma oranıdır ve Eşitlik 2.29. ile hesaplanmaktadır (26).

$$\mathrm{MCC} = \frac{\mathrm{DP\ x\ DN} - \mathrm{YP\ x\ YN}}{\sqrt{(\mathrm{DP} + \mathrm{YP})\ \mathrm{x}\ (\mathrm{YN} + \mathrm{DN})\ \mathrm{x}\ (\mathrm{DP} + \mathrm{YN})\ \mathrm{x}\ (\mathrm{YP} + \mathrm{DN})}} \qquad (2.29.)$$

Dengeli Doğruluk (Balanced Accuracy - BA): Bu performans ölçüsü, Duyarlılık ve Seçiciliğin aritmetik ortalaması alınarak hesaplanmaktadır. Dengeli Sınıflandırma Oranı olarak da adlandırılmaktadır. Dengeli Doğruluğun hesaplanmasında kullanılan formül Eşitlik 2.30.'da yer almaktadır (27).

$$\mathrm{BA} = \frac{\frac{\mathrm{DP}}{\mathrm{DP+YN}} + \frac{\mathrm{DN}}{\mathrm{DN+YP}}}{2} \qquad (2.30.)$$

Youden İndeksi: Duyarlılık ve Seçiciliğin bir fonksiyonu olan Youden İndeksi (J), genel teşhis etkinliğinin yaygın olarak kullanılan bir ölçüsüdür. Duyarlılık ve Seçiciliğe eşit ağırlık verildiğinde Youden İndeksi, biyobelirtecin (biomarker) ayırt etme yeteneğini optimize eden kesim noktasında ortaya çıkmaktadır. Bu indeks, Eşitlik 2.31. ile elde edilmektedir (28).

$$\mathrm{J} = \frac{\mathrm{DP}}{\mathrm{DP} + \mathrm{YN}} + \frac{\mathrm{DN}}{\mathrm{DN+YP}} - 1 \qquad (2.31.)$$

Karışıklık matrisi üzerinden hesaplanan model performans ölçütlerinin daha kolay anlaşılabilmesi için bu çalışmanın bir bölümünden bir kesit alınarak örnek sonuçlar ile hazırlanan karışıklık matrisi Tablo 2.2.'de yer almaktadır. Bu matris

üzerinden algoritmaların gen seçim performanslarını gösteren Pozitif Kestirim Değeri, Duyarlılık ve F1 Ölçüsü olmak üzere belirli model performans ölçütlerinin değerleri hesaplanmış ve sonuçlar yorumlanmıştır.

Tablo 2.2. Özellik seçiminde kullanılan karışıklık matrisi (örnek).

Tahmin Edilen Gen Sayıları	**Gerçek Durumdaki Gen Sayıları**		
	Önemli Gen Sayısı	Önemli Olmayan Gen Sayısı	**Toplam**
Önemli Gen Sayısı	7	2	**9**
Önemli Olmayan Gen Sayısı	3	188	**191**
Toplam	**10**	**190**	**200**

Tablo 2.2.'de yer alan ilk sütun toplamı olan 10 sayısı; benzetim çalışması ile anlamlı olarak üretilen toplamda 10 tane gen olduğunu göstermektedir. İkinci sütun toplamı olan 190 sayısı ise anlamsız olarak üretilen toplamda 190 tane gen olduğunu göstermektedir. Sütunlarda, üretilen genlerden gerçekte kaç tanesinin anlamlı olduğu biliniyor çünkü başlangıçta 10 gen anlamlı olacak şekilde üretilmiş, dolayısıyla 200 genden geriye kalan 190 gen de anlamsız olacak şekilde üretilmiştir. Satırlarda ise üretilen genlerden kaç tanesinin anlamlı olduğu görülmektedir. İlk sütundaki sayılar, en az 2 katmanda 7 genin anlamlı ve 3 genin anlamsız olduğunu göstermektedir. İlk satırdaki sayılar ise bu yöntemle 9 genin seçildiğini ancak seçilen bu 9 genin başlangıçta anlamlı olarak üretilen 10 genden 7'sini içerdiğini göstermektedir. Söz konusu karışıklık matrisinde yer alan sayılar kullanılarak elde edilen performans ölçüleri, kişileri sınıflandırma performans ölçülerini değil değişkeni doğru bulma performans ölçülerini göstermektedir.

Tablo 2.2.'de yer alan sayılar kullanılarak hesaplanan Pozitif Kestirim Değeri, Duyarlılık Oranı ve F1 Ölçüsü aşağıda yer almaktadır.

Duyarlılık = DP / (DP + YN)

= 7 / (7 + 3)

= 0,70

Duyarlılık, gerçekte anlamlı olan genlerin kaç tanesini modelin anlamlı bulduğunu göstermektedir. Diğer bir ifadeyle, model önemli genlerin %70'ini algılayabilmekte iken diğer %30'unu kaçırabilmektedir.

Pozitif Kestirim Değeri = DP / (DP + YP)

= 7 / (7 + 2)

= 0,78

Pozitif Kestirim Değeri, modelin anlamlı dediği genlerin gerçekte kaç tanesinin anlamlı olduğunu göstermektedir. Başka bir değişle, algoritma tarafından önemli gen olarak tanımlanan tüm genlerin %78'i gerçekten önemli olarak ifade edilmektedir.

F1 Ölçüsü = 2 / ((1/Duyarlılık + 1/Pozitif Kestirim Değeri))

= 2 / ((1/0,70 + 1/0,78))

= 0,74

Vaka-kontrol grupları arasındaki dağılım dengesiz olduğu durumlarda çoğunlukla Pozitif Kestirim Değeri ve Duyarlılık ölçütlerinin yanı sıra, bu iki performans ölçütünün harmonik ortalaması olan F1 Ölçüsünün incelenmesi de yarar sağlamaktadır.

2.4. K-katlı Çapraz Doğrulama

K-katlı Çapraz Doğrulama (K-fold Cross Validation - KCV) tekniği, model seçimi ve sınıflandırıcıların hata tahmini için sıklıkla kullanılan yaklaşımlardan birisidir. KCV, bir veri setini k sayıda alt kümeye bölmektedir. Yinelemeli olarak bunlardan bazıları modeli öğrenmek için kullanılmakta, diğerleri ise modelin performansını değerlendirmek için kullanılmaktadır (29).

3. GEREÇ VE YÖNTEM

3.1. Veri Setleri

Bu çalışmada böbrek, kolon, karaciğer, akciğer, prostat ve tiroid kanseri olmak üzere altı farklı kanser türüne ait, genetik verilerin bulunduğu TCGA veri tabanından elde edilen sekiz farklı gen veri seti kullanılmıştır. Veri setlerine Kanser Genom Atlası veri portalı üzerinden ulaşılabilmektedir.

Kanser Genom Atlası Projesi 2005 yılında başlatılmıştır. Günümüzde kullanılan akıllı kanser ilaçları gibi yenilikçi tedavilerin çıkış noktası bu projedir. 33 farklı kanser türünü ve 20 bini aşkın tümör çeşidini moleküler olarak tanımlayan bir genomik projedir. Kanserin tanı ve tedavisine yönelik çalışmalara imkân sağlayan proje kapsamındaki veriler araştırmacıların kullanımı için erişime açıktır (30).

R programında TCGAbiolinks R/Bioconductor paketi (31), ham gen sayılarını veri tabanından çekmek için kullanılmaktadır. Her bir gen veri seti 19.947 farklı genden oluşmaktadır. Bu veri setlerinin her biri kanser ve kontrol grubunu içeren örneklerden oluşmaktadır. Kontrollerden çok daha fazla vaka olduğu için oldukça dengesiz bir yapıya sahiptirler (4). Veri setleri ile ilgili genel bilgiler Tablo 3.1.'de yer almaktadır.

Kanser türleri arasında önemli genlerin bulunması için yeni nesil RNA dizileme veri setlerinin analizinde başta Boruta ve OmicsMarkeR paketleri olmak üzere; caret, genefilter, DESeq2 ve doParallel gibi birçok paket R programında kullanılarak analizler yapılmıştır. Bu analizlerde, çeşitli sınıflandırma yöntemlerine ve önemlilik ölçütlerine göre farklılaşan, Boruta Algoritması kapsamındaki yedi farklı yöntem ve Elastik Ağ yöntemi de dâhil olmak üzere toplamda sekiz farklı yöntem kullanılmıştır.

Tablo 3.1. Çalışmada kullanılan veri setlerinin özellikleri.

Veri Setleri	Proje Kimlikleri	n	Sınıf Boyutları	Sınıf Oranları
Böbrek kromofobisi	TCGA-KICH	91	66 / 25	2,64
Böbrek renal şeffaf hücreli karsinom	TCGA-KIRC	605	533 / 72	7,40
Kolon adenokarsinomu	TCGA-COAD	326	285 / 41	6,95
Karaciğer hepatoselüler karsinomu	TCGA-LIHC	421	371 / 50	7,42
Akciğer adenokarsinomu	TCGA-LUAD	574	515 / 59	8,73
Akciğer yassı hücreli karsinomu	TCGA-LUSC	553	502 / 51	9,84
Prostat adenokarsinomu	TCGA-PRAD	549	497 / 52	9,56
Tiroid karsinomu	TCGA-THCA	564	505 / 59	8,56

RNA dizileme veri setlerinde, değişken sayısı çok fazla iken (her bir veri setinde 19.947 gen) birey sayısı ise oldukça düşüktür. Bu çalışmada kullanılan sekiz veri setinde birey sayıları 91 ile 605 arasında değişmektedir. Örneğin; Tablo 3.1.'in ilk satırında yer alan veriler, 19.947 farklı gen hakkında bilgiler içeren Böbrek Kromofobisi veri setinin 91 kişiye ait olduğunu ve bu kişilerin 66 tanesinde vaka görüldüğünü, 25 tanesinde ise vaka görülmediğini (kontrol grubu) göstermektedir. Bu veri setinde kontrollerden çok daha fazla vaka olduğu için oldukça dengesiz bir yapıda olduğu anlaşılmaktadır. Benzer şekilde, çalışmada kullanılan diğer yedi veri setinin her birinin de dengesiz bir yapıda olduğu görülmektedir.

