DERİN ÖĞRENME VE TOPLULUK ÖĞRENME YÖNTEMLERİNE DAYALI BİLGİSAYAR DESTEKLİ TANI SİSTEMİNİN GELİŞTİRİLMESİ: OMİK TEKNOLOJİLERİ ÜZERİNE UYGULAMASI

Ahmet Kadir ARSLAN

BİYOİSTATİSTİK ve TIP BİLİŞİMİ ANABİLİM DALI

Tez Danışmanı
Prof. Dr. Cemil ÇOLAK

Doktora Tezi – 2021

T.C.

İNÖNÜ ÜNİVERSİTESİ

SAĞLIK BİLİMLERİ ENSTİTÜSÜ

DERİN ÖĞRENME VE TOPLULUK ÖĞRENME YÖNTEMLERİNE DAYALI BİLGİSAYAR DESTEKLİ TANI SİSTEMİNİN GELİŞTİRİLMESİ: OMİK TEKNOLOJİLERİ ÜZERİNE UYGULAMASI

Ahmet Kadir ARSLAN

Biyoistatistik ve Tıp Bilişimi Anabilim Dalı

Doktora Tezi

Tez Danışmanı

Prof. Dr. Cemil ÇOLAK

MALATYA

2021

İÇİNDEKİLER

TEŞEKKÜR

Değerli hocalarım Prof. Dr. Cemil ÇOLAK ve Prof. Dr. Saim YOLOĞLU'na akademik yaşamımda bana verdikleri emek ve desteklerden ötürü teşekkür ederim.

Hayatımın her alanında olduğu gibi, bu tez çalışmam süresince gösterdikleri üstün sabır ve desteklerden dolayı sevgili eşim Ayşegül'e ve biricik kızım Aysima'ya teşekkür ederim.

Son olarak beni yetiştirip bugünlere gelmemi sağlayan sevgili anne ve babama hürmetlerimi sunuyorum.

Arş. Gör. Ahmet Kadir ARSLAN

ÖZET

Derin Öğrenme ve Topluluk Öğrenme Yöntemlerine Dayalı Bilgisayar Destekli Tanı Sisteminin Geliştirilmesi: Omik Teknolojileri Üzerine Uygulaması

Amaç: Bu tez çalışmasında, kolorektal kanser hastalığına ilişkin açık erişimli deneysel metabolomik veri seti kullanılarak, çeşitli topluluk öğrenme ve derin öğrenme modelleri ile ilgili hastalığın sınıflandırılmasını sağlayabilecek bir ardışık kod sisteminin (pipeline) tasarlanması ve bu kapsamda yüksek çıktılı bir karar destek sisteminin geliştirilmesi amaçlanmıştır.

Materyal ve Metot: Bu tez çalışmasında, Washington Üniversitesi, Anesteziyoloji ve Algoloji Bölümü, Northwest Metabolomik Araştırma Merkezi'nde yürütülen PR000226 numaralı proje kapsamında üretilen veri seti kullanılmıştır. İlgili veri seti, iki denek grubu, 66 KRK hastası ve 92 sağlıklı kontrol olmak üzere toplam 158 örnekten oluşmaktadır. Değişken seçim yöntemleri olarak LASSO, Elastic-Net, Boruta ve BorutaShap yöntemleri kullanılmıştır. Sınıflandırma görevinde ise XGBoost, LightGBM, derin sinir ağları ve yığın otokodlayıcı modelleri kullanılmıştır.

Bulgular: Bulgular incelendiğinde, en zayıf sınıflandırma performansı gösteren modelin yığın otokodlayıcı olduğu görülmektedir. Modelin hiçbir değişken seçimi senaryosunda istenilen sınıflandırma performansını başarısını gösteremediği görülmektedir. LightGBM modeli hem eğitim hem de test veri setlerinin sınıflandırmasında, tüm performans ölçütleri açısından, en iyi sonuçları vermiştir. Ayrıca LightGBM modeli bu sınıflandırma başarımını tüm değişken seçim yöntemleri bazında elde etmiştir.

Sonuç: Bu tez çalışmasında, topluluk öğrenme yöntemlerinin, tüm değişken seçim senaryolarında, derin öğrenme yöntemlerine kıyasla çok daha iyi sınıflandırma sonuçları verdiği görülmüştür. Söz konusu topluluk öğrenme yöntemlerinin, büyük veri setlerinde daha hızlı sonuç verebilmeleri için grafik işlem birimi (GPU) destekli sürümlerinin kullanılmasının işlem ve zaman maliyetleri açısından daha verimli olacağı önerilebilir.

Anahtar kelimeler: Bilgisayar destekli tanı, Derin öğrenme, Topluluk öğrenme, Metabolomik, Sınıflandırma.

ABSTRACT

Development of Computer Aided Diagnosis System Based on Deep Learning and Ensemble Learning Methods: Application on Omics Technologies

Aim: In this study, it is aimed to design a pipeline system, by using open access experimental metabolomics data set on colorectal cancer disease, which can classify the related disease by various ensemble learning and deep learning models and to develop a high-throughput decision support system.

Material and Method: In this thesis, the data set produced within the scope of the project numbered PR000226 carried out at Washington University, Department of Anesthesiology and Algology, Northwest Metabolomics Research Center was used. The related data set consisted of a total of 158 samples, including two groups of subjects, 66 CRC patients and 92 healthy controls. As variable selection methods, LASSO, Elastic-Net, Boruta and BorutaShap methods were used. In the classification task, XGBoost, LightGBM, deep neural networks and stacked autoencoder models were utilized.

Results: When the findings were examined, it was seen that the model with the worst classification performance is the stacked autoencoder. It was seen that the model cannot achieve the desired classification performance in any variable selection scenario. The LightGBM model gave the best results for all performance measures in the classification of both training and test datasets. In addition, the LightGBM model has achieved this classification performance on the basis of all variable selection methods.

Conclusion: In this thesis study, it was seen that ensemble learning methods had much better classification results compared to deep learning methods in all variable selection scenarios. It can be suggested that the use of graphics processing unit (GPU) supported versions of these ensemble learning methods will be more efficient in terms of processing and time costs so that they can provide faster results in large data sets.

Keywords: Computer aided diagnosis, Deep learning, Ensemble learning, Metabolomics, Classification.

SİMGELER VE KISALTMALAR DİZİNİ

KRK : Kolorektal Kanser

VTBK : Veri Tabanlarında Bilgi Keşfi

LASSO : Least Absolute Shrinkage and Selection Operator

XGBoost : Extreme Gradient Boosting

LightGBM : Light Gradient Boosting Machine

DSA : Derin Sinir Ağları

YOK : Yığın Otokodlayıcı

AKD : Ardışık Kod Dizini

NMR : Nükleer Manyetik Rezonans

LC-MS : Liquid chromatography–mass spectrometry

GC-MS : Gas chromatography–mass spectrometry

KDD : Knowledge Discovery in Databases

SHAP : SHapley Additive exPlanations

LIME : Local Interpretable Model-agnostic Explanations

ROC : Receiving Operating Characteristic

mRMRe : Minimum Redundancy Maximum Relevance Ensemble

PCA : Principal Component Analysis

PTD : Pozitif Tahmin Değeri

NTD : Negatif Tahmin Değeri

EAKA : Eğri Altında Kalan Alan

ŞEKİLLER DİZİNİ

Şekil No **Sayfa No**

TABLOLAR DİZİNİ

Tablo No **Sayfa No**

1. GİRİŞ

Son yıllarda, çeşitli hastalıklara ilişkin biyobelirteçlerin tespit edilebilmesi için "omik" yaklaşımları ile bilimsel çalışmalar yapılmaktadır. Bu önemli yaklaşımlar arasında yer alan metabolomik; hücreler, biyolojik sıvılar, dokular veya organizmalar içerisinde metabolitler olarak adlandırılan küçük moleküllerin kapsamlı analizlerini içeren çalışmalardır. Metabolomik analizler ile belirlenecek olası biyobelirteçler, hastalıkların erken tanı ve takibi ile tedavi stratejilerinin yönlendirilmesi vb. konularda kullanılabilmektedir.

Kolorektal kanser (KRK), ileri evrelerde yüksek ölüm oranı ile dünya çapında en yaygın üçüncü kanserdir. Ancak KRK tek tip bir tümör değildir; patogenezi tümörün anatomik konumuna bağlıdır ve kolonun sağ ve sol tarafı arasında farklılık göstermektedir (1). KRK sadece sanayileşmiş ve gelişmiş ülkelerde önemli bir sağlık sorunu değildir, gelişmekte olan ülkelerde de görülme sıklığı artmakta ve bu da onu küresel ölçekte en yaygın kanser haline getirmektedir. Birincil önleme için risk faktörü modifikasyonunun sınırlı etkisi göz önüne alındığında, tarama ve erken teşhis şeklinde ikincil önleme, şu anda kolonoskopi kullanımı yoluyla KRK kaynaklı ölümleri azaltmak için en etkili yaklaşımdır (2).

Diğer kanser türlerinde olduğu gibi KRK'da da, biyobelirteç keşfinde bir araç olarak metabolomik teknolojisinden faydalanılmaktadır. Metabolitler, hücresel aktivitenin sadece biyolojik faktörlerden etkilenmediği, aynı zamanda çevreye duyarlı olduğu nihai ürünlerdir. KRK ve diğer kanser türlerinde değişen metabolik yolakları ortaya çıkarmak ve kanser metabolik yapısını anlamada daha iyi bilgiler sağlaması açısından son yıllarda metabolomik temelli çalışmaların sayısı artmaktadır (3).

Özellikle son 20 yıl içerisinde metabolomik deneyi temelli çalışmalar ve bu çalışmalardan üretilen metabolomik veri setleri gün geçtikçe artmış ve günümüzde de artmaya devam etmektedir. Özellikle, gaz ya da sıvı kromatografisi ile birlikte kullanılan kütle spektrometresi (GC-MS, LC-MS) ve nükleer manyetik rezonans (NMR) spektroskopisi tekniklerinin kullanılarak kanıt ve bilgi düzeyi anlamında yüksek çıktılı metabolomik verilerin elde edilmesi, metabolomik araştırmalara ilgiyi artırmaktadır. Bunun bir sonucu olarak, metabolomik veri analizi için ücretsiz ve açık erişimli iş akışları (workflow), ardışık kod düzenleri (pipeline), paketler/kütüphaneler ve masaüstü/web tabanlı yazılımlar geliştirmeye odaklanan bilimsel çalışmalar yapılmaktadır. Özellikle R

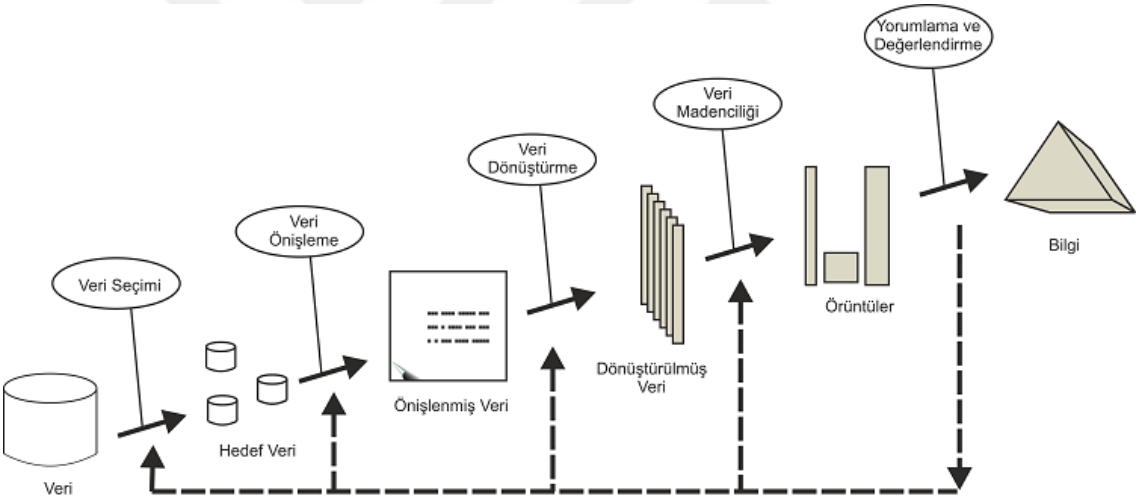
ve Python programlama dili kullanılarak oluşturulan bu araçların getirdiği otomatikleştirme, standartlaştırılma ve paylaşılabilme özellikleri sayesinde, araştırma ve sonuçların doğrulanabilirliği ve tekrarlanabilirliği büyük ölçüde iyileştirilebilir (4). İlgili omik veri setine ve hatta deney düzenine göre özelleştirilebilen, uyarlanabilen ayrıca çeşitli istatistik ve makine öğrenmesi yaklaşımlarını içeren bu araçların geliştirilmesi sayesinde, büyük metabolomik veri setlerinin analizinde ve çıktıların yorumlanabilmesinde büyük aşamalar kaydedilmiştir (5).

Bu tez çalışmasında, KRK hastalığına ilişkin açık erişimli deneysel metabolomik veriler kullanılarak ilgili hastalığın uygun biyobelirteçlerinin saptanması, topluluk öğrenme yöntemleri ve derin öğrenme mimarilerinin kullanılarak ilgili hastalığın sınıflandırılmasını sağlayabilecek bir sistemin tasarlanması ile bu kapsamda ilgili hastalığa yönelik bilgisayar destekli tanı/sınıflandırma için gerekli ardışık kod dizini geliştirilmesi amaçlanmıştır.

2. GENEL BİLGİLER

2.1. Veri Tabanlarında Bilgi Keşfi (VTBK) Süreci

1996 yılında Fayyad, Piatetski-Shapiro ve Smyth tarafından ortaya konulan Veri Tabanlarında Bilgi Keşfi (Knowledge Discovery in Databases, VTBK) süreci (6), büyük veri kümelerinin otomatik, keşif amaçlı bir analizi ve modellemesidir. VTBK, büyük ve karmaşık veri kümelerinden yeni, geçerli, kullanışlı ve anlaşılır kalıpları belirlemeye yönelik organize bir süreçtir. Veri Madenciliği, verileri araştıran, modeli geliştiren ve önceden bilinmeyen örüntüleri keşfeden algoritmaların çıkarımını içeren VTBK sürecinin özüdür (7). VTBK, Şekil 1'de gösterildiği gibi, aşağıdaki adımları içeren yinelemeli bir süreçtir.