3.2. Benzetim Çalışması

Makine öğrenimi algoritmalarının, gerçekte DEG olarak üretilen genleri seçme kapasitesini ortaya çıkarmak için bir Monte Carlo benzetim çalışması yapılmıştır. Bu amaçla benzetim veri setleri, Tablo 3.1.'de gösterilen sekiz gerçek veri setine dayalı olarak ssizeRNA R paketi (32) kullanılarak Negatif Binom Dağılımından üretilmiştir. Negatif Binom Dağılımı, kesikli bir yapıda olan RNA dizileme ham sayım verilerinin doğasının aşırı dağılımını yakalayabildiği için güçlü bir yapıya sahiptir (33). Veri setlerinin her biri için gen sayısı 19.947 olarak alınmıştır. Üretilen genler arasında ilk 10 gen, sim.counts fonksiyonu kullanılarak gruplar (vaka-kontrol) arasında önemli ölçüde farklı olacak şekilde üretilmiştir. Üretilen her veri seti için gerçek veri setlerinin

sınıf oranlarına dayalı olarak kanser ve kontrolü içeren bir ikili yanıt değişkeni oluşturulmuştur. Gerçek veri setlerine bağlı olarak gözlem sayısı dikkate alınmıştır. Ortalama vektör, her gen için aritmetik ortalama olarak elde edilmiştir.

Sekiz farklı RNA dizileme veri setine dayalı olarak toplam sekiz senaryo üzerinden çalışmalar yürütülmüştür. Benzetim senaryoları 1.000 kez tekrarlanmış ve tekrarların ortalaması alınarak performans ölçüleri elde edilmiştir. Benzetim çalışması ile üretilecek verilerin, gerçek veri setlerinin özelliklerine dayalı olması için Parametrik Bootstrap yöntemi kullanılmıştır.

3.3. Ön İşleme

Ham gen sayımlarını içeren RNA dizileme veri setlerinin, analiz edilmeden önce bazı ön işleme adımlarından geçirilmesi gerekmektedir (4). Bu çalışmada filtreleme, normalleştirme, dönüşüm ve tek değişkenli analiz yöntemleri kullanılarak ön işleme yapılmıştır.

3.3.1. Filtreleme

İlk olarak, caret paketi (34) kullanılarak sıfıra yakın varyans filtrelemesi uygulanmıştır. Bu işlem, R programında preProcess fonksiyonu kullanılarak yapılmıştır. Böylece, düşük miktarda gen dışlanmıştır (4).

3.3.2. Normalleştirme

Normalleştirme (normalizasyon), verilerin aralığının küçültülmesi veya ölçeklendirilmesi gibi verilerin standartlaştırılmasını ifade etmektedir. Min-maks normalleştirme, Z-skoru ile normalleştirme, ondalık ölçeklendirme ile normalleştirme ve Medyan Oran Normalleştirme (Median Ratio Normalization - MRN) gibi pek çok normalleştirme tekniği bulunmaktadır. MRN, transkriptomların göreli boyutundan kaynaklanan yanlılığa karşı tutarlı ve sağlam (robust) bir yapıya sahiptir (35, 36).

Bu kapsamda bir önceki filtreleme aşamasında filtrelenen ham sayım verilerine DESeq2 paketi (37) kullanılarak Medyan Oran Normalizasyonu uygulanmıştır.

3.3.3. Dönüşüm

Daha az çarpık bir dağılıma sahip olmak için normalleştirilmiş verilere Logaritmik Dönüşüm uygulanmıştır (4). Bu işlemin uygulanması için R programına özel bir paket yüklenmesine ihtiyaç bulunmamaktadır. Çünkü bu çalışmada kullanılan Logaritmik Dönüşüm işlemini gerçekleştiren log2 fonksiyonu, varsayılan olarak yüklü bulunan base paketinde yer almaktadır.

3.3.4. Tek Değişkenli Analiz

Tek değişkenli analiz, gen sayısını azaltmak ve koşullar arasında önemli ölçüde farklılık gösteren genleri belirlemek için kullanılmıştır. Student t testi sonuçlarına göre genler en düşük p-değerinden en yükseğe doğru sıralanmış ve ilk 200 gen seçilmiştir. Tek değişkenli analiz, genfilter paketindeki colttests işlevi (38) kullanılarak gerçekleştirilmiştir (4).

Burada oldukça önemli olan bir husus, benzetim çalışması yapılarak ilk 10 genin her veri seti için anlamlı olarak üretilmesidir. Söz konusu 10 geni elemeyecek filtrelerin kullanılmasına özen gösterilmiştir. Diğer bir ifadeyle; her bir veri setinde bulunan ve 19.947 olan gen sayısı, yapılan bu filtrelemeler sonucunda 200'e düşürüldüğünde anılan 10 genin de bu 200 gen içerisinde yer aldığı görülmüştür. Her bir veri setindeki toplam gen sayısı, 200'e düşürülerek ön işleme adımları tamamlanmıştır. Ön işleme adımları sonrasında elde edilen yeni veri setleri kullanılarak Boruta ve Elastik Ağ Algoritmalarının, anlamlı olarak üretilen bahse konu 10 genden kaç tanesini önerdiği incelenmiştir.

Düzenlenen algoritmalar ile yapılan işlemleri hızlandırmak amacıyla doParallel paketi kullanılarak paralel hesaplama yönteminden yararlanılmıştır. Bu yöntem sayesinde aynı anda beş çekirdekte beş farklı hesaplama yapılabilmiştir. Analizler 64 GB bellek, 12 çekirdek ve 2,4 GHz işlemciye sahip bir bilgisayar üzerinde paralel hesaplama yöntemi kullanılarak yapılmasına rağmen, yalnızca bir senaryonun tamamlanması yaklaşık on gün sürmüştür. Diğer taraftan, 1.000 tekrarlı analizlerin sonuçları anılan bilgisayarda yaklaşık üç aylık bir çalışma sonucunda elde edilmiştir.

4. BULGULAR

Bu tez çalışmasında; Böbrek Kromofobisi (Kidney Chromophobe – KICH), Böbrek Renal Şeffaf Hücreli Karsinomu (Kidney Renal Clear Cell Carcinoma - KIRC), Kolon Adenokarsinomu (Colon Adenocarcinoma - COAD), Karaciğer Hepatoselüler Karsinomu (Liver Hepatocellular Carcinoma - LIHC), Akciğer Adenokarsinomu (Lung Adenocarcinoma - LUAD), Akciğer Yassı Hücreli Karsinomu (Lung Squamous Cell Carcinoma - LUSC), Prostat Adenokarsinomu (Prostate Adenocarcinoma - PRAD) ve Tiroid Karsinomu (Thyroid Carcinoma - THCA) olmak üzere sekiz farklı RNA dizileme veri seti kullanılmıştır. Bu veri setlerine, çeşitli sınıflandırma yöntemlerine ve önemlilik ölçütlerine göre farklılaşan Boruta Algoritması kapsamındaki yöntemlerden yedi tanesi ve Elastik Ağ yöntemi de dâhil olmak üzere toplamda sekiz farklı yöntem uygulanmıştır. Sekiz farklı veri setine uygulanan söz konusu yöntemlerin özellik seçimine ilişkin performans ölçüleri Tablo 4.1., 4.2., 4.3. ve 4.4.'de yer almaktadır.

Söz konusu sekiz farklı RNA dizileme veri setlerinin her birinde kontrol grubundan çok daha fazla vaka grubu olduğu için oldukça dengesiz bir yapıya sahiptirler. Kullanılan RNA dizileme veri setlerinde, değişken sayısı çok fazla iken (her bir veri setinde 19.947 gen) birey sayısı ise oldukça düşüktür. En az birey, Böbrek Kromofobisi veri setinde yer almakta olup 91 kişidir. Bu kişilerin 66 tanesinde vaka görülmekte iken 25 tanesinde ise vaka görülmemekte olup oldukça dengesiz bir yapıya sahiptir. Benzer şekilde diğer yedi veri setinde de vaka-kontrol arasında dengesiz bir dağılım bulunmaktadır. Diğer taraftan, 605 kişi ile en çok bireyin bulunduğu veri seti Böbrek renal şeffaf hücreli karsinomudur. Diğer altı veri setinde yer alan birey sayıları 91 ile 605 arasında değişmektedir.

Her bir veri setinde bulunan 19.947 adet genin özellik seçimi çalışmasında modellerin performansını olumsuz yönde etkileyebileceği düşünüldüğünden, özellik boyutunu azaltmak için filtreleme, normalleştirme, dönüşüm ve tek değişkenli analiz olmak üzere dört farklı ön işleme adımı uygulanmıştır. İlk olarak, R programında caret paketi kullanılarak sıfıra yakın varyans filtrelemesi uygulanmıştır. Böylece, düşük miktarda gen dışlanmıştır. İlk filtreleme aşamasında filtrelenen ham sayım verilerine

yine R programı üzerinden DESeq2 paketi kullanılarak Medyan Oran Normalizasyonu uygulanmıştır. Ardından daha az çarpık bir dağılıma sahip olmak için normalleştirilmiş verilere Logaritmik Dönüşüm uygulanmıştır. Son ön işleme adımında, R programında genfilter paketindeki colttests işlevi kullanılarak gen sayısını azaltmak ve koşullar arasında önemli ölçüde farklılık gösteren genleri belirlemek için tek değişkenli analiz işlemi gerçekleştirilmiştir. Burada, Student t testi sonuçlarına göre genler en düşük p-değerinden en yükseğe doğru sıralanmış ve ilk 200 gen seçilmiştir. Diğer taraftan, yapılan benzetim çalışması ile ilk 10 gen, her bir veri seti için anlamlı olarak üretilmiştir. Söz konusu 10 geni elemeyecek filtrelerin kullanılmasına özen gösterilmiştir. Diğer bir ifadeyle, her bir veri setinde 19.947 olan gen sayısı yapılan bu filtrelemeler sonucunda 200'e düşürüldüğünde anılan 10 genin de bu 200 gen içerisinde yer aldığı görülmüştür. Her bir veri setinden toplam gen sayısı 200'e düşürülerek ön işleme adımları tamamlanmıştır.

Veri setleri, boyut azaltma ve dönüşüm işlemlerinden sonra özellik seçimi yöntemlerinin uygulanmasına hazır hale gelmiştir. Tüm veri setleri üzerinde Boruta Algoritması ve Elastik Ağ Genelleştirilmiş Doğrusal Modeller İçerisinde Determan'ın Algoritması uygulanarak makine öğrenimi tabanlı gen seçimi gerçekleştirilmiştir. Ön işleme adımları sonrasında elde edilen yeni veri setleri kullanılarak Boruta ve Elastik Ağ Algoritmalarının, anlamlı olarak üretilen bahse konu 10 genden kaç tanesini önemli gen olarak bulacağına yönelik inceleme yapılmıştır. Modellerin gen seçim performanslarının iyileştirilmesi için beş katlı çapraz doğrulama yapılarak en az iki katmanda önemli bulunan genler dikkate alınmıştır. Algoritmaların gen seçim performanslarının değerlendirilmesinde Duyarlılık, Pozitif Kestirim Değeri ve F1 Ölçüsü sonuçları üzerinde durulmuştur.

Üretilen her veri seti için kanser veya kontrolü gösteren bir ikili sınıf değişkeni de üretilmiştir. Gerçek veri setlerindeki sınıf oranı, çoğu omik verisinde olduğu gibi oldukça dengesizdir. Bu nedenle, sınıf oranının oldukça dengesiz olduğu durumlarda bile genleri öneren makine öğrenimi algoritmaları seçilmiştir. Benzetim sırasında, kanser ve kontrol numuneleri arasında diferansiyel olarak ifade edilen bazı genler üretilmiştir. Farklı makine öğrenimi algoritmalarının gen seçim performansları,

önceden işlenmiş benzetim veri setleri üzerinde karşılaştırılmıştır. Algoritmaların performansı, gerçek ve tahmin edilen ile önemli ve önemli olmayan gen sayılarını içeren karışıklık matrislerine dayalı olarak araştırılmıştır. Benzetim senaryoları 1.000 kez tekrarlanmış ve tekrarların ortalaması alınarak performans ölçüleri elde edilmiştir. Tüm senaryolar kapsamında sekiz farklı veri setine Boruta Algoritmasının yanı sıra Elastik Ağ Algoritması da uygulanarak algoritmaların performansları; Tablo 4.1., Tablo 4.2. ve Tablo 4.3.'de Pozitif Kestirim Değeri, Duyarlılık ve F1 Ölçüsü ile özetlenmiştir. Ayrıca, sonuçların yorumlanmasında kolaylık sağlamasını teminen her bir veri seti için Duyarlılık Oranları ve Pozitif Kestirim Değerleri Şekil 4.1. – Şekil 4.8.'de gösterilmiştir. İlaveten, diğer performans ölçülerini içeren daha ayrıntılı sonuçlar Tablo 4.4.'de mevcut bulunmaktadır.

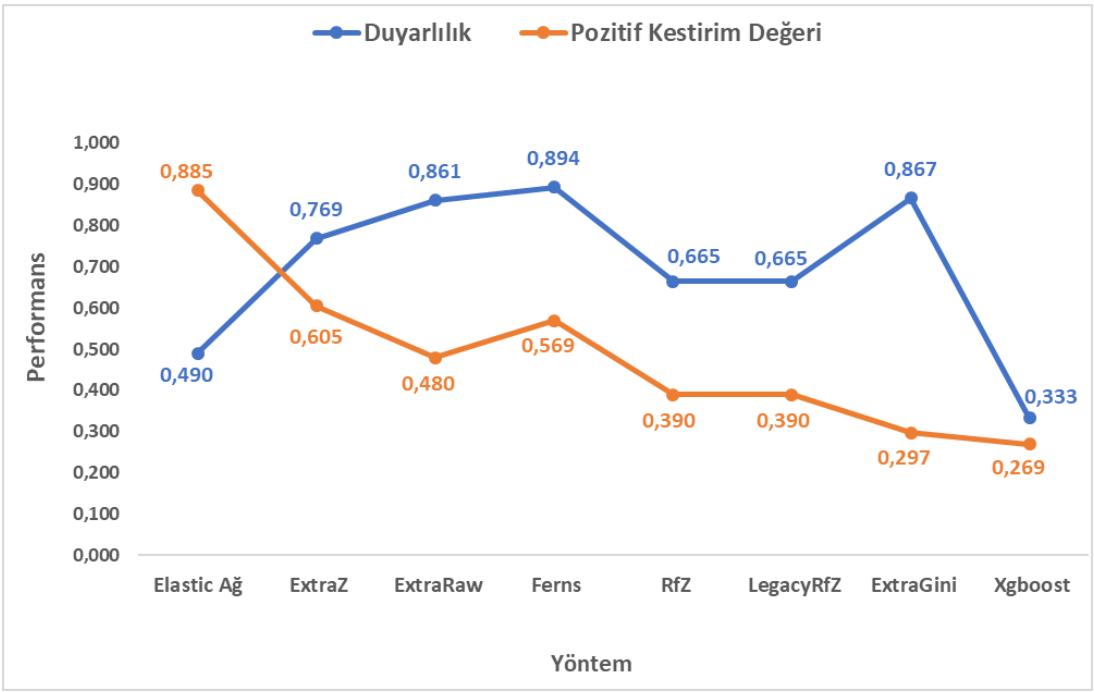
Tablo 4.1. Sekiz farklı RNA dizileme veri setinde, sekiz farklı yöntem çerçevesinde kurulan özellik seçimi modellerinin Duyarlılık Oranlarının karşılaştırılması.

Veri Seti	Elastic Ağ Algoritması	Boruta Algoritması						
		ExtraZ	ExtraRaw	Ferns	RfZ	LegacyRfZ	ExtraGini	Xgboost
KICH	0,490	0,769	0,861	**0,894**	0,665	0,665	0,867	0,333
KIRC	0,908	0,995	**0,996**	0,995	0,991	0,994	0,969	0,291
COAD	0,650	0,872	0,962	**0,992**	0,611	0,615	0,777	0,246
LIHC	0,745	0,958	0,992	**0,995**	0,799	0,804	0,831	0,252
LUAD	0,768	0,894	**0,913**	**0,913**	0,852	0,852	0,784	0,235
LUSC	0,682	0,833	0,918	**0,920**	0,637	0,644	0,605	0,194
PRAD	0,755	0,922	0,999	**1,000**	0,743	0,743	0,696	0,225
THCA	0,843	0,985	**1,000**	**1,000**	0,943	0,941	0,873	0,257

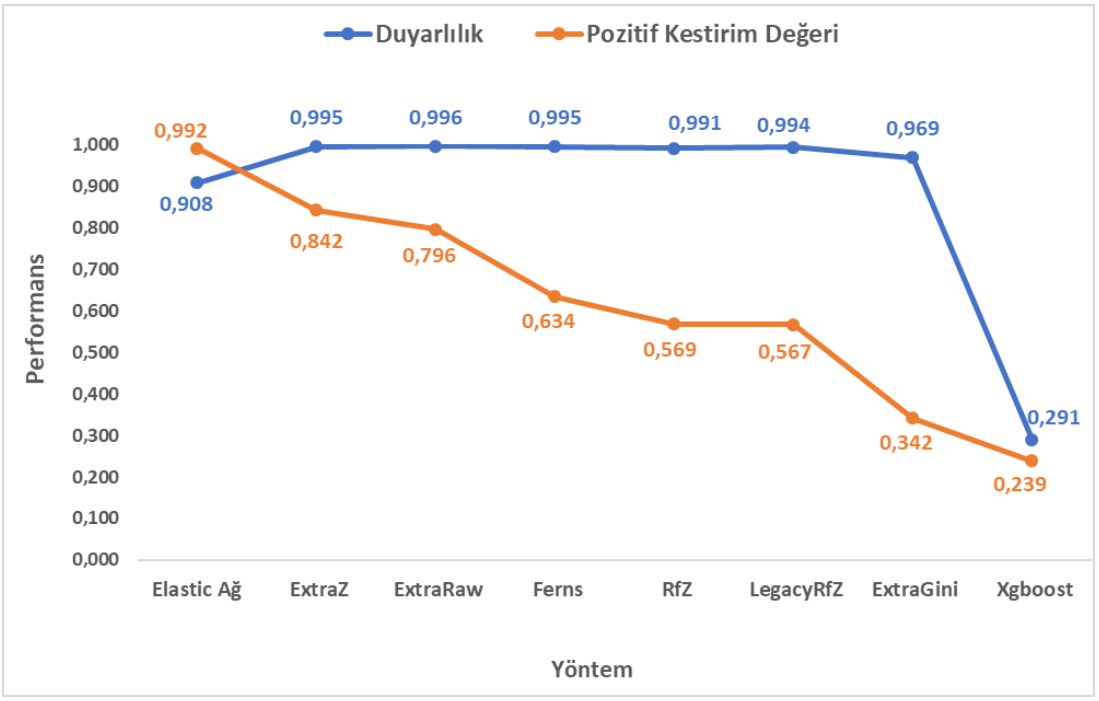
Tablo 4.2. Sekiz farklı RNA dizileme veri setinde, sekiz farklı yöntem çerçevesinde kurulan özellik seçimi modellerinin Pozitif Kestirim Değerlerinin karşılaştırılması.