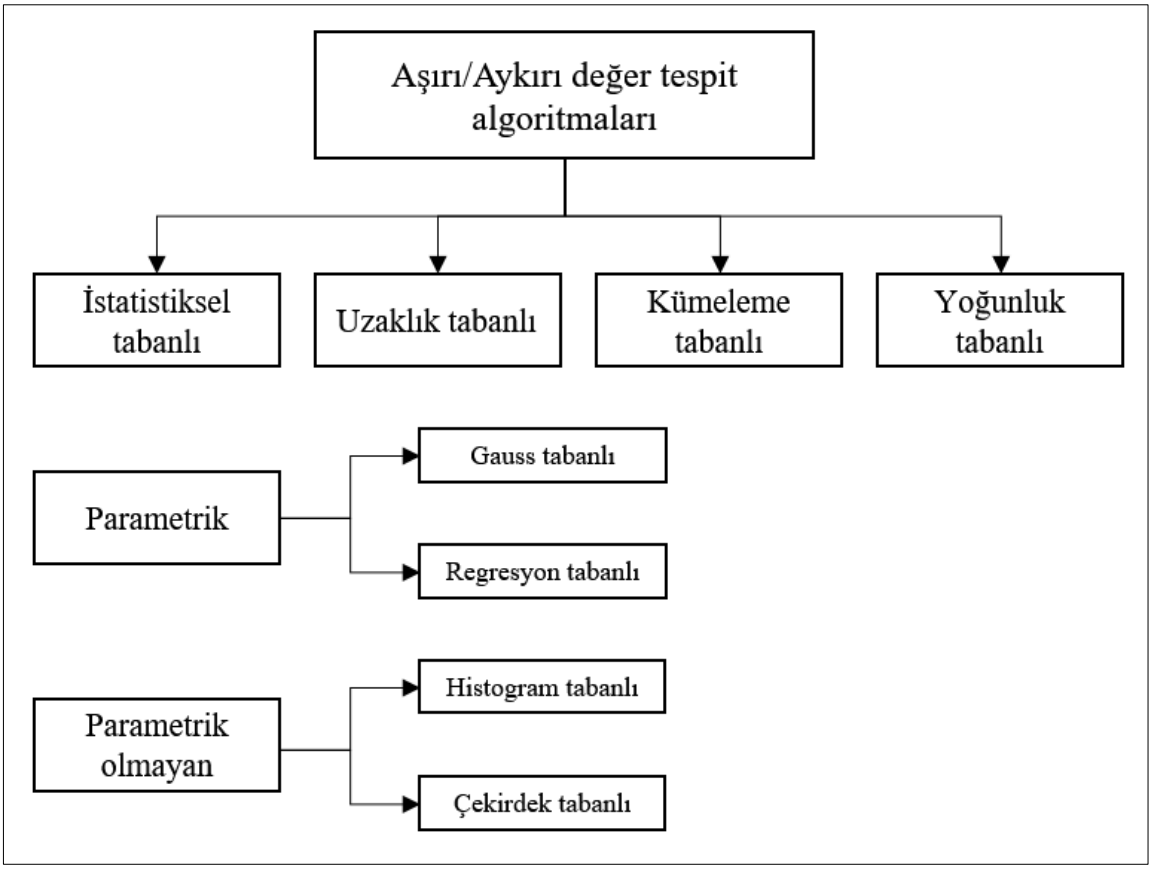
Şekil 1: VTBK sürecinin aşamaları

Veri seçimi aşaması, kaynaklarından ham verileri seçmekten oluşur. Bu veriler, eldeki sorunu çözmek için gerekli bilgileri sağlar nitelikte olmalıdır.

Veri önişleme, bilgi çıkarma için gerekli olan geçerli ve yapılandırılmış bir özellik kümesinin çıkarılmasını içerir. Ayrıca ilgisiz, gürültü, aşırı/aykırı veya eksik değerler gibi faktörlerin analizi ve veri setinden temizleme yöntemlerinin uygulanmasını ve verileri sonraki aşamaların algoritmalarının uygulanmasına uygun bir şekilde yapılandırmak için yeni özelliklerin türetilmesini içerir (7).

Kayıp değer ataması (missing value imputation), metabolomik verilerde kayıp değer problemi ile başa çıkmak sıklıkla başvurulan yaklaşımdır. Literatürde çoğunlukla k-en yakın komşuluk ataması, Random Forest atama ve tekil değer ayrıştırma (singular value decomposition) ataması gibi teknikler, metabolomik verilerde kayıp değer problemini ele almada kullanılmaktadır (8).

Aşırı/Aykırı değerler, bir veri kümesinin iyi tanımlanmış normlarından veya beklenen davranış kavramlarından sapma gösteren veri örnekleridir. Bazı durumlarda, veri analizini yanlış yönlendirdikleri kaldırılmaları tercih edilir, bazı durumlarda ise çok faydalı olabilirler ve onları tutmak en iyi çözüm olacaktır. Bu nedenle böylesi değerlerin analizi uzman bilgisi ve tecrübesi gerektirmektedir. Aşırı/Aykırı değer tespit algoritmaları, güvenlik, iş ve endüstri alanlarından bahsedebileceğimiz birçok alanda yaygın olarak uyarlanmıştır. Örneğin, aşırı/aykırı değer tespiti, bir kredi kartının şüpheli işlemlerini tespit edebilir veya şüpheli ağ saldırıları tespit edilebilir. Tıp alanında aşırı/aykırı değer tespit algoritmalarını uygulayarak, doktorların kanser tümörlerinin erken gelişimini ve çok daha fazlasını tespit etmesi sağlanabilir (9). Aşırı/Aykırı değer tespit yöntemleri genel olarak dört grupta incelenebilir. Bunlar; istatistiksel tabanlı, uzaklık tabanlı, kümeleme tabanlı ve yoğunluk tabanlı tespit yöntemleridir. Şekil 2'de aşırı/aykırı değer yöntemlerine ilişkin genel bir sınıflandırma çizelgesi verilmiştir (9).



Şekil 2: Aşırı/Aykırı değer yöntemlerine ilişkin genel bir sınıflandırma (9)

Veri dönüşümü ve indirgeme, veri madenciliği algoritmalarının uygulanmasını optimize etmek, hesaplama sürelerini azaltmak ve/veya sonuçlarını iyileştirmek için gerekli bir adımdır. Değişken alt kümesinin seçimi, boyutsallık azaltma (bir dizi özelliğin daha küçük bir alana izdüşümü) veya örnek kümesinin değiştirilmesini (örnek ekleme veya çıkarma) içerir.

Bir veri seti, örnek sayısı ve örnek başına özellik sayısı açısından çok büyük olabilir. Bu bağlamda, veri madenciliği algoritmalarının öğrenme kapasitesini engelleyen ve fazlalık içeren alakasız değişkenlerin modelin performansı üzerinde etkisiz olduğu bilinmektedir.

En basit değişken azaltma yöntemi, "ilgisiz" değişkenlerin manuel olarak kaldırılmasıdır. Bu değişkenler sonraki VTBK aşamalarından hariç tutulacağından, alan hakkında derin bir uzmanlık bilgisi gerektirir. Birçok veri madenciliği algoritması bazı değişken türleriyle (kategorik, sayısal) çalışamaz. Bu sorunla başa çıkmak için

araştırmacılar önişleme yöntemlerini uygulamak yerine ilgili değişkenleri veri setinden çıkarmayı tercih etmektedir. Boyutsallığın azaltılmasında başvurulan yaklaşımlardan biri olan değişken seçim (feature selection) analizi temel olarak daha basit ve daha anlaşılır modeller oluşturma, veri madenciliği performansını iyileştirme ve temiz, anlaşılır veriler hazırlamayı hedefler (10). Genel olarak değişken seçim yöntemleri, filtre, sarmalayıcı (wrapper) ve gömülü (embedded) veya hibrit yöntemler olmak üzere üç kategoride sınıflandırılabilir.

Sarmalayıcı yaklaşımlar, belirli bir değişkenin dâhil edilmesinin tahmine dayalı bir modelin performansı üzerindeki faydasını ölçer. Bu durum, modelin tahmin kapasitesini en üst düzeye çıkaran değişken alt kümesini arayan bir optimizasyon prosedürü olarak görülebilir. Bu nedenle ve tahmine dayalı modelin kalitesini ölçmek için verilerin etiketlenmesini gerektirir. Yaygın olarak kullanılan bazı sarmalayıcı tabanlı değişken seçim yöntemleri, genetik algoritmalar ve parçacık sürüsü optimizasyonu (particle swarm optimization) gibi evrimsel algoritmalara dayanır. Sarmalayıcı yöntemlerinin avantajları, değişken alt küme aramaları ile tahminleyici seçimi arasındaki etkileşimi ve değişken bağımlılıklarını hesaba katma yeteneğini içerir. Bununla birlikte, sarmalayıcı yöntemler, tahmin ediciyi eğitmek ve test etmek belirli bir hesaplama maliyeti gerektirdiğinden genellikle hesaplama açısından pahalıdır (11).

Sarma yöntemlerinin aksine, filtre yöntemlerinin sonuçları, bir sonraki VTBK aşamasında kullanılan tahminleyici modele bağlı değildir. Değişkenlerin uygunluğu, korelasyon, karşılıklı bilgi (mutual information), tutarlılık, varyans vb. ölçütler kullanılarak hesaplanır. Normalde, filtre yöntemleri hesaplama açısından sarmalayıcı yöntemlerden daha kolaydır ve etiketlenmemiş veriler veya çok sayıda değişken içeren veri kümeleri için daha uygundur (12).

Son olarak, gömülü değişken seçimi yaklaşımları, model eğitim aşamasının bir parçası olarak tahmin algoritmalarına entegre edilmiş filtre veya sarmalayıcı yöntemlerdir. Bu yaklaşımlar, verilen bir öğrenme algoritmasına bağlı olmaları bakımından sarmalayıcı yaklaşımlara benzerler. Bununla birlikte, bu yöntemler sınıflandırıcı ile etkileşime girebilirken, sarmalayıcı yöntemlerden hesaplama açısından daha az yoğundur. Bu nedenle, gömülü yöntemlerin filtrelerin verimliliğini sarmalayıcıların doğruluğu ile birleştirmesi beklenir. C4.5 ve C5.0 karar ağacı modelleri ile LASSO (Least Absolute Shrinkage and Selection Operator) düzenlileştirme

(regularization) tekniği en bilinen gömülü değişken seçimi yöntemlerinden birkaçıdır (13).

Değişken izdüşümü teknikleri, değişken seçim yöntemlerinin aksine orijinal değişken kümesinin bir alt kümesini seçmez. Bunun yerine, değişkenlerin daha düşük boyutlu bir uzaya izdüşümünü elde etmek için kullanılırlar. Orijinal değişken vektörlerinin bilgilerinin çoğunu koruyan doğrusal ve doğrusal olmayan değişken kombinasyonlarını bulan istatistiksel ve matematiksel fonksiyonların kullanımına dayanan bir sıkıştırma şeklidir (14). Bu haliyle, değişken izdüşümü tekniklerinin aynı zamanda bir veri dönüşümü (transformation) tekniği olduğu da söylenebilir.

Temel Bileşen Analizi (Principal component analysis; PCA) en yaygın boyut azaltma prosedürlerinden biridir. PCA, değişkenleri doğrusal olarak ilişkisiz (uncorrelated) bileşenlerden oluşan bir uzaya eşlemek için orijinal açıklayıcı değişkenlere doğrusal bir dönüşüm uygular. Boyut azaltma, orijinal değişken sayısından daha küçük bir dizi bileşen seçilerek gerçekleştirilir. İdeal olarak, daha geniş değer aralıklarına sahip değişkenlere daha fazla ağırlık verilmesini önlemek için PCA uygulanmadan önce veriler ölçeklenmeli ve aşırı/aykırı değerlerden arındırılmalıdır. PCA'nın temel eksikliği, değişkenler arasındaki karmaşık, doğrusal olmayan ilişkileri yakalayamamasıdır (14, 15).

Özellik projeksiyonu için bir tür derin sinir ağı türü olan otokodlayıcılar (autoencoders) da önerilmektedir. Çıktı katmanında girdi değerlerini (orijinal değişkenleri) yeniden üretmeye çalışırlar. Boyut azaltma, girdinin ara katmanlarda daha küçük bir alana yansıtılmasıyla sağlanır. Otokodlayıcılar, girdi verilerindeki karmaşık ilişkileri modelleyebilir, doğrusal ilişkilerle sınırlı değildir (14, 16).

Veri madenciliği VTBK sürecinin özüdür, ancak önceki adımlar düzgün bir şekilde gerçekleştirilirse başarılı olabilir. Veri kümelerinde gizli kalmış örüntü ve ilişkileri bularak verilerden bilgi çıkarmak için makine öğrenmesi algoritmaları da dâhil olmak üzere uygun tekniklerin seçimi ve uygulamasından oluşur. Belirli bir veri madenciliği algoritmasının seçimi, eldeki probleme ve mevcut verilerin doğasına bağlıdır. Veri madenciliği özetle aşağıdaki algoritma türlerini içerir:

- Sınıflandırma algoritmaları, veri kümesindeki bağımsız değişkenlere dayalı olarak bir veya daha fazla kategorik değişkeni tahmin eder.
- Regresyon algoritmaları, veri kümesindeki bağımsız değişkenlere dayalı olarak bir veya daha fazla kesikli/ sürekli sayısal değişkeni tahmin eder.

- Kümeleme/Segmentasyon/Bölümleme algoritmaları, verileri benzer özelliklere sahip öğelerden oluşan gruplara veya kümelere böler.
- İlişkilendirme algoritmaları, bir veri kümesindeki farklı değişkenler arasındaki bağıntıları bulur. Bu tür bir algoritmaların en yaygın uygulaması, pazar sepeti (market basket) analizinde kullanılabilecek birliktelik kuralları oluşturmak içindir.

Yorumlama ve değerlendirme aşaması, bir dizi performans ölçütü kullanılarak, veri madenciliği veya makine öğrenmesi algoritmasının performansını ve etkinliğini ölçmekten oluşur. Ayrıca sonuç görselleştirme ve raporlamayı da içerir.