Veri Seti	Elastic Ağ Algoritması	Boruta Algoritması						
		ExtraZ	ExtraRaw	Ferns	RfZ	LegacyRfZ	ExtraGini	Xgboost
KICH	**0,885**	0,605	0,480	0,569	0,390	0,390	0,297	0,269
KIRC	**0,992**	0,842	0,796	0,634	0,569	0,567	0,342	0,239
COAD	**0,938**	0,755	0,701	0,620	0,371	0,371	0,270	0,152
LIHC	**0,969**	0,805	0,747	0,626	0,458	0,459	0,293	0,171
LUAD	**0,982**	0,844	0,798	0,643	0,532	0,533	0,311	0,197
LUSC	**0,962**	0,833	0,802	0,663	0,424	0,428	0,255	0,156
PRAD	**0,964**	0,836	0,801	0,657	0,449	0,454	0,268	0,165
THCA	**0,981**	0,841	0,792	0,637	0,532	0,533	0,314	0,196

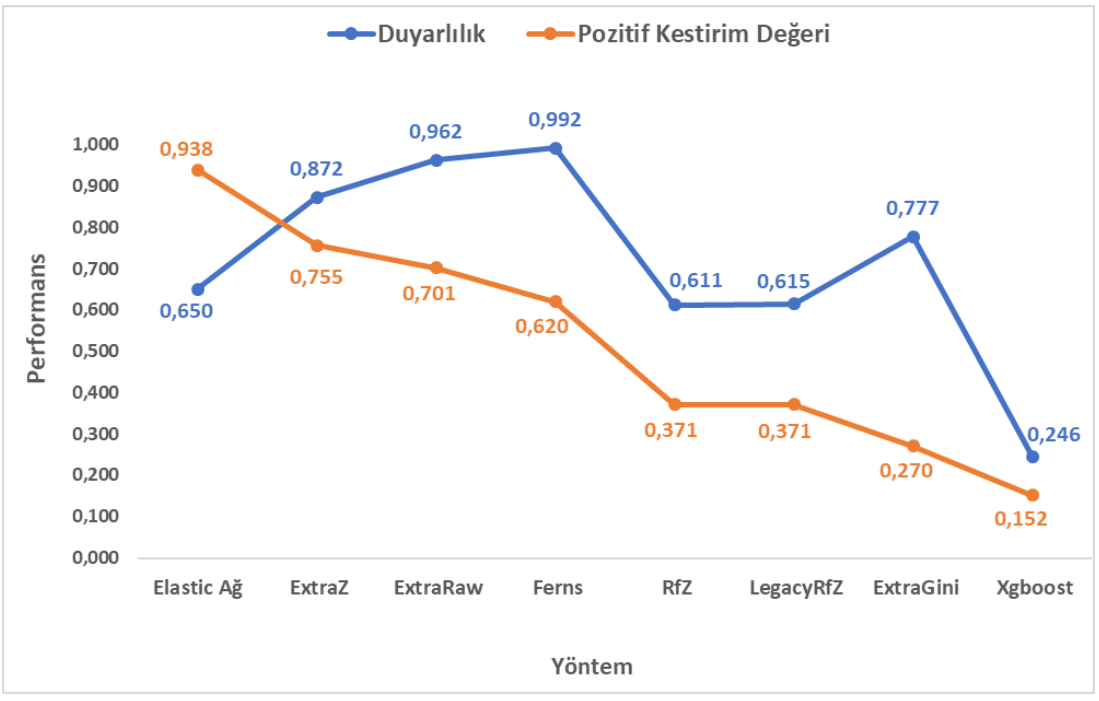
Sonuçların yorumlanmasında kolaylık sağlamasını teminen her bir veri seti için Duyarlılık Oranları ve Pozitif Kestirim Değerleri Şekil 4.1. – Şekil 4.8.'de grafiklerle gösterilmiştir.



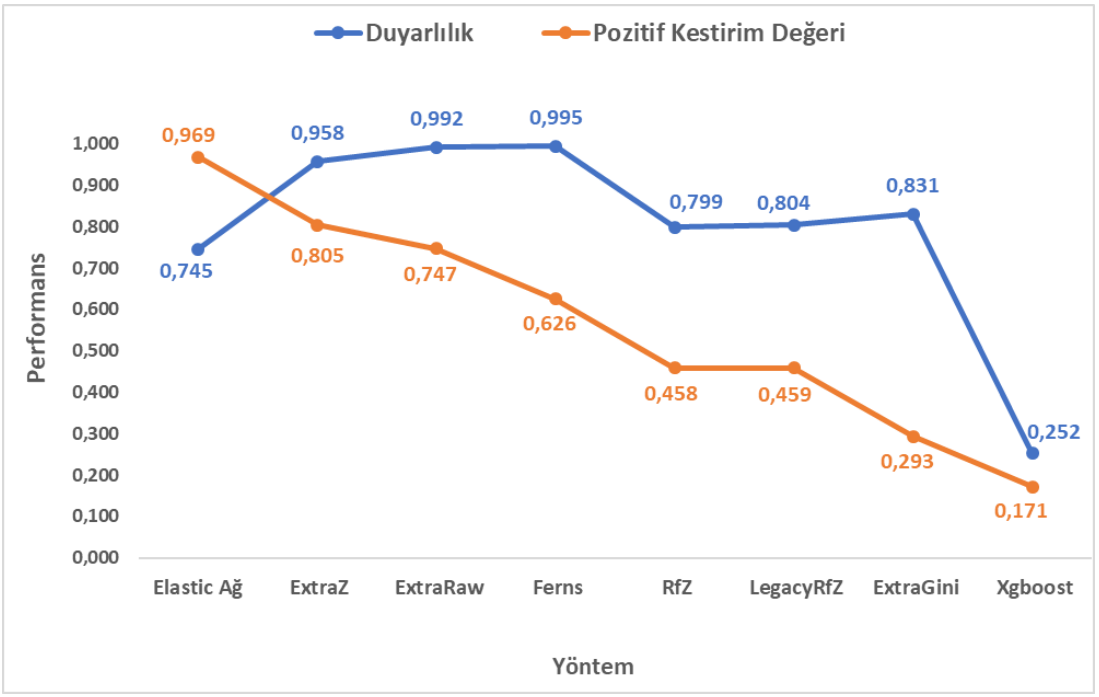
Şekil 4.1. Böbrek Kromofobisi veri seti üzerinden algoritmaların Duyarlılık Oranları ve Pozitif Kestirim Değerlerinin performanslarının karşılaştırılması.



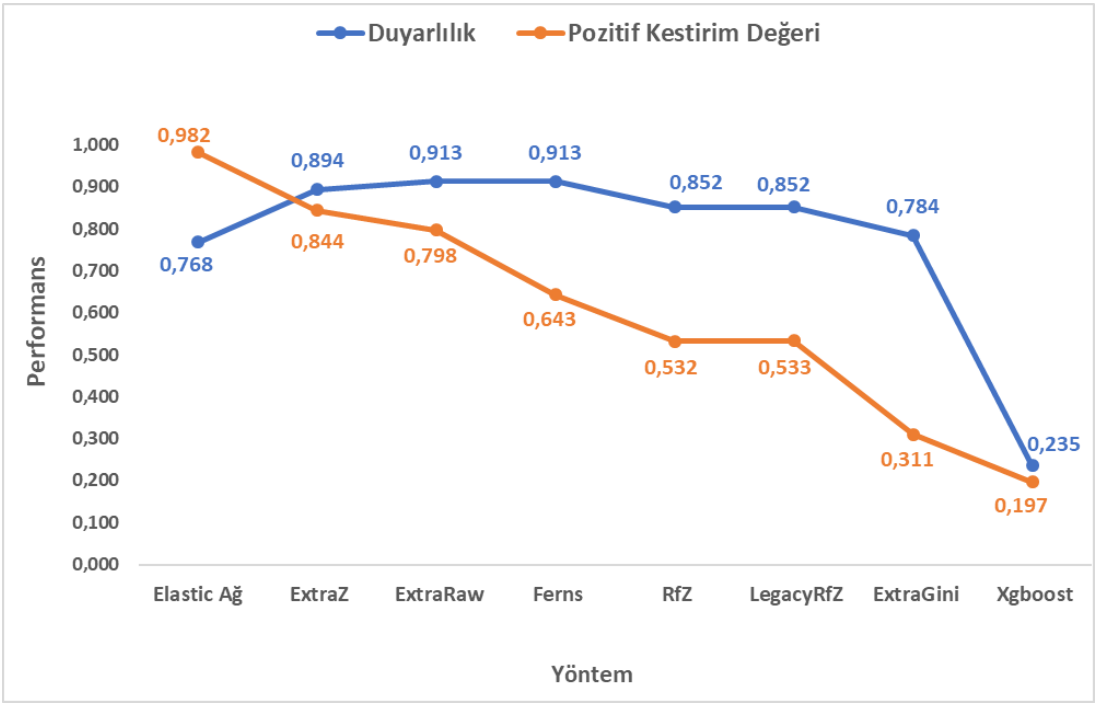
Şekil 4.2. Böbrek renal şeffaf hücreli karsinom veri seti üzerinden algoritmaların Duyarlılık Oranları ve Pozitif Kestirim Değerlerinin performanslarının karşılaştırılması.



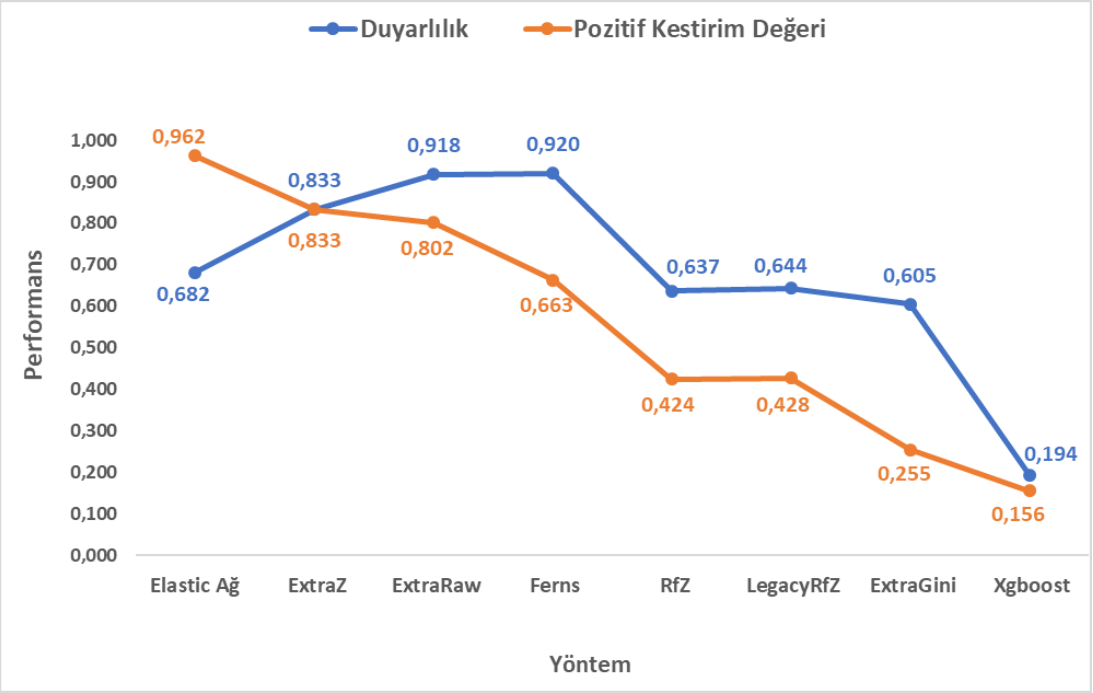
Şekil 4.3. Kolon adenokarsinomu veri seti üzerinden algoritmaların Duyarlılık Oranları ve Pozitif Kestirim Değerlerinin performanslarının karşılaştırılması.