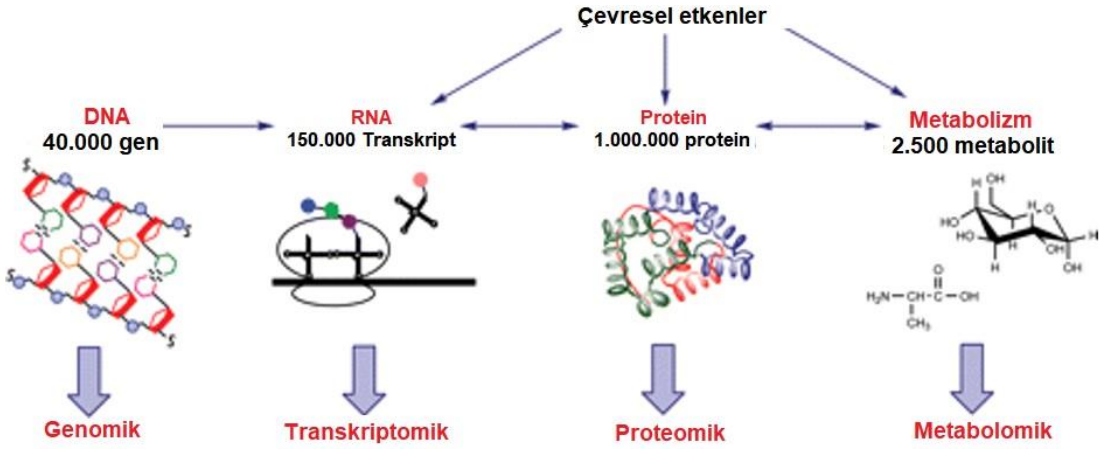
VTBK sürecinin yinelemeli doğası göz önüne alındığında, tatmin edici sonuçlar elde etmek için bazı adımların yeniden uygulanması ve sonuçların bir kereden fazla ele alınması gerekebilir. Örneğin, veri dönüştürme adımı sırasında başlangıçta hariç tutulan bir değişken, belirli bir veri madenciliği algoritmasının performansını artırmak için daha sonra eklenebilir (6, 7).

2.2. Omik Teknolojileri

Biyoloji bilimlerinde -omik soneki, büyük biyolojik molekül kümelerinin çalışmasına atıfta bulunmak için kullanılır. Moleküler biyoloji alanının, izole edilmiş biyolojik molekülleri incelemekten büyük biyolojik molekül kümelerinin geniş bir analizine doğru ilerlemesi gerektiği fikri, 2001 yılında insan genom projesinin (Human Genome Project, HGP) tamamlanmasıyla vurgulanmıştır. Omik alanı, genomik (genom odaklı) ile proteomik (büyük protein kümelerine, proteom odaklı) ve metabolomiklere (büyük küçük molekül kümelerine, metabolom odaklı) kadar uzanır (17, 18).

Omik teknolojileri, büyük biyolojik molekül kümelerini incelemek için uygulanabilen biyobelirteç keşif araçlarıdır. Omik tekniklerinin sayısı sürekli artmasına rağmen, en gelişmiş beş omik teknolojisi genomik, transkriptomik, epigenomik, proteomik ve metabolomiktir (18).

Şekil 3'de, bir organizmada biyolojik bilginin akışını tarif eden 'Omik' basamakları açıklanmaktadır (19).



Şekil 3: Bir organizmada biyolojik bilginin akışını tarif eden 'Omik' basamakları (19)

2.2.1. Genomik

En yaygın olarak kullanılan omik teknolojilerinden biri olan genomik, canlıların ve bazı virüslerin biyolojik gelişimleri için gereken genetik bilgiyi taşıyan DNA'nın yapı ve fonksiyonlarını detaylı olarak inceleyen bilim dalı şeklinde tanımlanmaktadır (20). Bir organizmanın genomundaki bütün genetik yapıların belirlenmesi, dizi analizinin yapılıp haritasının çıkarılması gibi DNA yapı ve işlevinin kapsamlı olarak incelenmesi genomik teknolojisi olarak açıklanmaktadır (21).

Bir genom, bir organizmanın tüm genleri de dâhil olmak üzere eksiksiz bir DNA setidir. Genomik ayrıca, tüm genomların işlevini ve yapısını bir araya getirmek ve analiz etmek için yüksek verimli DNA dizilimi ve biyoinformatik kullanımı yoluyla genomların dizilenmesini ve analizini içerir. Genomikteki ilerlemeler, beyin gibi en karmaşık biyolojik sistemlerin bile anlaşılmasını kolaylaştırmak için keşfe dayalı araştırma ve sistem biyolojisinde bir devrimi tetiklemiştir. Genomik teknolojisinin tıp alanındaki temel hedefi hastalık önleyicilik, tanı ve tedavinin bireye özgü düzenlenmesidir (22-24).

Genomik kısaca özetlenecek olursa (25);

- Hastalıkla ilgili genleri tanımlamaya,
- Genler arasındaki etkileşimlerin ortaya çıkartılmasına,
- Genlerin ekspresyon profillerinin ortaya çıkartılmasına,
- Farklı organizmaların genetik zeminde karşılaştırılmasına izin verir.

2.2.2. Transkriptomik

Transkriptom, bir hücre veya hücre popülasyonu tarafından üretilen mRNA moleküllerinin tamamlayıcısı olarak tanımlanır. Terim ilk olarak 1996'da Charles Auffray tarafından önerilmiştir (26). "-om" ekini edinen teknolojilerin çoğunun aksine, "Transkriptom" uzun bir soyağacına sahiptir ve gerçek bir omik teknolojisinin gereksinimlerini karşılar (27).

Bir organizmanın tüm RNA içeriğinin incelenmesi transkriptomiktir. DNA'da bulunan bilgiler, hücrenin o andaki aktivitesinin anlık görüntüsünü temsil eden transkripsiyon yoluyla ifade edilir.

Transkriptomik, post-genomik çağda en gelişmiş alanlardan biridir. Transkriptom, haberci RNA, transfer RNA, ribozomal RNA ve diğer kodlamayan RNA'lar dâhil olmak üzere belirli bir gelişim aşamasında ve/veya belirli bir fizyolojik koşul altında belirli bir hücre tipi veya dokusundaki RNA transkriptlerinin eksiksiz setidir. Transkriptomik, RNA düzeyinde gen ekspresyonuna odaklanır ve özellikli biyolojik süreçlerde yer alan moleküler mekanizmaları ortaya çıkarmak için gen yapısı ve gen işlevi hakkında genom çapında bilgi sunar. Yeni nesil yüksek verimli dizileme teknolojisinin gelişmesiyle birlikte, transkriptom analizi, RNA tabanlı gen düzenleyici ağ anlayışımızı aşamalı olarak geliştirmektedir (28).

Modern transkriptomik, farklı fizyolojik veya patolojik koşullarda çoklu transkriptlerin ifadesini analiz etmek için yüksek verimli yöntemler kullanır ve bu durum, transkriptom ve fenotip arasındaki ilişkileri geniş bir canlı varlık yelpazesinde hızla genişletmektedir (29).

2.2.3. Proteomik

Proteomik, proteinlerin sistematik ve büyük ölçekli analizidir. Belirli bir hücre veya organizma tarafından tanımlanmış bir dizi koşul altında üretilen eksiksiz bir protein seti olarak olarak tanımlanan proteom kavramına dayanmaktadır. Proteinler hemen hemen her biyolojik süreçte doğrudan yer alır, bu nedenle hücredeki proteinlerin kapsamlı analizi, bu moleküllerin nasıl etkileşime girdiğine ve çalışan bir biyolojik sistem oluşturmak ve sürdürmek için nasıl işbirliği yaptığına dair benzersiz bir küresel bakış açısı sağlar. Hücre, proteinlerinin seviyesini ve aktivitesini düzenleyerek iç ve dış

değişikliklere yanıt verir, bu nedenle proteomdaki nitel veya nicel değişiklikler, bu düzenleyici ağın eylem halindeki bir anlık görüntüsünü sağlar (30).

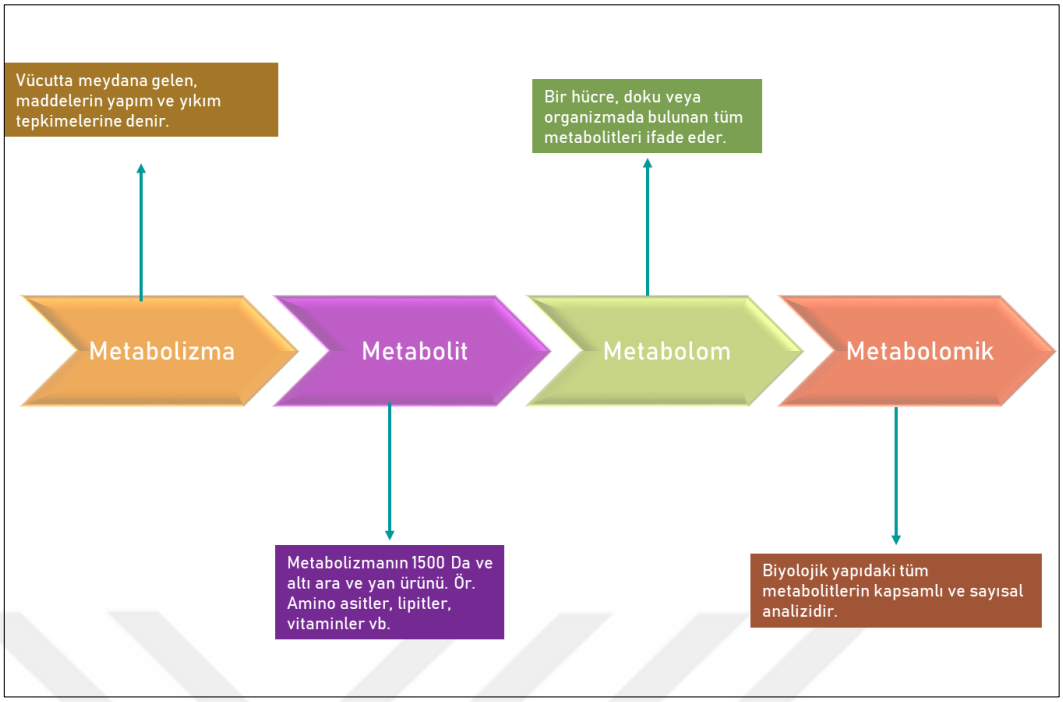
Proteinler, herhangi bir biyolojik sistemdeki birincil işlevsel varlıklardır. Bu nedenle, nasıl işlediklerini ve biyoloji ve tıpta oynadıkları rolleri anlama arayışı, biyokimyanın önemli bir meşguliyeti olmuştur. Fonksiyonu anlamak için temel bir ön koşul, herhangi bir proteini, hatalı iddialardan ve sonuçlardan kaçınmak için doğru ve yeterince sıkı bir şekilde tanımlama yeteneğidir. Proteinler, genomik ebeveyn benzerlerinin DNA'sından farklı olarak, yalnızca doğrusal bir aminoasit dizisi içermez, aynı zamanda 28'den fazla bilinen translasyon sonrası modifikasyondan etkilenen sayısız şekle katlanırlar ve her biri onlarca potansiyele sahiptir. Bu ve diğer komplikasyonlar, kütle spektrometrisine dayalı proteomiklerin ortaya çıkmasına kadar proteinlerin tanımlanmasını özellikle zorlaştırmıştır (31).

Proteomik, bir hücre veya biyolojik bir örnekteki proteinlerin ifade (ekspresyon) seviyelerini, translasyon sonrası modifikasyonlarını veya lokalizasyonunu ölçmek için kullanılabilen bir dizi teknik içerir (32).

2.2.4. Metabolomik

Metabolomik veya metabolom analizi, hücre içi metabolitlerin eş zamanlı olarak belirlenmesini ve nicel analizini gerçekleştirmeyi amaçlar. Metabolomik, hücresel aktivitenin substratları ve ürünleri olan küçük moleküller ile ilgilendiğinden, biyolojik sistem/çevre arayüzünün (fenotipin) doğrudan ve anında araştırılmasına izin verir Metabolomik, çoklu biyolojik düzeyde toplanan bilgileri entegre etmeyi amaçlayan bir alan olan sistem biyolojisinde giderek daha önemli bir rol oynamaktadır. Şu anda mikrobiyoloji, tanısal biyobelirteç keşfi, toksikolojik test, yiyecek ve içecek analizi, bitki ve hayvan fenotiplemesi ve ilaç keşfi ve geliştirme dâhil birçok uygulamada yaygın olarak kullanılmaktadır (33). Metabolomik analizlerini gerçekleştirmek için kullanılan iki analitik teknik, kütle spektrometresi (Mass spectrometry, MS) ve nükleer manyetik rezonans (Nuclear magnetic resonance, NMR) spektroskopisidir. Her iki teknik de tek bir ölçümde birçok farklı molekül hakkında bilgi verir ve metabolitlerin yapılarını ve konsantrasyonlarını belirlemek için kullanılabilir (34).

Metabolizmadan metabolomiğe olan akışı gösteren bir şema Şekil 4'de verilmiştir.



Şekil 4: Metabolizmadan metabolomiğe olan akışı gösteren bir şema

Metabolomik, hücrelerde, biyolojik sıvılarda, dokularda veya organizmalarda bulunan metabolitlerin tanımlanması ve miktarının belirlenmesine yönelik geniş ölçekli bir bilimsel çalışmadır. Metabolitler, metabolizma sırasında kimyasal olarak değiştirilen ve hücrelerin büyümesi, devamlılığı ve normal fonksiyonu için gerekli olan düşük moleküler ağırlıklı küçük moleküllerdir. Metabolitler, hücresel durumun işlevsel bir okumasını verir ve amino asitler, yağ asitleri, organik asitler, şekerler vb. olabilir. Bir biyolojik numunede metabolik tepkimeye giren metabolitlerin toplam nicel koleksiyonu metabolom olarak bilinir. Metabolom, homeostaz değişiminin doğrudan bir göstergesi olarak düşünülür. Belirgin metabolitlerin oluşumundaki değişim, metabolik yolda bir değişikliği gösterir; bu nedenle, metabolomik, yol analizi için tutarlı bir yaklaşım sağlar. Genlerin ve proteinlerin işlevi sırasıyla epigenetik düzenleme ve translasyon sonrası modifikasyonla ilgilidir, ancak metabolitler biyolojik aktivitenin doğrudan bir izlenimini sağlar, bu nedenle biyolojik sistemin moleküler fenotipi ile ilişkilidir. Metabolit profili çıkarma, klinik teşhis için yaygın olarak kullanılan bir yaklaşım haline gelmiştir (35, 36).

Günümüzde teknolojideki gelişmeler, bir örnekte/numunede bulunan ve hücrelerin metabolik durumu hakkında ayrıntılı bilgi veren daha fazla sayıda metabolitin profilini çıkarmayı mümkün kılmıştır. Son on yılda, metabolomik alanı, biyokimyasal

değişimi fenotiple ilişkilendirerek mekanik bilgi veren yeni araçlar sunarak önemli ilerleme kaydetmiştir. Kütle spektroskopisinin detaylandırılması nedeniyle, küçük miktarlarda numune kullanarak binlerce metabolit aynı anda işlenebilir. Metabolomik deney tasarımı (numunenin hazırlanması ve cihazın seçimi) metabolitlerin toplamına ve kimyasal bileşimlerine bağlıdır. Bu teknolojiler, metabolitlerin kantifikasyonu ve tanımlanması için pik kalıpları veren spektrumlar sağlar. Bu modeller, spektral veri tabanları ve otomatik veri analizi yoluyla metabolomik profiller oluşturur. Metabolomik, yüksek verimli bir analitik yöntemdir, bu nedenle veri analizi, metabolomikte çok önemli bir adımdır (37, 38).

Metabolomik, büyük miktarda veri üretir. Yeni teknolojilerin ortaya çıkmasıyla birlikte veri yorumlama ana konu haline geliyor. Spektrometre teknolojisinin gelişmesiyle, biyokimyasal spektrumların yorumlanması için NMR spektrumlarının görünür analizi yeterli değildir. Bu büyük miktardaki verinin yorumlanması için örüntü tanıma ve istatistiksel yöntemler gereklidir. Bu yöntemler, verilerin karmaşıklığını azaltarak veri kümesindeki yerel örüntülerin özelliklerini verir. Metabolomik verilerin analizi, ham veri ön işleme istatistiksel analizini ve kalıbın tanınmasını içerir. İşlemenin ilk adımında gürültü giderme ve pik toplama, işlenen verilerin kalitesini artırır (39, 40).

Loke ve ark. (2018), KRK ve normal doku örneklerinin işlev ve taksonomisindeki farkı bulmak için LC-MS metabolomikleri ve 16S rRNA yeni nesil dizilimi kullandı ve mikrobiyom disbiyozunun kanser bölgelerinde bir varyasyon nedeni olabileceğini destekledi. Bu çalışma, bir metabolit olan S-adenosil- l - homosisteinin (SAH), normal dokulara kıyasla tümörde yüksek konsantrasyonda bulunduğunu ortaya koymaktadır (41).

KRK üzerine bir çalışmada, hedeflenmemiş yaklaşımın ardından LC-MS metabolomikleri kullanılarak hedeflenen metabolik profilleme ve ardından MarkerView yazılımı kullanılarak verilerin analizi sonucunda çapraz doğrulamada hem hedeflenen hem de hedeflenmeyen yaklaşımdaki sınıflandırma modeli % 97.2 başarı oranı vermiştir (42).

KRK için rapor edilen metabolit belirteçlerinin kısa bir özeti aşağıda gösterilmektedir (43).

Tablo 1: KRK için rapor edilen metabolit belirteçlerinin kısa bir özeti (43)

Metabolitler	Örnek	Teknoloji	Metabolit konsantrasyonu
Asetoasetat, glutamin, guanidoasetat, cis-akonitat, trans-akonitat ve homosistein	İdrar	^{1}H-NMR	Yüksek
Kreatinin, kolin, dimetil sülfon, asparagin, alanin, metilamin	İdrar	^{1}H-NMR	Düşük
Taurin, alanin, 3-aminoizobutirat ve valin	İdrar	^{1}H-NMR	Yüksek
Treonin, gliserol, hippurat, askorbat, kreatinin ve sitrat	İdrar	^{1}H-NMR	Düşük
Betain aldehit, N-metildietanolamin, Adenilosüksinat, İzovalerat, Valerat, N1-metil 2-piridon-5-karboksamid	Biyopsi	GC-MS ve UPLC-MS/MS	Yüksek
2-aminoadipat, Stearoil sfingomiyelin, 4-hidroksifenilpiruvat, Sorbitol, Alfa hidroksiizovalerat, Cys-gly, oksitlenmiş, Triptofilglisin, Deoksikolat, 7-ketodeoksikolat, Asparagin, Aspartilvalin, Aspartiltriptofan 6-fosfat, Glukoz 6-fosfat6	Biyopsi	GC-MS ve UPLC-MS/MS	Düşük
SCFA'lar (asetat, propiyonat ve butirat), glikoz ve glutamin	Dışkı	^{1}H-NMR	Düşük
Prolin, süksinat, izolösin, lösin, valin, alanin, glutamat, dimetilglisin ve laktat	Dışkı	^{1}H-NMR	Yüksek
KRK biyobelirteci (çoklu lojistik regresyon modeliyle seçilir): - 2-hidroksibutirat, aspartik asit, kynurenine ve sistamin	Serum	Yüksek GC/MS	Yüksek
İzolösin, 3-hidroksibutirat, laktat, asetat, glutamat, kolin, glisin, serin ve glikoz	Biyopsi	^{1}H-NMR	Yüksek
KRK biyobelirteci: Sfinganin, endokannabinoidler	Serum	LC-HRMS	Yüksek
3-hidroksibütirat, asetat, format, gliserol, lipid (-CH2-OCOR), glikoproteinlerin N-asetil sinyali, fenilalanin ve prolin	Biyopsi	^{1}H-NMR	Yüksek

3. MATERYAL VE METOT

3.1. Veri seti

Bu tez çalışmasında, Washington Üniversitesi, Anesteziyoloji ve Algoloji Bölümü, Northwest Metabolomik Araştırma Merkezi'nde yürütülen PR000226 numaralı proje kapsamında üretilen veri seti (44) kullanılmıştır. İlgili veri seti, açık erişimli metabolomik veri setlerini bünyesinde barındıran Metabolomics Workbench (45) isimli veri tabanından indirilmiştir. Orijinal veri seti, potansiyel önemi olan 113 metabolit değişken (bundan sonra sadece değişken olarak adlandırılacaktır) ile üç denek grubundan (66 KRK hastası, 76 polip hastası ve 92 sağlıklı kontrol) toplam 234 serum örneğini içermektedir. Fakat bu tez çalışmasında, yalnızca KRK hastalığına ilişkin biyobelirteçlerin tespit edilmesine imkân tanıyan bir karar destek sisteminin oluşturulması amaçlandığından ilgili veri setini oluşturan yalnızca iki denek grubu (66 KRK hastası ve 92 sağlıklı kontrol olmak üzere toplam 158 örnek) dâhil edilmiştir.

Çalışmada, KRK hastalığının tahmini ve olası biyobelirteçlerin tespiti amaçlı VTBK tabanlı bir ardışık kod dizini (AKD, pipeline) oluşturulmuştur. İlgili AKD şu aşamaları içermektedir:

1. Veri dosyasının R yazılımına yüklenmesi,
2. Aşırı, aykırı ve/veya gürültülü verilerin tespiti ve analizi,
3. Kayıp değer analizi. → Random Forest tabanlı kayıp değer atama yöntemi kullanılmıştır.
4. Değişken seçimi → 5 farklı değişken seçim yöntemi/yaklaşımı kullanılmıştır.
5. Model eğitim süreçleri → Çeşitli topluluk (ensemble) ve derin öğrenme modelleri kullanılmıştır.
6. Çıktıların değerlendirilmesi ve yorumlanması → Çeşitli performans ölçütleri kullanılarak sonuçların değerlendirilmesi ve yorumlanması gerçekleştirilmiştir.

AKD'nin oluşturulmasında R programlama dili kullanılmış, gerekli yerlerde Python programlama dilinde yazılmış bazı programlar da kullanılmıştır.

3.2. Önişleme Aşamaları

3.2.1. Aşırı/Aykırı Değer Analizi

Aşırı/Aykırı değerler, diğer veri noktalarından uzak olan veri noktalarıdır. Başka bir deyişle, bir veri kümesindeki olağandışı değerlerdir. Aykırı değerler, testlerin önemli bulguları kaçırmasına veya gerçek sonuçları çarpıtmasına neden olabileceğinden, birçok istatistiksel analiz için sorunludur.

Aşırı/Aykırı değerleri kesin olarak belirlemek için katı istatistiksel kurallar yoktur. Aykırı değerlerin bulunması, konu alanı bilgisine ve veri toplama sürecinin anlaşılmasına bağlıdır. Kesin bir matematiksel tanım olmasa da, aykırı adayları bulmak için kullanabileceğiniz kılavuzlar ve istatistiksel testler vardır. Bu tez çalışmasında değişken bazında Tukey'in aşırı/aykırı değer testi kullanılmıştır.

3.2.2. Kayıp Değer Analizi

Rubin, 2004'de yayınladığı bir çalışmada (46) kayıp veriler üç kategoride tanımlamıştır:

- Eksik değerlerin gözlenen verilerden bağımsız olduğu tamamen rastgele kayıp (Missing Completely at Random, MCAR),
- Kayıp değerlerin yalnızca gözlemlenen verilere bağlı olduğu ve gözlemlenmemiş verilerden koşullu olarak bağımsız olduğu rastgele kayıp (Missing at Random, MAR),
- Rasgele olmayan kayıp (Missing not at Random, MNAR). Aynı zamanda gözlenen ve gözlemlenmeyen verilere bağlı olan eksik veri kalıplarının göz ardı edilemeyen eksik veri veya yapısal eksik veri olarak da bilinir.

Genellikle, kayıp değer sorunu çözülebilir kılmak için MCAR veya MAR ile çalışıldığı varsayılır (47). Bu tez çalışmasında da kayıp değerler için benzer durum varsayılmıştır.

Bu tez çalışmasında, kayıp değerlerin yerine değer ataması süreci, missForest (48) R yazılım kütüphanesi ile gerçekleştirilmiştir. missForest, temel olarak her türlü veri için kullanılabilen parametrik olmayan bir atama yöntemidir. Karma (nicel-nitel) tipte değişkenler, değişkenler arası doğrusal olmayan ilişkiler, karmaşık etkileşimler ve yüksek

boyutluluk (p≫n) ile baş edebilir. Söz konusu algoritma Random Forest tekniğine dayalı çalışır ve bunun için bir R yazılımdaki randomForest (49) kütüphanesini kullanılır.

3.3. Değişken seçimi

3.3.1. LASSO (Least Absolute Shrinkage and Selection Operator)

1996 yılında Robert Tibshirani tarafından geliştirilen LASSO (50), en küçük kareler yönteminin amaç fonksiyonuna L_1 normlu bir ceza terimi ekler. Buna göre LASSO tahmin edicisi,

$$\hat{\beta}_{LASSO} = min\|y - X\beta\|_2^2 + \lambda\|\beta\|_1$$

denklemi ile belirtilir. Burada,

$$\|\beta\|_1 = \sum_{i=1}^{p} |\beta_i|$$

dir. LASSO, tanımından da anlaşılacağı üzere değişken seçimi yapmaktadır ve bu özelliği onu değişken sayısının çok olduğu biyolojik veri setlerinin analizinde en çok tercih edilen değişken seçim yöntemlerinden birisi yapmıştır. Ancak LASSO yüksek boyutlu veri setleri için yararlı olsa da, çoklu bağlantı varlığında çok tavsiye edilmemektedir (51). Bu tez çalışmasında LASSO değişken seçimi yöntemi, R programlama dili tabanlı glmnet (52) kütüphanesinden faydalanılarak uygulanmıştır.

3.3.2. Elastic-Net

Ridge ve LASSO regresyon yöntemlerinin melezlenmesi olarak ifade edilebilecek bu yöntem, 2005 yılında Zou ve Hastie (53) tarafından önerilmiştir. Elastic-net tahmin edicisi,

$$\hat{\beta}_{e-net} = min\|y - X\beta\|_2^2 + \lambda[(1-\alpha)\|\beta\|_2^2 + \alpha\|\beta\|_1]$$

denklemi ile belirtilir. Burada λ büzülme (shrinkage) parametresini ifade etmektedir. α bir cezalandırma parametresi olup, α=1 olduğunda LASSO tahmin edicisi, α=0 seçildiğinde ise Ridge tahmin edicisi elde edilir. Elastic-net tahmin edicisi için α'nın 0 ile 1 arasında (0 ve 1'den farklı) değerler alması gereklidir. Bu durum Elastic-net yönteminin, hem belli oranda çoklu bağlantı problemini çözdüğünü hem de değişken seçimi görevi üstlendiğini göstermektedir. λ ve α parametrelerinin optimizasyonu için

genellikle k-katlı çapraz geçerlilik (k-fold cross-validation) tekniğinden faydalanılmaktadır (51). Bu tez çalışmasında Elastic-net değişken seçimi tekniğinin kullanılabilmesi için R programlama dili tabanlı glmnet (52) kütüphanesinden faydalanılmıştır. Parametre optimizasyonu için 5-katlı çapraz geçerlilik tekniği kullanılmıştır.

3.3.3. Boruta

İlk olarak Kursa ve Rudnicki tarafından 2010 yılında duyurulan Boruta (54), Random Forest (55) tekniğinin temel çalışma prensibine sahip bir değişken seçim algoritmasıdır. Boruta aşağıdaki adımlardan oluşur (56):

1. Orijinal veri setinin kopyaları oluşturulur ve bu kopya veri setlerindeki değişken değerleri (gölge değişkenler olarak adlandırılır) karıştırılır ve orijinal ve karıştırılmış veriler, değişken önemliliğini ölçen bir RF modelini eğitmek için birleştirilir.
2. Her değişken için Z skoru hesaplanır. Z skoru, Random Forest modelinden hesaplanan değişken önemlilik değerlerinin standardize edilmiş halidir.
3. Gölge değişkenler içinde en yüksek Z skoruna sahip olanı belirlenir (Bunu maxZF olarak adlandıralım).
4. Z skoru maxZF'den büyük olan orijinal değişkenler önemli, Z skoru maxZF'den küçük olan değişkenler ise önemsiz olarak etiketlenir.
5. Tüm değişkenler etiketlenene kadar yukarıdaki işlemler tekrarlanır.