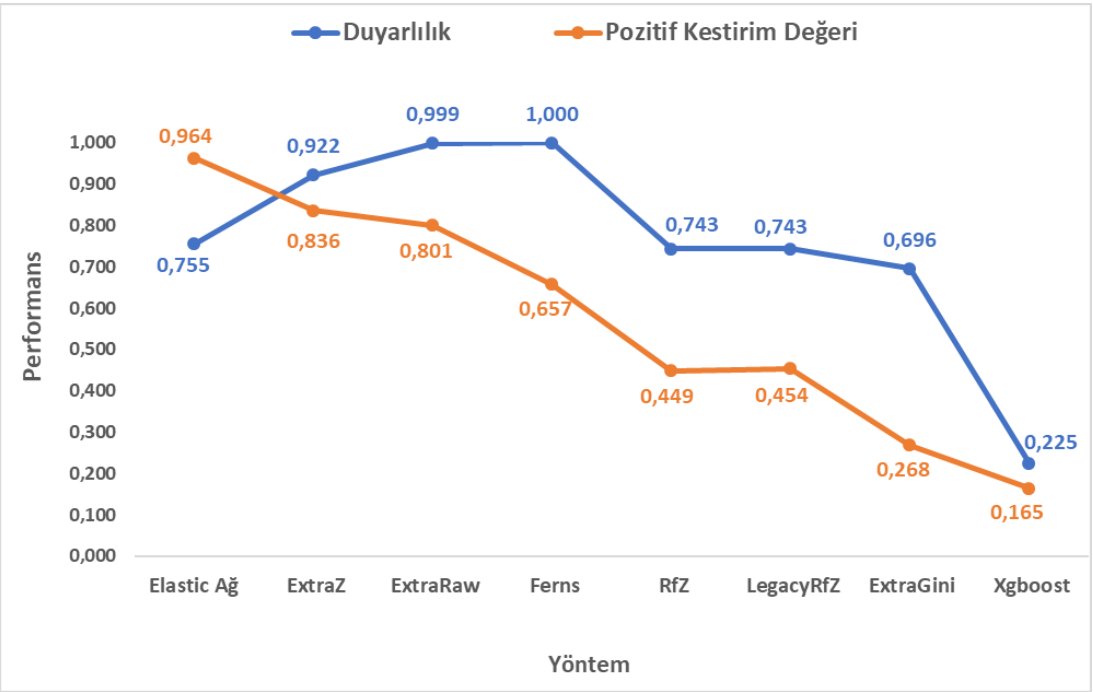
Şekil 4.4. Karaciğer hepatoselüler karsinomu veri seti üzerinden algoritmaların Duyarlılık Oranları ve Pozitif Kestirim Değerlerinin performanslarının karşılaştırılması.



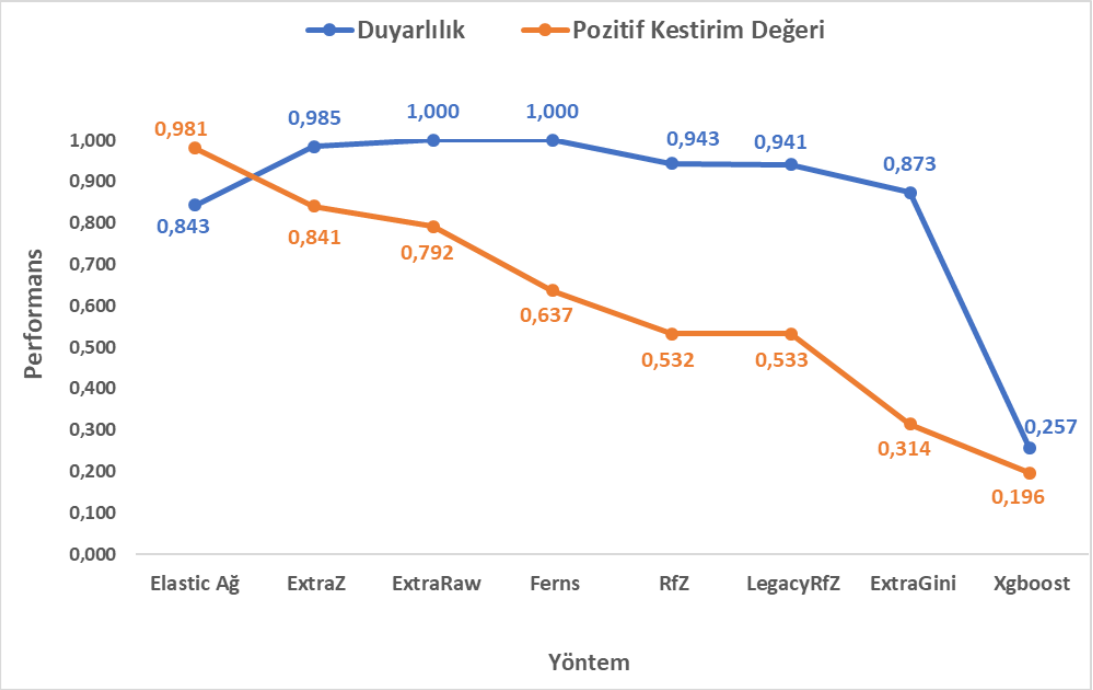
Şekil 4.5. Akciğer adenokarsinomu veri seti üzerinden algoritmaların Duyarlılık Oranları ve Pozitif Kestirim Değerlerinin performanslarının karşılaştırılması.



Şekil 4.6. Akciğer yassı hücreli karsinomu veri seti üzerinden algoritmaların Duyarlılık Oranları ve Pozitif Kestirim Değerlerinin performanslarının karşılaştırılması.



Şekil 4.7. Prostat adenokarsinomu veri seti üzerinden algoritmaların Duyarlılık Oranları ve Pozitif Kestirim Değerlerinin performanslarının karşılaştırılması.



Şekil 4.8. Tiroid karsinomu veri seti üzerinden algoritmaların Duyarlılık Oranları ve Pozitif Kestirim Değerlerinin performanslarının karşılaştırılması.

Tablo 4.3. Sekiz farklı RNA dizileme veri setinde, sekiz farklı yöntem çerçevesinde kurulan özellik seçimi modellerinin F1 Ölçülerinin karşılaştırılması.

Veri Seti	Elastic Ağ Algoritması	Boruta Algoritması						
		ExtraZ	ExtraRaw	Ferns	RfZ	LegacyRfZ	ExtraGini	Xgboost
KICH	0,623	0,672	0,612	**0,691**	0,488	0,489	0,441	0,293
KIRC	**0,949**	0,911	0,883	0,773	0,721	0,720	0,504	0,259
COAD	0,765	0,806	**0,809**	0,761	0,459	0,461	0,400	0,186
LIHC	0,841	**0,873**	0,850	0,766	0,580	0,582	0,432	0,202
LUAD	**0,903**	**0,903**	0,884	0,779	0,674	0,675	0,454	0,220
LUSC	0,833	0,862	**0,886**	0,794	0,522	0,527	0,366	0,176
PRAD	0,843	0,872	**0,886**	0,789	0,556	0,560	0,385	0,187
THCA	**0,904**	**0,904**	0,880	0,775	0,677	0,677	0,460	0,219

Tablo 4.4. Sekiz farklı RNA dizileme veri setinde, sekiz farklı yöntem çerçevesinde kurulan özellik seçimi modellerinin çeşitli performans ölçülerinin karşılaştırılması.

Veri Setleri	Ölçütler	Elastik Ağ Algoritması	Boruta Algoritması						
			ExtraZ	ExtraRaw	Ferns	RfZ	LegacyRfZ	ExtraGini	Xgboost
COAD	Seçicilik	**0,998**	0,984	0,977	0,966	0,944	0,944	0,887	0,926
	Doğruluk	**0,980**	0,979	0,976	0,968	0,927	0,927	0,882	0,892
	Kappa	0,751	0,791	**0,793**	0,741	0,421	0,423	0,350	0,132
	MCC	0,771	0,800	**0,811**	0,771	0,440	0,442	0,414	0,138
	Dengeli Doğruluk	0,824	0,928	0,970	**0,979**	0,778	0,779	0,832	0,586
	Youden index	0,647	0,856	0,939	**0,958**	0,555	0,558	0,665	0,172
KICH	Seçicilik	**0,996**	0,972	0,949	0,963	0,944	0,944	0,889	0,951
	Doğruluk	**0,971**	0,962	0,945	0,959	0,930	0,930	0,888	0,920
	Kappa	0,610	0,652	0,585	**0,671**	0,454	0,454	0,395	0,251
	MCC	0,642	0,660	0,616	**0,693**	0,474	0,474	0,467	0,256
	Dengeli Doğruluk	0,743	0,871	0,905	**0,928**	0,804	0,804	0,878	0,642
	Youden index	0,487	0,741	0,810	**0,856**	0,609	0,609	0,757	0,284
KIRC	Seçicilik	**1,000**	0,989	0,986	0,968	0,959	0,958	0,899	0,949
	Doğruluk	**0,995**	0,990	0,986	0,969	0,960	0,960	0,903	0,916
	Kappa	**0,942**	0,902	0,873	0,754	0,698	0,698	0,463	0,214
	MCC	**0,948**	0,911	0,884	0,782	0,735	0,734	0,544	0,218
	Dengeli Doğruluk	0,954	**0,992**	0,991	0,982	0,975	0,976	0,934	0,620
	Youden index	0,908	**0,984**	0,982	0,963	0,950	0,952	0,868	0,240
LIHC	Seçicilik	**0,999**	0,987	0,981	0,967	0,949	0,949	0,892	0,934
	Doğruluk	**0,986**	0,985	0,982	0,968	0,941	0,941	0,889	0,900
	Kappa	0,829	**0,861**	0,837	0,747	0,549	0,551	0,385	0,150
	MCC	0,843	**0,871**	0,853	0,776	0,578	0,581	0,453	0,155
	Dengeli Doğruluk	0,872	0,972	**0,987**	0,981	0,874	0,876	0,862	0,593
	Youden index	0,744	0,945	**0,973**	0,962	0,748	0,752	0,723	0,186
LUAD	Seçicilik	**0,999**	0,991	0,987	0,972	0,959	0,959	0,906	0,948
	Doğruluk	**0,988**	0,986	0,983	0,969	0,954	0,954	0,900	0,912
	Kappa	**0,821**	0,820	0,801	0,698	0,595	0,596	0,375	0,158
	MCC	**0,903**	0,902	0,885	0,788	0,683	0,684	0,477	0,176
	Dengeli Doğruluk	0,884	0,942	**0,950**	0,942	0,905	0,906	0,845	0,591
	Youden index	0,767	0,884	**0,900**	0,885	0,811	0,811	0,690	0,183
LUSC	Seçicilik	**0,999**	0,991	0,987	0,974	0,953	0,953	0,906	0,943
	Doğruluk	0,983	0,983	**0,984**	0,971	0,938	0,938	0,891	0,905
	Kappa	0,760	0,787	**0,809**	0,718	0,452	0,456	0,292	0,116
	MCC	0,835	0,859	**0,887**	0,801	0,509	0,515	0,361	0,129
	Dengeli Doğruluk	0,840	0,912	**0,952**	0,947	0,795	0,799	0,756	0,568
	Youden index	0,681	0,824	**0,905**	0,894	0,590	0,597	0,511	0,137

Tablo 4.4. Sekiz farklı RNA dizileme veri setinde, sekiz farklı yöntem çerçevesinde kurulan özellik seçimi modellerinin çeşitli performans ölçülerinin karşılaştırılması (devamı).

Veri Setleri	Ölçütler	Elastik Ağ Algoritması	Boruta Algoritması						
			ExtraZ	ExtraRaw	Ferns	RfZ	LegacyRfZ	ExtraGini	Xgboost
PRAD	Seçicilik	**0,998**	0,990	0,986	0,971	0,951	0,952	0,898	0,938
	Doğruluk	0,986	0,986	**0,987**	0,972	0,940	0,941	0,888	0,902
	Kappa	0,836	0,865	**0,879**	0,776	0,526	0,531	0,338	0,138
	MCC	0,845	0,869	**0,887**	0,797	0,548	0,551	0,385	0,141
	Dengeli Doğruluk	0,877	0,956	**0,993**	0,985	0,847	0,847	0,797	0,581
	Youden index	0,754	0,912	**0,985**	0,971	0,694	0,695	0,594	0,163
THCA	Seçicilik	**0,999**	0,989	0,985	0,968	0,955	0,955	0,897	0,942
	Doğruluk	**0,991**	0,989	0,986	0,970	0,954	0,954	0,896	0,908
	Kappa	**0,900**	0,898	0,873	0,760	0,655	0,655	0,417	0,172
	MCC	**0,904**	0,903	0,882	0,784	0,688	0,687	0,485	0,175
	Dengeli Doğruluk	0,921	0,987	**0,993**	0,984	0,949	0,948	0,885	0,599
	Youden index	0,843	0,974	**0,985**	0,968	0,898	0,896	0,770	0,199

5. TARTIŞMA

Tüm veri setlerine uygulanan özellik seçimi yöntemlerinin Duyarlılık performansları Tablo 4.1. üzerinden incelendiğinde, Boruta Algoritmasının Ferns Yönteminin önemli genlerin neredeyse tamamını doğru olarak bildiği anlaşılmaktadır. Çünkü bu yöntemin tüm veri setleri üzerindeki Duyarlılık Oranlarının %89,4 ile %100 arasında değiştiği görülmektedir. 1.000 tekrar sonucunda hesaplanan ortalamalara göre, Ferns Yöntemi daha önce anlamlı olarak üretilen 10 genin neredeyse tamamını önemli olarak önermekte ancak, söz konusu 10 geni içerse dahi çok fazla gen önermiş olma ihtimali bulunmaktadır. Diğer bir ifadeyle, Ferns Yönteminin önerdiği genlerin hepsi önemli olmayabilmektedir. Böyle durumlarda, Algoritmaların performanslarının Pozitif Kestirim Değeri ile de değerlendirilmesi önem arz etmektedir. Tablo 4.1.'de yer alan en yüksek Duyarlılık Oranının (%100), Boruta Algoritmasının Ferns yöntemi tarafından PRAD ve THCA veri setleri ile ExtraRaw yöntemi tarafından da THCA veri seti üzerinden elde edildiği görülmektedir.

Benzer şekilde, tüm veri setlerine uygulanan özellik seçimi yöntemlerinin Pozitif Kestirim Değeri performansları Tablo 4.2. üzerinden incelendiğinde; Elastik Ağ Genelleştirilmiş Doğrusal Modeller İçerisinde Determan'ın Algoritması, hesaplamalar sonucunda bulduğu genleri önemli gen olarak önerdiğinde bu sonuçların oldukça güvenilir olduğu anlaşılmaktadır. Çünkü Tablo 4.2.'de yer alan %88,5 ile %99,2 arasında değişen Pozitif Kestirim Değeri performans ölçüleri, Elastik Ağ Algoritmasının önerdiği genlerin gerçekte anlamlı olarak üretilen genlerin arasında bulunma oranının oldukça yüksek olduğunu göstermektedir. Tablo 4.2.'de yer alan en yüksek Pozitif Kestirim Değeri, Elastik Ağ Algoritması tarafından KIRC veri seti üzerinden elde edilmiştir.

Şekil 4.1. – Şekil 4.8.'de Duyarlılık Oranı ve Pozitif Kestirim Değeri ölçülerinin birlikte değişimi gösterilmektedir. Söz konusu şekiller incelendiğinde, Elastik Ağ Algoritmasının Pozitif Kestirim Değeri açısından tüm senaryolarda en yüksek performansa sahip olduğu açık bir şekilde görülmektedir. Diğer taraftan, senaryoların tamamında Boruta Algoritmasının Xgboost Yöntemi tarafından hem Pozitif Kestirim

Değeri hem de Duyarlılık Oranı ölçüleri bakımından en düşük performans değerinin elde edildiği net bir şekilde görülmektedir. Bu sonuçlara göre, benzer şartlar altında çalışma yapılacak olması durumunda Boruta Algoritmasının Xgboost Yöntemini kullanmanın gerekli olmadığı düşünülmektedir.

Tablo 4.3.'de yer alan F1 Ölçüsü sonuçları kapsamında, Duyarlılık Oranları ve Pozitif Kestirim Değerlerinin harmonik ortalaması olan F1 Ölçülerinde, Boruta Algoritmasının ExtraZ, ExtraRaw ve Ferns yöntemlerinin 8 veri setinin 7'sinde en yüksek performans değerini elde ettiği görülmüştür. Diğer taraftan, Tablo 4.3.'de yer alan en yüksek F1 Ölçüsü, Elastik Ağ Algoritması tarafından KIRC veri seti üzerinden elde edilmiştir. Ayrıca, Xgboost yönteminin senaryoların tamamında Duyarlılık Oranları ve Pozitif Kestirim Değerlerinde olduğu gibi bu ölçütlerin harmonik ortalaması olan F1 Ölçüleri bakımından tüm veri setlerinde en düşük performans değerlerini elde ettiği görülmüştür.

Tüm veri setlerinde, Tablo 4.4.'de yer alan Doğruluk performans sonuçlarına göre en düşük performansı ExtraGini Yöntemi göstermektedir. Yine aynı tablolarda bulunan Doğruluk haricindeki diğer yedi performans ölçüsünde, tüm veri setlerinde en düşük performans Xgboost Yöntemi tarafından gösterilmektedir. Diğer bir ifadeyle, özellik seçimi söz konusu olduğunda sıklıkla tercih edilen Pozitif Kestirim Değerleri, Duyarlılık Oranları ve F1 Ölçülerinin Tablo 4.1., Tablo 4.2. ve Tablo 4.3.'de yer alan performans sonuçlarına göre tüm veri setlerinde en düşük performansa Xgboost Yönteminin sahip olduğu görülmektedir. Tablo 4.4.'de yer alan Xgboost Yönteminin tüm veri setleri üzerinden özellik seçiminde performans sonuçları; Duyarlılık Oranları %19,4 ile %33,3 arasında, Pozitif Kestirim Değerleri %15,2 ile %26,9 arasında ve F1 Ölçüleri %17,6 ile %29,3 arasında değişmektedir. Bu bilgiler ışığında, özellik seçimi çalışmalarında, diğer yedi yöntemlerden en az bir tanesinin kullanılması halinde Xgboost Yönteminin kullanılmasına gerek olmadığı düşünülmektedir.