Boruta tekniği ile değişken seçim işlemi, R yazılımı için oluşturulmuş Boruta (54) kütüphanesi ile gerçekleştirilmiştir. Boruta değişken seçimi yönteminde değişken önemliliklerini hesaplamak için, ortalama doğruluk düşüşü (mean decrease accuracy) değerlerinin standardize edilmiş istatistikleri kullanılmıştır.

3.3.4. BorutaShap

BorutaShap (57), Boruta değişken seçim algoritmasını, Shapley değerleriyle birleştiren bir sarmalayıcı (wrapper) değişken seçim yöntemidir. Oyun teorisi alanında bir çözüm yaklaşımı olan Shapley değeri, son yıllarda, uygun makine öğrenmesi modellerine açıklanabilir/yorumlanabilir özellik kazandırılmasında kullanılan, açıklanabilir/yorumlanabilir yapay zekâ alanında popüler bir yöntemdir (58). Burada "Shap" Shapley Additive ExPlanations ifadesinin kısaltmasıdır. Shap (59), Lundberg ve Lee tarafından 2016 yılında ilgili makine öğrenmesi modelinin bireysel tahminlerini açıklamaya yönelik geliştirilmiş bir yöntemdir. Daha açık bir ifadeyle bu yöntemin temel amacı, her bir değişkenin model tahminine katkısını hesaplayarak, ilgili modelin, veri setindeki bir x örneğinin tahminini açıklamaktır (60). Bu değişken seçimi yönteminin uygulanabilmesi için Python programlama dilinde oluşturulmuş BorutaShap (57) kütüphanesi kullanılmıştır. Bu kütüphanenin R yazılımında kullanılabilmesi için reticulate (61) kütüphanesi kullanılmıştır. Bu kütüphane, R oturumuna bir Python oturumu yerleştirerek sorunsuz, yüksek performanslı birlikte çalışabilirlik sağlayan faydalı bir yazılımdır. BorutaShap yönteminin parametrelerinden olan yineleme (iterasyon) sayısı 200 olarak belirlenmiştir.

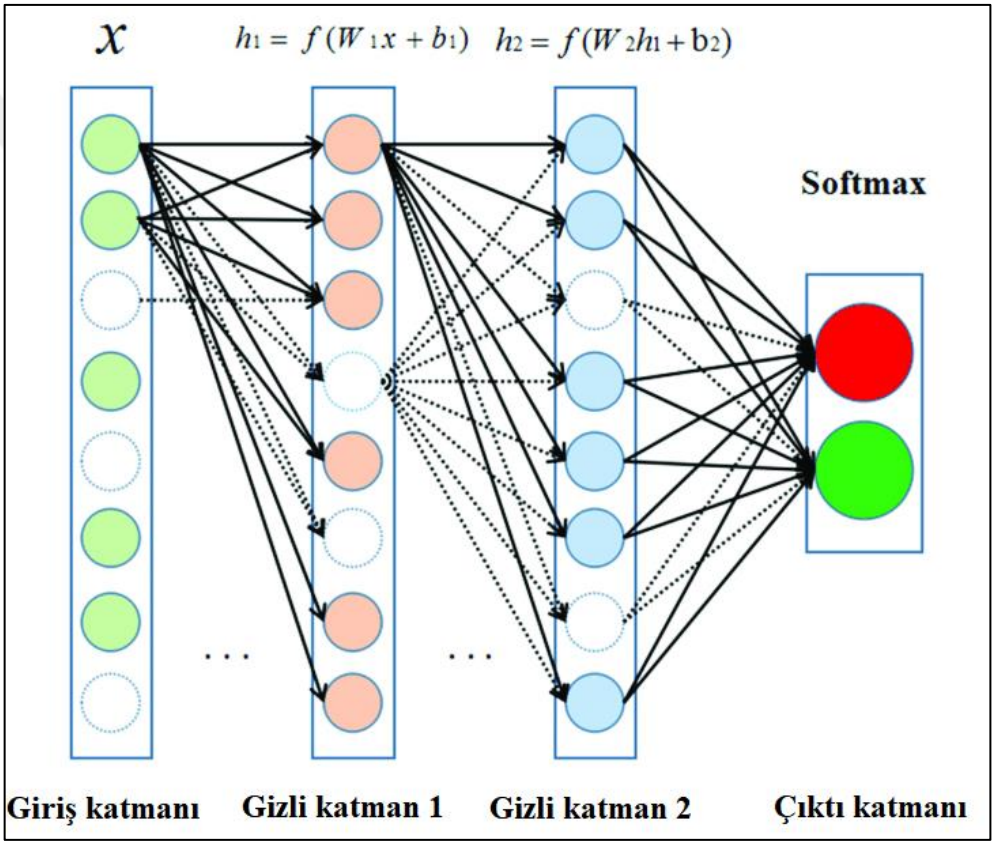
3.3.5. Uzlaşmacı (Konsensüs) Değişken Seçimi

Her değişken seçim yöntemi çeşitli kıstaslar ışığında, kendine göre en önemli değişkenleri seçer. Uzlaşmacı (Konsensüs) Değişken Seçim yöntemi, n adet bağımsız değişken seçim yönteminin, üzerinde uzlaşmaya vardığı seçilmiş değişkenleri seçer (62). Bir başka deyişle, n adet bağımsız değişken seçim yönteminin bireysel olarak seçtiği değişken kümelerinin kesişimini almaktadır. Buradan hareketle, ilgili yöntemin bir meta değişken seçim yöntemi olduğu söylenebilir. Bu kapsamda, Uzlaşmacı (Konsensüs) Değişken Seçimi, daha önce bahsedilen LASSO, Elastic-net, Boruta ve BorutaShap yöntemlerinin bir araya getirilmesi ile oluşturulmuştur.

3.4. Modelleme Aşaması

3.4.1. Derin Sinir Ağları

Bu tez çalışmasında, bir derin sinir ağı (DSA) modelinin eğitilebilmesi için Java tabanlı olan fakat R programlama dili için de uyarlaması mevcut olan H_2O (63) kütüphanesi kullanılmıştır. Bu kütüphanede kullanılan DSA, geri yayılım (back-propagation) kullanılarak stokastik gradyan inişi ile eğitilmiş çok katmanlı bir ileri beslemeli yapay sinir ağına dayanmaktadır. DSA'ya yönelik tipik bir mimari örneği Şekil 5'de verilmiştir (64).



Şekil 5: Tipik bir derin sinir ağı mimarisi (64)

Bu çalışmada oluşturulan DSA modeli, bir giriş katmanı, 200'er tane düğümden (node) oluşan 2 tane gizli (hidden) katman ve bir çıktı katmanından oluşmaktaydı. Aktivasyon fonksiyonu olarak ReLU (Rectified Linear Unit) kullanılmıştır. Öğrenme oranı (learning rate) hiperparametresi varsayılan başlangıç değeri olan 0.005 olarak alınmıştır. Hiperparametre optimizasyonu için 5-katlı çapraz geçerlilik tekniği kullanılmıştır.

3.4.2. Extreme Gradient Boosting (XGBoost)

Chen ve Guestrin tarafından 2016 yılında tanıtılan XGBoost (65), gradyan geliştirmede güçlü bir rol oynayabilen bir gradyan yükseltme ağacı (Gradient Boosting Trees, GBM) tabanlı bir algoritmadır. XGBoost algoritmasının, regresyon ve sınıflandırma problemleri için çok etkili bir yöntem olduğu belirtmiştir (66). Bu modelin uygulanmasında R programlama dilinde oluşturulmuş xgboost (67) kütüphanesi kullanılmıştır.

İlgili çalışmada oluşturulan xgboost modelinde, yükseltme (boosting) yineleme sayısı 1000 olarak belirlenmiştir. Öğrenme oranının kontrol eden eta parametresi ve maksimum ağaç derinliği varsayılan olarak sırasıyla 0.3 ve 6 olarak seçilmiştir.

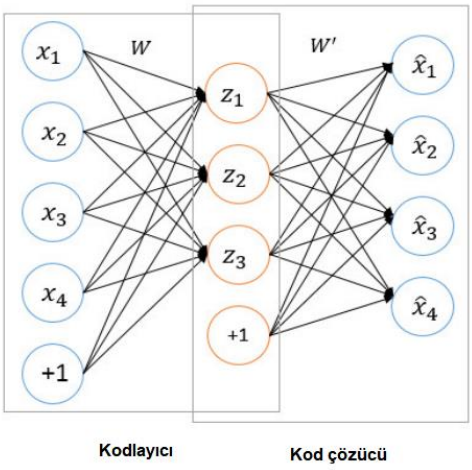
3.4.3. Light Gradient Boosting Machine (LightGBM)

Gradyan yükseltme karar ağaçları (Gradient Boosting Decision Tree, GBDT) ve XGBoost gibi bazı yükseltme algoritmaları, en iyi segmentasyon noktası elde edilirken her değişken için tüm örneklem noktalarının taranması gibi ortak bir zayıflığa sahiptir. Bu durum, çok zaman alıcı ve hesaplama açısından maliyetlidir. Model eğitiminde zaman ve hesaplama maliyetini azaltmak için LightGBM modeli önerilmiştir (68). LightGBM iki ana algoritma içerir: Gradyan Tabanlı Tek Yönlü Örnekleme (Gradient-Based One-Side Sampling, GOSS) ve Özel Özellik Paketleme (Exclusive Feature Bundling, EFB). LightGBM modeli hakkında daha detaylı bilgiler için (69) numaralı makaleden faydalanılabilir. İlgili model için R yazılımı için geliştirilen lightgbm (70) kütüphanesi kullanılmıştır. LightGBM modeli oluşturulurken yineleme (iterasyon) sayısı 1000 olarak belirlenmiştir.

3.4.4. Yığın (Stacked) Otokodlayıcı (Autoencoder)

Otokodlayıcı, çok boyutlu veriyi gizli katmanda sıkıştırdıktan sonra bu sıkıştırılan katmandan veriyi tekrar oluşturmak için kullanılan bir denetimsiz öğrenme algoritmasıdır. Otokodlayıcının temel amacı, veri boyutunun indirgenmesi, sıkıştırılması, birleştirilmesi ve daha birçok işlem için girdi verilerinin öğrenilmesi ve kodlanmasıdır. Bir otomatik kodlayıcı iki bölümden oluşur: kodlayıcı ve kod çözücü. Kodlama aşamasında girdi örnekleri genellikle daha düşük boyutlu değişken uzayına haritalanır.

Bu yaklaşım, istenen özellik boyutsal alana ulaşılıncaya kadar tekrar edilebilir. Kod çözme aşamasında, asıl değişkenler daha düşük boyutlu değişken uzayından ters işleme ile yeniden elde edilir (71). Şekil 6'da bir otokodlayıcı diyagramı verilmiştir.



Şekil 6: Örnek bir otokodlayıcı diyagramı (72)

Yığın otokodlayıcılar (YOK) ise otokodlayıcıların sınıflandırma görevleri için özelleştirilmiş halidir. YOK temel olarak üç adımdan oluşur (73):

1. İlk otokodlayıcı girdi verileriyle eğitilir ve öğrenilen öznitelik vektörü (feature vector) elde edilir,
2. Bir önceki katmanın öznitelik vektörü bir sonraki katman için girdi olarak kullanılır ve bu işlem eğitim tamamlanana kadar tekrarlanır,
3. Tüm gizli katmanlar eğitildikten sonra, maliyet fonksiyonunun minimizasyonu, parametre optimizasyonu ve etiketli eğitim seti ile ağırlıkları güncellemek için geri yayılım algoritması (Backpropagation) kullanılır.

YOK modelinin uygulanabilmesi için R programlama dilinde geliştirilmiş deepnet (74) kütüphanesi kullanılmıştır.

Bu tez çalışmasında modellerin öğrenme performanslarının test edilebilmesi için basit geçerlilik (split validation) yaklaşımı kullanılmıştır. Bu kapsamda, kullanılan veri seti %80 eğitim ve %20 test olmak üzere rasgele olarak ikiye ayrılmıştır.

3.4.5. Kullanılan Performans Ölçütleri

Bu tez çalışmasında, modellerin sınıflandırma performanslarını ölçmede kullanılan sınıflandırma matrisi, ölçütler ve bu ölçütlere ilişkin formüller Tablo 2-3'de verilmiştir.

Tablo 2: Sınıflandırma matrisi

	Referans	
Tahmin	**KRK**	**Sağlıklı kontrol**
KRK	DP	YP
Sağlıklı kontrol	YN	DN

Tablo 3: Kullanılan sınıflandırma performans ölçütleri

Ölçüt	**Formül/Yaklaşım**
Doğruluk	$\frac{DP + DN}{DP + DN + YN + YP}$
Duyarlılık	$\frac{DP}{DP + YN}$
Seçicilik	$\frac{DN}{DN + YP}$
Pozitif tahmin değeri (PTD)	$\frac{DP}{DP + YP}$
Negatif tahmin değeri (NTD)	$\frac{DN}{DN + YN}$
F_1 skor	$\frac{2 * PTD * Duyarlılık}{PT + Duyarlılık}$
ROC eğrisi altında kalan alan (EAKA)	Yamuk (Trapezoidal) kuralı*

*: EAKA'nın hesaplanmasında kullanılan çeşitli yöntemler mevcut olup, bu tez çalışmasında kullanılan yöntem belirtilmiştir.