Çalışmada kullanılan sekiz yöntem arasından en yüksek özellik seçimi performansının; Elastik Ağ, ExtraZ, ExtraRaw ve Ferns yöntemleri arasında performans ölçütlerine ve veri setlerine bağlı olarak değiştiği görülmektedir. Doğruluk

ölçüsü bakımından, 8 veri setinin 6'sında en yüksek performansa Elastik Ağ yönteminin sahip olduğu görülmekte olup diğer 2 veri setinde (LUSC ve PRAD) ise ExtraRaw yönteminin en yüksek performansa sahip olduğu görülmektedir. Dengeli Doğruluk ölçüsü açısından ise 8 veri setinin 5'inde en yüksek performansa ExtraRaw yönteminin ulaştığı görülmekte olup diğer 3 veri setinin 2'sinde (COAD ve KICH) Ferns yöntemi, 1'inde (KIRC) ise ExtraZ yönteminin en yüksek performansa ulaştığı görülmektedir. Seçicilik ölçüsü bakımından, Tablo 4.4.'de yer alan sonuçlara göre tüm veri setlerinde en yüksek performansa Elastik Ağ yönteminin sahip olduğu görülmektedir. Seçicilik ölçüsünün veri setlerine göre performans değerleri %99,6 ile %100,0 arasında değişmektedir. Kappa ölçüsü bakımından, tüm veri setlerinde genel olarak baskın performans gösteren bir yöntem bulunmamakla birlikte, en yüksek performanslarla öne çıkan yöntemler arasında Elastik Ağ ve ExtraRaw yöntemleri yer almaktadır. 8 veri setinin 6'sında en yüksek performansa söz konusu iki yöntemin sahip olduğu görülmekte olup diğer 2 veri setinin 1'inde (LIHC) ExtraZ yönteminin, diğerinde (KICH) Ferns yönteminin en yüksek performansa sahip olduğu görülmektedir. MCC ölçüsü kapsamında, tüm veri setlerinde Kappa ölçüsüne benzer bir performans elde edilmiştir. Her ne kadar performans değerlerinde küçük farklılıklar da olsa, en yüksek ve en düşük performansa aynı veri setleri üzerinden aynı yöntemlerin sahip olduğu dikkat çekmektedir. Youden İndeks ölçüsü performans sonuçlarına göre, tüm veri setleri üzerinden Dengeli Doğruluk ölçüsüne benzer bir performans elde edilmektedir. Performans oranlarında küçük farklılıklar olmakla birlikte, en yüksek ve en düşük performansa aynı yöntemlerin aynı veri setlerini kullanarak sahip olduğu fark edilmektedir. Diğer taraftan, Tablo 4.4.'de Dengeli Doğruluk ve Youden İndeksi ölçüleri açısından Boruta Algoritması altında bulunan 3 farklı yöntemin en yüksek performansa sahip olduğu ve aynı zamanda Boruta Algoritmasının, Elastik Ağ Algoritmasından daha yüksek performansa sahip olduğu görülmektedir.

Tablo 4.4.'de yer alan sonuçlar çerçevesinde, tüm veri setlerinde üzerinden sekiz farklı yöntemin performanslarını gösteren dokuz farklı performans ölçütü arasından çalışmanın en dikkat çekici sonucu, Pozitif Kestirim Değerleri bakımından

Elastik Ağ Yöntemi tarafından elde edilmiştir. Elastik Ağ Yöntemine ait Pozitif Kestirim Değerleri tüm veri setlerinde %88,5 ile %99,2 arasında değişmektedir. Pozitif Kestirim Değeri, modelin anlamlı dediği genlerin gerçekte kaç tanesinin anlamlı olduğunu göstermektedir. Bu kapsamda, bahse konu sonuçlar çerçevesinde; Elastik Ağ Algoritmasının, önemli olarak tahmin ettiği genlerin gerçekte önemli olma ihtimalinin diğer yedi yönteminkinden daha yüksek olduğu anlaşılmaktadır. Duyarlılık Oranlarına ilişkin 8 veri setinin 7'sinde en yüksek özellik seçimi performansına Ferns yönteminin sahip olduğu görülmekte olup diğer 1 veri setinde ise (KIRC) ExtraRaw yönteminin en yüksek performansa sahip olduğu görülmektedir. F1 ölçüsü bakımından, tüm veri setlerinde genel olarak etkili performans gösteren bir yöntem bulunmamakla birlikte, en yüksek performanslarla Elastik Ağ, ExtraZ ve ExtraRaw yöntemleri öne çıkmaktadır.

Ayrıca, 91 gözlemin bulunduğu en küçük veri seti olan KICH'den başlayarak 605 gözlemin yer aldığı en büyük veri seti olan KIRC'e kadar büyüklük sırasına dikkat ederek yapılan incelemede, tüm performans ölçüleri bakımından performans değeri ile veri setinin büyüklüğü arasında bir ilişkiye rastlanmamıştır. Çünkü veri setinin büyüklüğü artarken tüm ölçütler için performans değerinin zaman zaman arttığı ya da azaldığı görülmüş olup doğru orantılı bir artış ya da azalış görülmemiştir.

Çalışma sonuçlarına bakıldığında, RNA dizileme veri setleri kullanılarak özellik seçimi yapılmasında Determan'ın Elastik Ağ Genelleştirilmiş Doğrusal Modeller İle Optimal Gen Seçim Algoritması ve Boruta Algoritması bünyesinde yer alan ExtraZ, ExtraRaw ve Ferns Yöntemlerinin başarılı bir performans gösterdiği söylenebilmektedir. Cao ve ark. (2022), Soneson ve Robinson (2018) tarafından yapılan çalışmada sağlanan üç adet üretilmiş gen veri setini kullanmıştır. Dört makine öğrenimi algoritmasının gen seçim performanslarını incelemeleri sonucunda, Elastik Ağ Algoritmasının gen seçiminde en yüksek performansa sahip olduğu bulgusuna ulaşmışlardır. DEG'lerin tespiti için birçok makine öğrenimi algoritması kullanılmaktadır. Ancak seçilen genler, algoritmaya bağlı olarak önemli ölçüde farklılık gösterebilmektedir. Wenric ve Shemirani (2018), çeşitli kanser örneklerini içeren RNA dizileme veri setlerinde diferansiyel olarak ifade edilen genleri bulmak için Rastgele Orman Algoritmasını kullanmıştır. Vener ve ark. (2022), önemli genleri tanımlamak

için Rastgele Orman, Destek Vektör Makineleri ve Elastik Ağ Algoritmalarını kullanmıştır. Kasikci ve Dag (2023), Determan'ın Elastik Ağ Gen Seçim Algoritmasının gen seçimi açısından Biosigner Algoritması, GMDH Tipi Sinir Ağı Algoritması, Determan'ın Rastgele Orman İle Optimal Gen Seçim Algoritması ve Determan'ın Destek Vektör Makineleri İle Optimal Gen Seçim Algoritmasından daha iyi performans gösterdiğini belirtmiştir.

6. SONUÇ VE ÖNERİLER

Bu çalışmada, kapsamlı bir Monte Carlo benzetim çalışması ile Boruta ve Elastik Ağ Algoritmalarının gen seçim performanslarının değerlendirilmesi amaçlanmıştır. Benzetim senaryoları, sekiz gerçek RNA dizileme veri setine dayalı olarak tasarlanmıştır. Söz konusu veri setleri, kanserin mevcut olup olmadığını gösteren sınıf değişkenini içermektedir. Bu çalışmadaki veri setlerinin çoğunluğu kanser hastalarından oluştuğu için sınıf dağılımları oldukça dengesizdir. Bu nedenle, veri setleri bu sınıf dağılımları esas alınarak üretilmiştir. Toplam sekiz senaryo bulunmakta ve bu senaryolar çerçevesinde gen seçimi için iki farklı makine öğrenimi algoritması uygulanmıştır. Tüm araştırmamızın ışığında Elastik Ağ Algoritmasının özellik seçimi performansı, Pozitif Kestirim Değerleri açısından tüm senaryolarda dikkat çekicidir. Bu bulgulara dayanarak Elastik Ağ Yönteminin, diferansiyel olarak ifade edilen genlerin tanımlanması için umut verici olduğu ve tercih edilebilir olduğu değerlendirilmektedir. Tabii ki, bu bulguların klinik değerlendirme ve doğrulamaya ihtiyacı bulunmakta, ancak yine de araştırmacılar için rehberlik sağlayabilmektedirler.

Sonuç olarak;

- Elastik Ağ Algoritması, genel olarak tüm veri setlerinde Pozitif Kestirim Değeri açısından öne çıkmaktadır. Pozitif Kestirim Değeri, model tarafından anlamlı olarak tahmin edilen genlerin gerçekte kaç tanesinin anlamlı olduğunu göstermektedir.

- Boruta Algoritmasının Ekstra Trees ve Ferns tabanlı yöntemleri, Duyarlılık Oranları açısından Elastik Ağ Algoritmasından daha iyi performans göstermektedir. Duyarlılık, gerçekte anlamlı olan genlerin kaç tanesini modelin anlamlı olarak bulduğunu göstermektedir.

- Eğer çalışmanın amacı, belirli bir hastalık veya duruma ilişkin tüm ilgili genleri belirlemekse Duyarlılık, Pozitif Kestirim Değerinden daha önemlidir. Çünkü önemli bir genin kaçırılması teşhis veya tedavi fırsatlarının kaçırılmasına neden olabilmektedir.

- Eğer çalışmanın amacı, daha fazla deneysel doğrulama için küçük bir gen setini tanımlamaksa Pozitif Kestirim Değeri, Duyarlılıktan daha önemlidir. Çünkü çok fazla ilgisiz genin seçilmesi, zaman ve kaynak açısından maliyetli olabilmektedir. Bu durumda, çalışma kapsamında Pozitif Kestirim Değeri bakımından en iyi sonuçlar Elastik Ağ Algoritması ile elde edildiğinden benzer durumlarda anılan yöntemin tercih edilmesinin daha yararlı olacağı düşünülmektedir.

Diğer taraftan, günümüz teknolojileri çerçevesinde oldukça yüksek özelliklere sahip bir bilgisayarda (64 GB bellek, 12 çekirdek ve 2,4 GHz işlemci) gözlem sayısı en fazla 605 ve daha az olan RNA dizileme veri setleri kullanılarak 1.000 tekrarlı benzetim çalışması yapılabilmiştir. Daha yüksek yazılım ve donanım teknolojilerine erişim sağlanabilmesi durumunda, gözlem sayısı özellikle 1.000'in üzerindeki veri setleri için ileride yapılacak çalışmalarda aynı benzetim çalışmasının uygulanabileceği düşünülmektedir.

7. KAYNAKLAR

1. Feature Selection in R with the Boruta R Package [Internet]. 2018 [Erişim Tarihi 8 Mayıs 2023]. Erişim adresi: https://www.datacamp.com/tutorial/feature-selection-R-boruta
2. Kursa MB, Rudnicki WR. Feature Selection with the Boruta Package. Journal of Statistical Software. 2010;Volume36,Issue11.
3. Perlato, A. Feature Selection using Boruta Algorithm [Internet]. 2023 [Erişim Tarihi 17 Haziran 2023]. Erişim adresi: https://www.andreaperlato.com/mlpost/feature-selection-using-boruta-algorithm/
4. Kaşıkcı M, Dağ O. A Comprehensive Real Data-Based Simulation Study on Gene Selection Performance of Machine Learning Algorithms. (under review). 2023.
5. Mazzanti, S. Looking under the hood of Boruta, one of the most effective feature selection algorithms [Internet]. 2020 [Erişim Tarihi 7 Mayıs 2023]. Erişim adresi: https://towardsdatascience.com/boruta-explained-the-way-i-wish-someone-explained-it-to-me-4489d70e154a
6. Bhalla, D. Feature Selection: Select Important Variables With Boruta Package [Internet]. 2017 [Erişim Tarihi 17 Haziran 2023]. Erişim adresi: https://www.listendata.com/2017/05/feature-selection-boruta-package.html.
7. Breiman L. Random Forests. Machine Learning. 2001;45:5-32.
8. Stoppiglia H, Dreyfus G, Dubois R, Oussar Y. Ranking a Random Feature for Variable and Feature Selection. Journal of Machine Learning Research. 2003;3,1399-1414.
9. D'Agostino, A. Feature Selection with Boruta in Python [Internet]. 2021 [Erişim Tarihi 17 Haziran 2023]. Erişim adresi: https://towardsdatascience.com/feature-selection-with-boruta-in-python-676e3877e596.
10. Rigatti, S. J. Random Forest [Internet]. 2017 [Erişim Tarihi 9 Temmuz 2023]. Erişim adresi: https://meridian.allenpress.com/jim/article-abstract/47/1/31/131479/Random-Forest.
11. Muratlar, E. R. Boruta Algoritması ile Değişken Seçimi [Internet]. 2021 [Erişim Tarihi 17 Haziran 2023]. Erişim adresi: https://www.veribilimiokulu.com/boruta-algoritmasi/
12. Dutta, D. How to perform feature selection using Boruta Package in R? [Internet]. 2022 [Erişim Tarihi 7 Mayıs 2023]. Erişim adresi: https://www.analyticsvidhya.com/blog/2016/03/select-important-variables-boruta-package/.
13. Liaw A, Wiener M. Classification and Regression by randomForest. R News. 2002;2(3),18-22.

14. Kursa, M.B. Wrapper Algorithm for All Relevant Feature Selection [Internet]. 2022 [Erişim Tarihi 19 Ağustos 2023]. Erişim adresi: https://search.r-project.org/CRAN/refmans/Boruta/html/00Index.html
15. Determan C. Optimal algorithm for metabolomics classification and feature selection varies by dataset. International Journal of Biology. 2015;7(1):100–115.
16. Zou H, Hastie T. Regularization and variable selection via the elastic net. Royal Statistical Society. 2004;67,Part2,pp.301–320.
17. Cui L, Bai L, Wang Y, Jin X, Hancock ER. Internet financing credit risk evaluation using multiple structural interacting elastic net feature selection. Pattern Recognition. 2021;107835.
18. Dağ O. Binary Classification Via GMDH-Type Neural Network Algorithm [PhD thesis]. Ankara: Hacettepe University; 2018.
19. Marafino BJ, Boscardin WJ, Dudley RA. Efficient and sparse feature selection for biomedical text classification via the elastic net: Application to ICU risk stratification from nursing notes. Journal of Biomedical Informatics. 2015;114–120.
20. Mola CD, Vito ED, Rosasco L. Elastic-net regularization in learning theory. Journal of Complexity. 2009;201-230.
21. Sokolova M, Lapalme G. A Systematic Analysis of Performance Measures for Classification Tasks. Information Processing and Management. 2009;45:427-437.
22. Kaşıkcı M. Transkriptom Veri Seti Üzerinde Derin Öğrenme Yöntemi İle Klasik Veri Madenciliği Yöntemlerinin Sınıflama Performanslarının Karşılaştırılması[Yüksek Lisans Tezi]. Ankara: Hacettepe Üniversitesi; 2019.
23. Arık Ö. Mikrodizi Gen İfade Verilerinde Farklı Öznitelik Seçim Yöntemleri İle Sınıflama Yöntemlerinin Performanslarının Değerlendirilmesi [Doktora Tezi]. Ankara: Hacettepe Üniversitesi; 2020.
24. Arie BD. Comparison of Classification Accuracy Using Cohen's Weighted Kappa. Expert Systems with Applications. 2008;34:825-832.
25. Alpar R. Spor, Sağlık ve Eğitim Bilimlerinden Örneklerle Uygulamalı İstatistik ve Geçerlik-Güvenirlik. 4. Baskı. Ankara: Detay Yayıncılık; 2016.
26. Chicco D, Jurman G. The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. BMC Genomics. 2020;doi.org/10.1186/s12864-019-6413-7.
27. Chicco D, Tötsch N, Jurman G. The Matthews correlation coefficient (MCC) is more reliable than balanced accuracy, bookmaker informedness, and markedness in two-class confusion matrix evaluation. BioData Mining. 2021;doi.org/10.1186/s13040-021-00244-z.
28. Schisterman EF, Perkins NJ, Liu A, Bondell H. Optimal Cut-point and Its Corresponding Youden Index to Discriminate Individuals Using Pooled Blood Samples. Epidemiology. 2005;Volume16.

29. Anguita D, Ghelardoni L, Ghio A, Oneto L, Ridella S. The 'K' in K-fold Cross Validation. Esann. 2012;11A,I-16145.

30. Özdoğan, M. Kanser Genom Atlası [Internet]. 2022 [Erişim Tarihi 15 Temmuz 2023]. Erişim adresi: https://www.drozdogan.com/kanser-genom-atlasi-nedir-tarihi-ve-onemli-donum-noktalari/.

31. Mohamed M, Lucchetta M, Silva TC, Olsen C, Bontempi G, Chen X, et al. New functionalities in the tcgabiolinks package for the study and integration of cancer data from gdc and gtex. PLoS computational biology. 2019;15(3):e1006701.

32. Bi R, Liu P. (2019, version 1.3.2) ssizeRNA: Sample Size Calculation for RNA-Seq Experimental Design, R Package. https://cran.r-project.org/web/packages/ssizeRNA/index.html.

33. Ren X, Kuan PF. Negative binomial additive model for RNA-Seq data analysis. BMC Bioinformatics. 2020;21:171.

34. Kuhn M. (2020, version 6.0-86) caret: Classification and Regression Training, R Package. https://cran.r-project.org/web/packages/caret/index.html.

35. Patro SGK, Sahu KK. Normalization: A Preprocessing Stage. arXiv. 2015;1503.06462.

36. Maza E, Frasse P, Senin P, Bouzayen M, Zouine M. Comparison of normalization methods for differential gene expression analysis in RNA-Seq experiments. Communicative & Integrative Biology. 2013;6:6.e25849.

37. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for rna-seq data with deseq2. Genome Biology. 2014;15:550,DOI10.1186/s13059-014-0550-8.

38. Gentleman R, Carey V, Huber W, Hahne F. (2019, version 1.68.0) genefilter: genefilter: methods for filtering genes from high-throughput experiments, R Package.
https://www.bioconductor.org/packages/release/bioc/html/genefilter.html.

Özgür Saman Yüksek Lisans Tezi

ORİJİNALLİK RAPORU

%4	%4	%1	%1
BENZERLİK ENDEKSİ	İNTERNET KAYNAKLARI	YAYINLAR	ÖĞRENCİ ÖDEVLERİ

BİRİNCİL KAYNAKLAR

1	www.openaccess.hacettepe.edu.tr:8080 İnternet Kaynağı	%2
2	acikbilim.yok.gov.tr İnternet Kaynağı	<%1
3	ekutup.dpt.gov.tr İnternet Kaynağı	<%1
4	9lib.net İnternet Kaynağı	<%1
5	docplayer.biz.tr İnternet Kaynağı	<%1
6	www.ulusaltezmerkezi.net İnternet Kaynağı	<%1
7	Submitted to Hacettepe University Öğrenci Ödevi	<%1
8	nek.istanbul.edu.tr:4444 İnternet Kaynağı	<%1
9	docplayer.com.br İnternet Kaynağı	<%1

EK-2: Dijital Makbuz

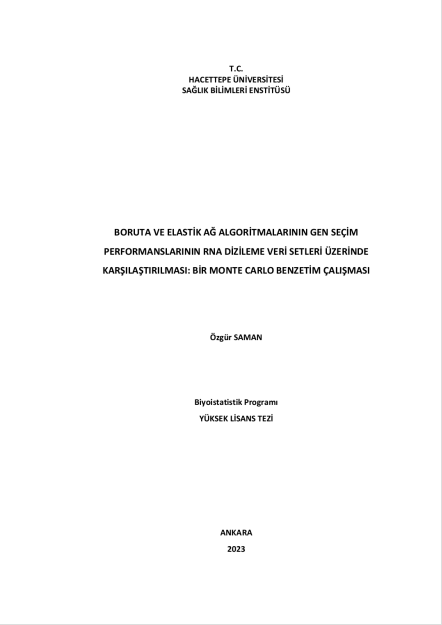
turnitin

Dijital Makbuz

Bu makbuz ödevinizin Turnitin'e ulaştığını bildirmektedir. Gönderiminize dair bilgiler şöyledir:

Gönderinizin ilk sayfası aşağıda gönderilmektedir.

Gönderen:	Özgür Saman
Ödev başlığı:	Özgür Saman Yüksek Lisans Tezi
Gönderi Başlığı:	Özgür Saman Yüksek Lisans Tezi
Dosya adı:	Tez_Ozgur_Saman.pdf
Dosya boyutu:	1.57M
Sayfa sayısı:	45
Kelime sayısı:	9,938
Karakter sayısı:	65,251
Gönderim Tarihi:	24-Ağu-2023 11:45ÖÖ (UTC+0300)
Gönderim Numarası:	2150419277



T.C.
HACETTEPE ÜNİVERSİTESİ
SAĞLIK BİLİMLERİ ENSTİTÜSÜ

BORUTA VE ELASTİK AĞ ALGORİTMALARININ GEN SEÇİM PERFORMANSLARININ RNA DİZİLEME VERİ SETLERİ ÜZERİNDE KARŞILAŞTIRILMASI: BİR MONTE CARLO BENZETİM ÇALIŞMASI

Özgür SAMAN

Biyoistatistik Programı
YÜKSEK LİSANS TEZİ

ANKARA
2023

9. ÖZGEÇMİŞ

Kişisel Bilgiler

Adı Soyadı : Özgür SAMAN

Doğum Yeri ve Tarihi :

İletişim Bilgileri :

Eğitim

Yüksek Lisans :

Lisans :