Performans ölçütlerine ilişkin güven aralıklarının hesaplanmasında parametrik olmayan bootstrap örnekleme tekniği kullanılmıştır. Bootstrap yönteminin temel fikri, örnek verilerden bir anakütle hakkında çıkarımın, örnek verileri yeniden örnekleyerek ve yeniden örneklenen verilerden bir örnek hakkında çıkarım gerçekleştirerek modellenebilmesidir. Anakütle bilinmediği için, bir örneklem istatistiğindeki anakütle değerine karşı gerçek hata bilinmemektedir. Bootstrap örneklerinde, anakütle aslında örneklemdir ve bu bilinir; dolayısıyla yeniden örneklenen "gerçek" örneğin çıkarım

kalitesi ölçülebilir. Bootstrap güven aralıklarının hesaplanmasında R programlama dilinde geliştirilmiş boot (75) kütüphanesi ve bu kütüphanede tanımlanmış boot ve boot.ci fonksiyonları kullanılmıştır. Güven düzeyi %95, yeniden örnekleme tekrar sayısı ise 1000 olarak belirlenmiştir.

3.4.6. Ardışık Kod Dizini (Pipeline) Tasarımı

KRK hastalığına ilişkin alt grupların metabolomik veriler kullanılarak sınıflandırılabilmesi için topluluk öğrenme ve derin öğrenme mimarileri kullanılmıştır. Değişken seçim (feature/attribute selection) işlemlerinde ise LASSO, Elastic-Net, Boruta ve BorutaShap gibi gömülü (embedded) ve sarmalayıcı (wrapper) yöntemler kullanılmıştır.

Bu tez çalışmasının önemli hedeflerinden birisi, topluluk öğrenme ve derin öğrenme algoritmalarının sınıflandırma tahminlerini vererek yüksek çıktılı bir karar destek sistemi oluşturmaktır. Bu çerçevede, topluluk öğrenme yöntemleri, varyansı azaltmak, yanlılığı düşürmek veya tahminleri iyileştirmek amacıyla önceden açıklanan derin öğrenme mimarilerini bir tahmin modelinde birleştirmek için kullanılmış ve en uygun model tespit edilmiştir. Ardışık kod dizini R programlama dili kullanılarak oluşturulmuş ve tez ekinde paylaşılmıştır.

4. BULGULAR

Bu tez çalışmasında kullanılan veri seti, 77'si (%48.7) kadın, 81'i (%52.3) erkek olmak üzere toplam 158 bireyden oluşmaktadır. Bu 158 bireye ilişkin genel yaş ortalaması ve standart sapması ise 55.9±13.6 olarak hesaplanmıştır. Yaş ve cinsiyet demografik değişkenlerinin gruplara göre dağılımları Tablo 4-5'de verilmiştir.

Tablo 4: Bireylerin gruplara göre yaş dağılımına ilişkin tanımlayıcı istatistikler

Değişken	**KRK (*n*=66)**	**Sağlıklı kontrol (*n*=92)**	***p****
Yaş			
Art.ort±std.sapma	57.92±13.72	54.45±13.4	0.202
Ortanca (min. - maks.)	58.5 (27-88)	56.5 (18-80)	

*: Mann-Whitney U testi.

Tablo 5: Bireylerin gruplara göre cinsiyet dağılımına ilişkin tanımlayıcı istatistikler

Cinsiyet (n (%))	**Grup (n (%))**		**Toplam**	***p****
	KRK	**Sağlıklı kontrol**		
Kadın	30 (39)	47 (61)	77 (100)	
Erkek	36 (44.4)	45 (56.6)	81 (100)	0.485
Toplam	66 (41.8)	92 (58.2)	158 (100)	

*: Pearson ki-kare testi.

Veri setindeki metabolit değişkenlerinin, KRK ve sağlıklı kontrol grupları açısından yoğunluk (intensity) değerlerine ilişkin aritmetik ortalama ve standart sapma değerleri ile bağımsız örneklemlerde t-testi ile hesaplanan *p* değerleri Tablo 6'da verilmiştir. Bu bulgulara göre, 2'-Deoxyuridine, Adenylosuccinate, Allantoin, Alpha-Ketoglutaric Acid, Arginine, Aspartic Acid, Creatinine, Dimethylglycine, Epinephrine, Erythrose, Fumaric, G16BP, gama-Aminobutyrate, Glucose, Glutamic acid, Glutamine, Glyceraldehyde, Histidine, Homogentisate, Hydroxyproline/Aminolevulinate, Hyppuric Acid, IMP, Kynorenate, lactate, Linoleic Acid, Linolenic Acid, Lysine, Maleic Acid, Margaric Acid, Methionine, N-AcetylGlycine, OH-Phenylpyruvate, Oxalic Acid, PEP, Proline, Trimethylamine-N-oxide ve Urate değişkenleri açısından KRK ve sağlıklı kontrol grupları arasında istatistiksel olarak anlamlı farklılık bulunmaktaydı ($p<0.05$).

Tablo 6: Veri setindeki metabolit değişkenlerinin, KRK ve sağlıklı kontrol grupları açısından yoğunluk (intensity) değerlerine ilişkin aritmetik ortalama ve standart sapma değerleri ile *p* değerleri

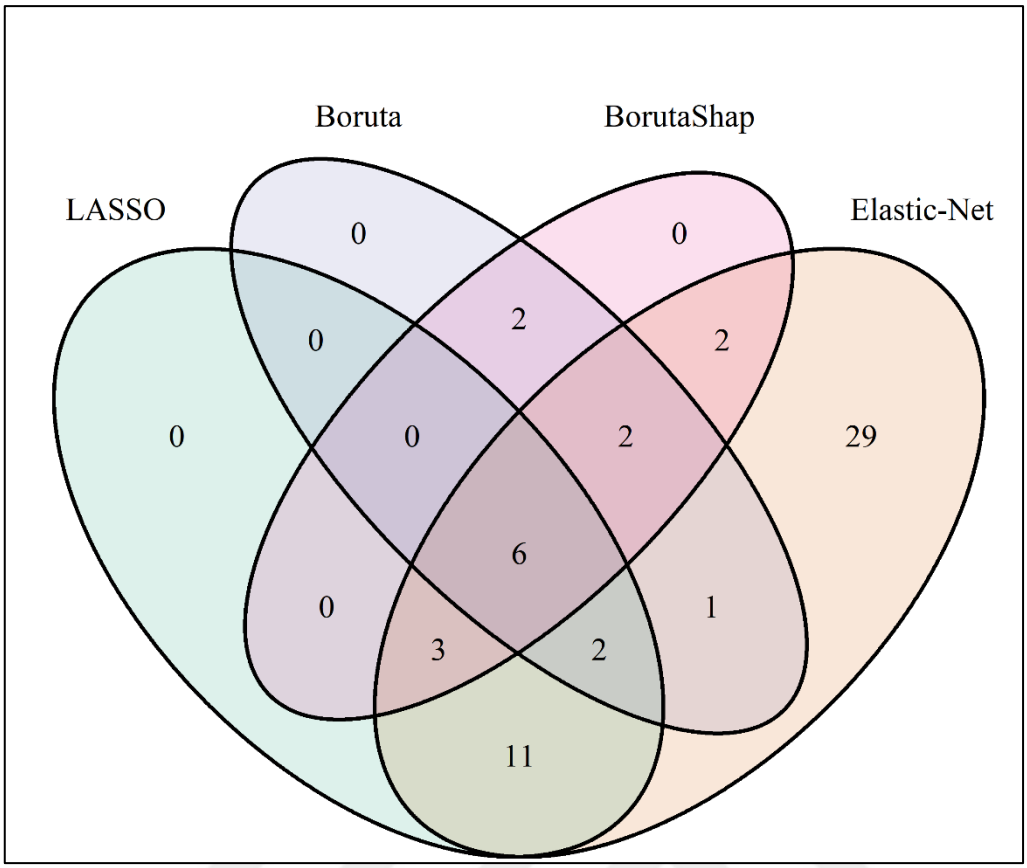
Metabolit (Yoğunluk x 10^{-3})	Grup		*p**
	KRK (n=66)	Sağlıklı kontrol (n=92)	
1-Methyladenosine	439.92±93.06	423.39±153.73	0.44
1-Methylhistamine	30.84±8.93	30.66±7.69	0.89
2-Aminoadipate	206.86±117.69	239.91±147.22	0.13
2'-Deoxyuridine	11.35±1.77	12.51±2.08	**<0.01**
3-Nitro-tyrosine	86.82±19.31	85.48±16.35	0.64
4-Pyridoxic acid	366.38±31.51	363.08±20.98	0.43
5-Hydroxytryptophan	79.69±6.68	79.15±7.79	0.65
Acetoacetate	1016.26±1380.65	807.88±957.1	0.26
Acetylcholine	1755.3±610.2	1868.9±611.78	0.25
Aconitate	531.42±186.03	587.69±270.05	0.15
Adenosine	44.87±82.37	81.86±178.1	0.12
Adenylosuccinate	58.94±22.83	48.78±20.29	**<0.01**
Adipic Acid	242.18±265.14	378.9±903.91	0.24
Alanine	5640.81±1378.35	6001.05±1204.94	0.08
Allantoin	120.44±58.39	97.1±61.11	**0.02**
Alpha-Ketoglutaric Acid	1639.28±253.66	1765.81±227.53	**<0.01**
Aminoisobutyrate	48.85±9.91	48.13±9.31	0.64
AMP	221.84±20.54	221.33±18.76	0.87
Anthranilate	74.66±37.75	84.35±49.64	0.18
Arginine	15366.05±3181.8	14282.65±2421.95	**0.02**
Asparagine	656.88±125.58	679.2±106.69	0.23
Aspartic Acid	1636.52±862.62	1195.67±516.73	**<0.01**
Betaine	35402.74±8198.75	35608.63±7741.41	0.87
Biotin	123.13±31.52	126.04±32.03	0.57
Carnitine	95.01±16.51	97.06±16.89	0.45
Choline	9169.31±3099.39	8879.1±2765.05	0.54
Citraconic Acid	753.24±267.58	846.5±387.92	0.09
Citrulline	8854.62±2993.43	8916.1±2676.43	0.89
Creatine	9606.78±4960.46	8581.75±4415.92	0.17
Creatinine	9762.78±2174.5	10819.17±3449.77	**0.03**
Cystamine	171.64±226.94	146.77±230.47	0.5
Cystathionine	17.59±22.74	12.15±13.93	0.07
Cystine	211±28.34	206.58±29.15	0.34
Cytidine	25.43±13.08	26.39±16.75	0.7
D-GA3P/DHAP	966.03±64.44	963.4±64.97	0.8
Dimethylglycine	1068.28±455.93	1299.64±472.94	**<0.01**
D-Leucic Acid	104.89±206.81	72.9±109.05	0.21
DTMP	218.4±14.66	221.94±16.76	0.17
Epinephrine	374.36±70.68	352.31±53.64	**0.03**
Erythrose	140.24±31.88	129.27±31.76	**0.03**
F16BP/F26BP	1081.21±196.58	1064.98±213.75	0.63
Fructose	425.76±623.85	317.48±159.62	0.11
Fumaric	2128.87±532.82	1955.54±368.78	**0.02**
G16BP	619.44±106.82	587±88.47	**0.04**
gama-Aminobutyrate	24.96±6.04	26.97±4.68	**0.02**
Glucoronate	230.8±170.11	277.25±274.12	0.23
Glucose	73383.59±15906.74	68103.16±13117.53	**0.02**
Glutamic acid	2683.58±1133.66	2205.35±1301.58	**0.02**
Glutamine	29168.84±4169.19	31649.29±3941.78	**<0.01**
Glutaric Acid	435.08±130.89	425.71±170.52	0.71
Glyceraldehyde	76.66±47.85	57.37±15.54	**<0.01**
Glycerate	323.3±128.81	355±157.91	0.18
Glycerol-3-P	206.22±102.42	209.96±74.61	0.79
Glycine	209.54±72.71	207.02±66.18	0.82
Glycochenodeoxycholate	638.08±1031.78	448.59±1240.88	0.31
Glycocholate	152.36±402.97	88.72±202.68	0.19
Guanidinoacetate	42.33±11.01	42.56±10.57	0.9
Guanosine	33.38±36.49	37.97±42.66	0.48
Histidine	14905.49±3699.23	18327.21±4661.29	**<0.01**
Homogentisate	1123.67±208.89	1216.85±241.54	**0.01**
Homovanilate	841.94±181.2	806.22±174.63	0.21
Hydroxyproline/Aminolevulinate	1457.78±1527.51	1062.38±515.82	**0.02**
Hypoxanthine	1841.4±711.27	1878.54±649.06	0.73
Hyppuric Acid	685.5±1305.03	251±475.03	**<0.01**
IMP	93.9±39.2	141.22±169.83	**0.03**
Inosine	107.21±123.23	122.12±140.61	0.49

Kynorenate	56.64±18.93	48.92±14.54	**<0.01**
lactate	36183.81±9619.14	33067.98±7241.21	**0.02**
Leucine/iso-Leucine	24183.03±5297	24733.88±5007.14	0.51
Linoleic Acid	47.9±22.28	54.66±17.97	**0.04**
Linolenic Acid	662.33±292.03	851.73±262.83	**<0.01**
L-Kynurenine	74.75±28.18	67.16±26.47	0.09
Lysine	8703.9±1875.26	9875.25±1897.08	**<0.01**
Malate	5060.27±1582.41	5349.5±1859.18	0.31
Maleic Acid	2034.39±497.28	1796.02±339.37	**<0.01**
Malonic Acid/3HBA	19395.62±19824.65	24809.63±18633.27	0.08
Margaric Acid	56.89±11.45	61.86±16.65	**0.04**
Methionine	617.98±127.89	702.51±144.87	**<0.01**
MethylSuccinate	1207.7±242.23	1272.29±208.09	0.07
N2,N2-Dimethylguanosine	14.38±5.03	14.42±9.99	0.97
N-AcetylGlycine	383.15±263.89	509.94±323.18	**0.01**
Niacinamide	47.26±33.92	48.51±30.45	0.81
OH-Phenylpyruvate	78.05±18.49	72.59±15.73	0.05
Ornithine	3801.41±1164.11	3561.25±996.41	0.17
Orotate	171.77±456.8	101.31±94.75	0.15
Oxalic Acid	137.65±39.99	123.13±28.2	**0.01**
Oxaloacetate	408.92±196.69	413.96±203.09	0.88
Pentothenate	235.84±440.12	215.16±148.16	0.68
PEP	121.72±21.7	136.55±36.51	**<0.01**
Phenylalanine	10812.31±2751.76	10781.82±2139.75	0.94
Proline	33100.64±7493.93	29986.8±6573.07	**0.01**
Propionate	12.24±5.04	13.07±5.5	0.33
Pyridoxal-5-P	146.21±123.1	160.89±144.07	0.5
Pyroglutamic Acid	76.46±25.71	75.14±21.46	0.73
Pyruvate	199.2±85.01	173.11±81.43	0.05
Reduced glutatione	79.27±12.4	82.49±11.54	0.1
Salicylurate	28.73±9.76	26.63±8.95	0.17
Serine	2728.81±617.63	2726.29±500.82	0.98
Shikimic Acid	21118.84±21383.43	22856.57±21897.99	0.62
Sorbitol	85.99±22.12	88.28±21.62	0.52
Succinate/Methylmalonate	2229.77±519.47	2181.73±527.04	0.57
Taurine	1196.69±327.95	1182.67±272.96	0.77
Threonine	264.55±73.68	274.24±61.91	0.37
Trimethylamine-N-oxide	5911.84±5739.02	4149.55±3926.06	**0.02**
Tryptophan	3451.36±878.27	3586.35±1036.54	0.39
Tyrosine	2491.38±633.98	2603.41±637.87	0.28
Urate	54054±8471.01	57981.92±9961.75	**0.01**
Uridine	17.72±5.09	19.28±5.27	0.06
Valine	2720.05±671.34	2810.44±687.64	0.41
Xanthine	1779.69±2325.68	1569.01±446.98	0.4
Xanthosine	85.59±403.47	26.22±29.86	0.16
Xanthurenate	90.77±24.29	95.68±21.25	0.18

*: Bağımsız örneklemlerde t-testi.

Veri önişleme aşamasında ilk olarak veri setini oluşturan değişkenlere sıfır/sıfıra yakın varyanslı değişken taraması yapılmış, bu tarama sonunda veri setinden herhangi bir değişken çıkarılmamıştır.

Yapılan taramada 6 tane değişkenin 1 (bir) değerini içerdiği görülmüştür. Bu değişkenlerden 27 (%17) tane 1 (bir) değerini içeren değişken veri setinden çıkarılmıştır. Diğer değişkenlerdeki 1 değerleri ise veri setinden çıkarılarak yerlerine değer ataması yapılmıştır. Değer ataması, Random Forest algoritması tabanlı kayıp değer atama yöntemi ile gerçekleştirilmiştir. Kayıp değer ataması yapıldıktan sonra veri setine sırasıyla LASSO, Elastic-Net, Boruta ve BorutaShap değişken seçim yöntemleri uygulanmıştır. İlgili değişken seçim yöntemlerince seçilen değişkenleri gösterir venn şeması Şekil 7'de sunulmuştur.



Şekil 7: Farklı değişken seçim yöntemleri tarafından seçilen değişken sayıları

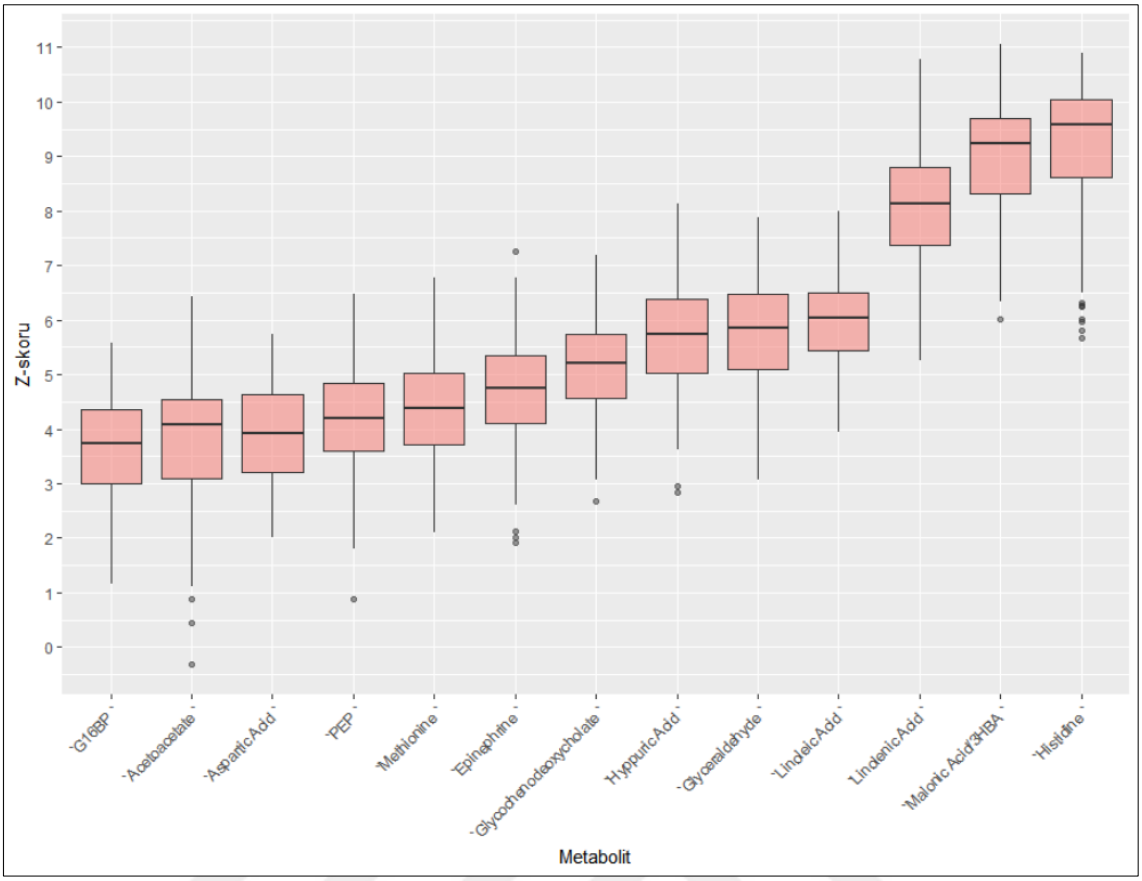
Değişken seçim yöntemlerince seçilen değişkenlerin sayısı Tablo 7'de verilmiştir.

Tablo 7: Değişken seçim yöntemlerince seçilen değişkenlerin sayısı

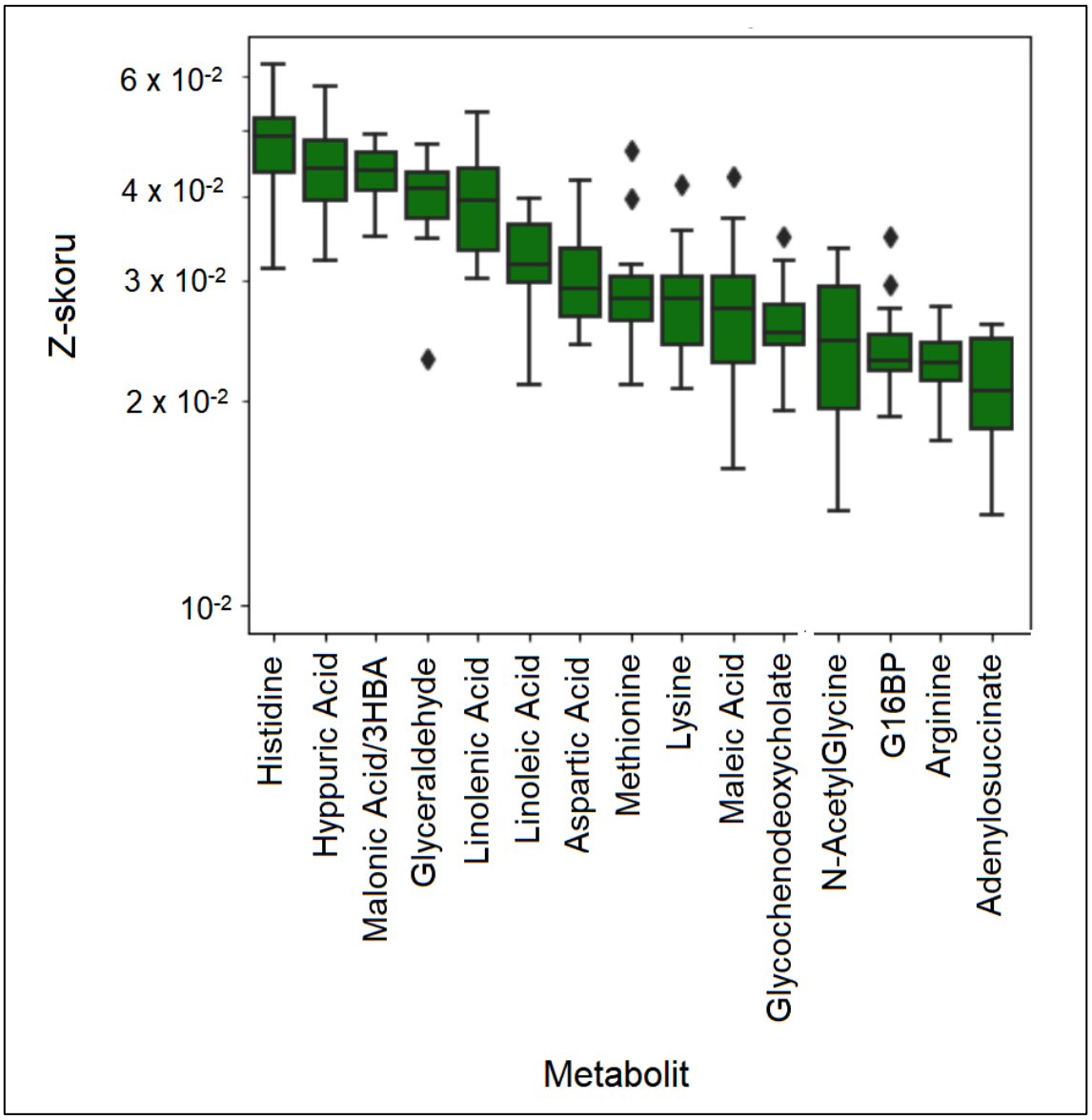
Yöntem	Seçilen değişken sayısı
Elastic-Net	56
LASSO	22
BorutaShap	15
Boruta	13
Uzlaşmacı (Konsensüs)	6

Değişken seçimi analizlerine göre, en çok değişken Elastic-Net yöntemi ile seçilirken, en az sayıda değişken ise Uzlaşmacı (Konsensüs) yöntemi ile seçilmiştir.

Boruta ve BorutaShap yöntemlerine ilişkin, her bir değişken için hesaplanan Z-skor tabanlı önemlilik değerlerinin kutu-çizgi grafik gösterimleri sırasıyla Şekil 8-9'da verilmiştir.



Şekil 8: Boruta yöntemine göre seçilen değişkenlere ilişkin önemlilik değerlerinin dağılımlarına ilişkin kutu-çizgi grafiği



Şekil 9: BorutaShap yöntemine göre seçilen değişkenlere ilişkin önemlilik değerlerinin dağılımlarına ilişkin kutu-çizgi grafiği

XGBoost, derin sinir ağları, LightGBM ve YOK modellerinin farklı değişken seçim yöntemlerinin sonuçlarına göre eğitim ve test veri setleri üzerindeki sınıflandırma performanslarını gösteren değerler, %95 güven aralıkları ile sırasıyla Tablo 8-13'de verilmiştir. Ayrıca, aynı modellerin, kullanılan farklı değişken seçim yöntemleri bazında, eğitim ve test veri setleri için çizilen ROC grafikleri sırasıyla Şekil 10-13'de sunulmuştur.

Tablo 8'de XGBoost modelinin eğitim veri seti üzerindeki sınıflandırma performansları incelendiğinde, değişken seçim yöntemleri uygulansa da uygulanmasa da doğruluk, duyarlılık, seçicilik, PTD, NTD, F_1 skor ve EAKA için sırasıyla, 1.00 (0.97 - 1.00), 1.00 (0.93 – 1.00), 1.00 (0.95 – 1.00), 1.00 (0.93 – 1.00), 1.00 (0.95 – 1.00), 1.00 (0.93 – 1.00) ve 1.00 (1.00 – 1.00) olarak hesaplandığı görülmüştür.

Tablo 9 incelendiğinde, XGBoost modelinin test veri seti üzerindeki en iyi sınıflandırma performanslarını, hiçbir değişken seçimi uygulanmadığı ve Boruta değişken seçimi uygulandıktan sonraki durumda elde ettiği görülmüştür. İlgili bulgular, hiçbir değişken seçimi uygulanmadığı durumda doğruluk, duyarlılık, seçicilik, PTD, NTD, F_1 skor ve EAKA için sırasıyla, 0.87 (0.70 - 0.96), 0.85 (0.55 - 0.98), 0.89 (0.65 - 0.99), 0.85 (0.55 - 0.98), 0.89 (0.65 - 0.99), 0.85 (0.72 - 1.00) ve 0.94 (0.85 – 1.00) olarak hesaplanmıştır. Boruta değişken seçim yöntemi uygulandıktan sonraki durum için sırasıyla, 0.87 (0.70 - 0.96), 0.77 (0.46 - 0.95), 0.94 (0.73 - 1.00), 0.91 (0.59 - 1.00), 0.85 (0.62 - 0.97), 0.83 (0.70 – 1.00) ve 0.91 (0.80 – 1.00) olarak hesaplanmıştır.

Tablo 10'da DSA modelinin eğitim veri seti üzerindeki sınıflandırma performansı ele alındığında, LASSO değişken seçimi yöntemi sonrası elde edilen performans metriklerinin doğruluk, duyarlılık, seçicilik, PTD, NTD, F_1 skor ve EAKA için sırasıyla, 0.93 (0.87 - 0.97), 0.92 (0.82 - 0.98), 0.93 (0.85 - 0.98), 0.91 (0.80 - 0.97), 0.95 (0.87 - 0.98), 0.92 (0.86 - 0.98) ve 0.93 (0.87 – 0.99) olarak hesaplanmıştır.

Tablo 11 incelendiğinde, DSA modelinin test veri seti üzerindeki en yüksek sınıflandırma performansı LASSO değişken seçimi yöntemi sonrası elde edilmiştir. İlgili sınıflandırma performans bulguları, doğruluk, duyarlılık, seçicilik, PTD, NTD, F_1 skor ve EAKA için sırasıyla, 0.87 (0.70 - 0.96), 0.85 (0.55 - 0.98), 0.89 (0.65 - 0.99), 0.85 (0.55 - 0.98), 0.89 (0.65 - 0.99), 0.85 (0.73 - 1.00) ve 0.85 (0.66 – 1.00) olarak elde edilmiştir.

LightGBM modelinin eğitim performansı Tablo 12 üzerinde incelendiğinde, Elastic-Net, Boruta ve BorutaShap değişken seçim yöntemleri sonrası sınıflandırma performans metriklerinin en yüksek değerleri aldığı görülmektedir. İlgili performans ölçütlerinin değerleri doğruluk, duyarlılık, seçicilik, PTD, NTD, F_1 skor ve EAKA için

sırasıyla, 1.00 (0.97 - 1.00), 1.00 (0.93 - 1.00), 1.00 (0.95 - 1.00), 1.00 (0.93 - 1.00), 1.00 (0.95 - 1.00), 1.00 (0.93 - 1.00) ve 1.00 (1.00 – 1.00) olarak hesaplanmıştır.

Tablo 13 incelendiğinde, LightGBM modelinin test performansı Boruta değişken seçim yöntemi sonrası doğruluk, duyarlılık, seçicilik, PTD, NTD, F_1 skor ve EAKA için sırasıyla, 0.90 (0.74 - 0.98), 0.85 (0.55 - 0.98), 0.94 (0.73 - 1.00), 0.92 (0.62 - 1.00), 0.89 (0.67 - 0.99), 0.88 (0.76-1.00) ve 0.88 (0.73 – 1.00) olarak elde edilmiştir.

Tablo 14 ve 15 birlikte incelendiğinde, YOK modeline ilişkin en iyi performans ölçütlerinin, test veri seti üzerinde hiçbir değişken seçim yönteminin uygulanmadığı durum için elde edildiği görülmüştür. İlgili durum için doğruluk, duyarlılık, seçicilik, PTD, NTD, F_1 skor ve EAKA değerleri sırasıyla, 0.58 (0.39 - 0.75), 0.00 (0.00 - 0.25), 1.00 (0.81 - 1.00), hesaplanamadı, 0.58 (0.39 - 0.75), hesaplanamadı, 0.57 (0.39 - 0.74) olarak elde edilmiştir.

Tablo 8: XGBoost modelinin eğitim performansı

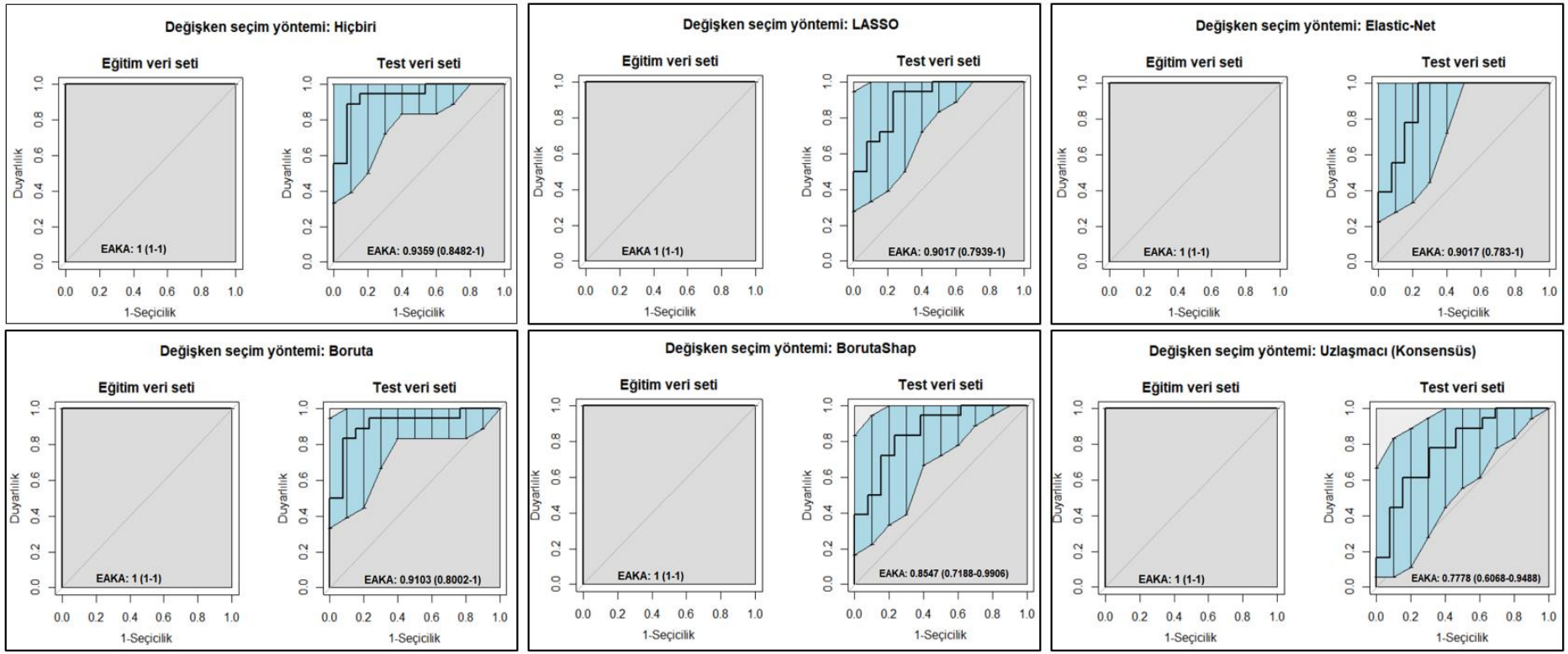
Değişken seçim yöntemi	Performans ölçütleri (%95 GA)						
	Doğruluk	**Duyarlılık**	**Seçicilik**	**PTD**	**NTD**	**F_1 skor**	**EAKA**
Hiçbiri	1.00 (0.97 - 1.00)	1.00 (0.93 – 1.00)	1.00 (0.95 – 1.00)	1.00 (0.93 – 1.00)	1.00 (0.95 – 1.00)	1.00 (0.93 – 1.00)	1.00 (1.00 – 1.00)
LASSO	1.00 (0.97 - 1.00)	1.00 (0.93 – 1.00)	1.00 (0.95 – 1.00)	1.00 (0.93 – 1.00)	1.00 (0.95 – 1.00)	1.00 (0.93 – 1.00)	1.00 (1.00 – 1.00)
Elastic-Net	1.00 (0.97 - 1.00)	1.00 (0.93 – 1.00)	1.00 (0.95 – 1.00)	1.00 (0.93 – 1.00)	1.00 (0.95 – 1.00)	1.00 (0.93 – 1.00)	1.00 (1.00 – 1.00)
Boruta	1.00 (0.97 - 1.00)	1.00 (0.93 – 1.00)	1.00 (0.95 – 1.00)	1.00 (0.93 – 1.00)	1.00 (0.95 – 1.00)	1.00 (0.93 – 1.00)	1.00 (1.00 – 1.00)
BorutaShap	1.00 (0.97 - 1.00)	1.00 (0.93 – 1.00)	1.00 (0.95 – 1.00)	1.00 (0.93 – 1.00)	1.00 (0.95 – 1.00)	1.00 (0.93 – 1.00)	1.00 (1.00 – 1.00)
Uzlaşmacı (Konsensüs)	1.00 (0.97 - 1.00)	1.00 (0.93 – 1.00)	1.00 (0.95 – 1.00)	1.00 (0.93 – 1.00)	1.00 (0.95 – 1.00)	1.00 (0.93 – 1.00)	1.00 (1.00 – 1.00)

PTD: Pozitif tahmin değeri, NTD: Negatif tahmin değeri, EAKA: Eğri altında kalan alan, GA: Güven aralığı.

Tablo 9: XGBoost modelinin test performansı

Değişken seçim yöntemi	Performans ölçütleri (%95 GA)						
	Doğruluk	**Duyarlılık**	**Seçicilik**	**PTD**	**NTD**	**F_1 skor**	**EAKA**
Hiçbiri	0.87 (0.70 - 0.96)	0.85 (0.55 - 0.98)	0.89 (0.65 - 0.99)	0.85 (0.55 - 0.98)	0.89 (0.65 - 0.99)	0.85 (0.72 - 1.00)	0.94 (0.85 – 1.00)
LASSO	0.74 (0.55 - 0.88)	0.77 (0.46 - 0.95)	0.72 (0.47 - 0.90)	0.67 (0.38 - 0.88)	0.81 (0.54 - 0.96)	0.71 (0.55 - 0.96)	0.90 (0.79 – 1.00)
Elastic-Net	0.81 (0.63 - 0.93)	0.77 (0.46 - 0.95)	0.83 (0.59 - 0.96)	0.77 (0.46 - 0.95)	0.83 (0.59 - 0.96)	0.77 (0.61 - 0.99)	0.90 (0.78 – 1.00)
Boruta	0.87 (0.70 - 0.96)	0.77 (0.46 - 0.95)	0.94 (0.73 - 1.00)	0.91 (0.59 - 1.00)	0.85 (0.62 - 0.97)	0.83 (0.70 – 1.00)	0.91 (0.80 – 1.00)
BorutaShap	0.77 (0.59 - 0.90)	0.77 (0.46 - 0.95)	0.78 (0.52 - 0.94)	0.71 (0.42 - 0.92)	0.82 (0.57 - 0.96)	0.74 (0.58 - 0.98)	0.85 (0.72 – 0.99)
Uzlaşmacı (Konsensüs)	0.71 (0.52 - 0.86)	0.54 (0.25 - 0.81)	0.83 (0.59 - 0.96)	0.70 (0.35 - 0.93)	0.71 (0.48 - 0.89)	0.61 (0.42 - 0.88)	0.78 (0.61 – 0.95)

PTD: Pozitif tahmin değeri, NTD: Negatif tahmin değeri, EAKA: Eğri altında kalan alan, GA: Güven aralığı.



Şekil 10: XGBoost modeline ilişkin ROC grafikleri

Tablo 10: DSA modelinin eğitim performansı

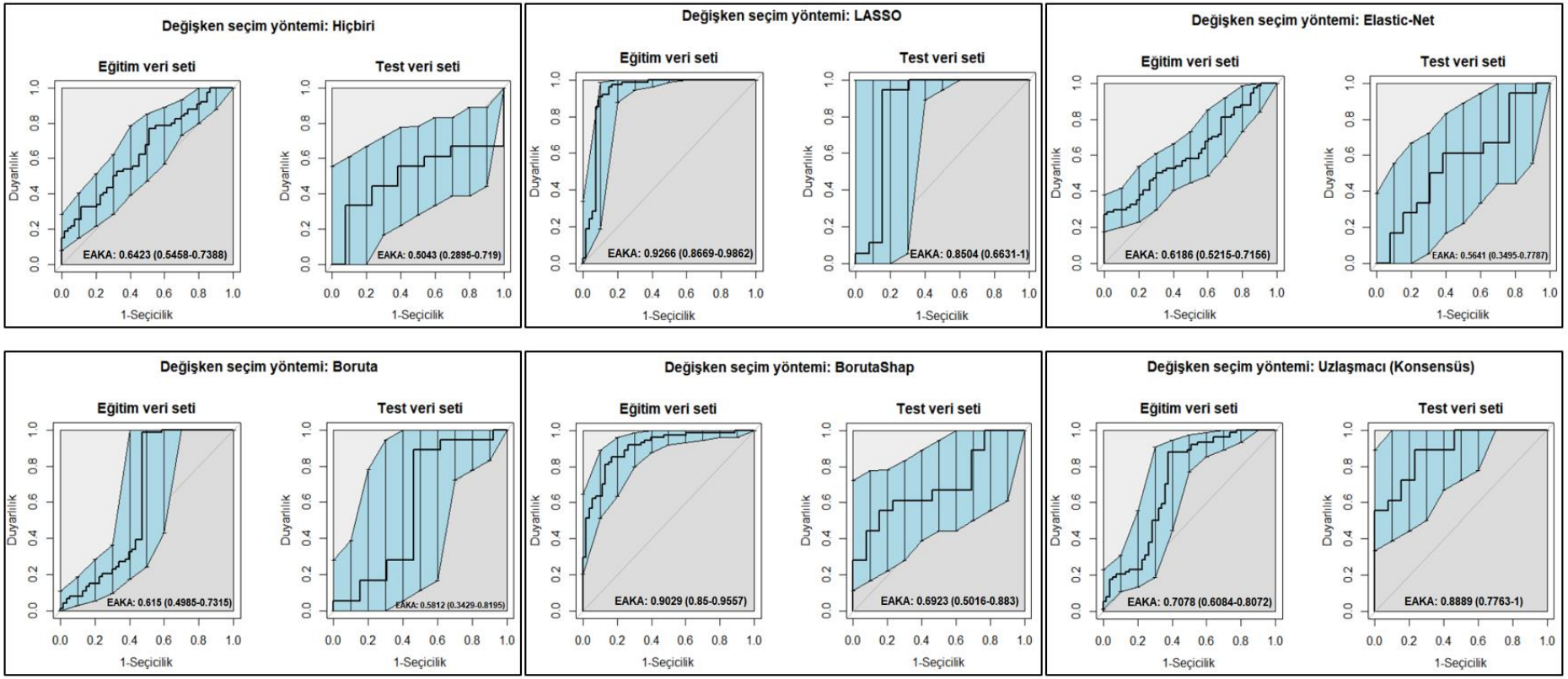
Değişken seçim yöntemi	Performans ölçütleri (%95 GA)						
	Doğruluk	**Duyarlılık**	**Seçicilik**	**PTD**	**NTD**	**F_1 skor**	**EAKA**
Hiçbiri	1 (0.97 - 1)	1.00 (0.93 - 1.00)	1.00 (0.95 - 1.00)	1.00 (0.93 - 1.00)	1.00 (0.95 - 1.00)	1.00 (0.93 - 1.00)	0.64 (0.55 – 0.74)
LASSO	0.93 (0.87 - 0.97)	0.92 (0.82 - 0.98)	0.93 (0.85 - 0.98)	0.91 (0.80 - 0.97)	0.95 (0.87 - 0.98)	0.92 (0.86 - 0.98)	0.93 (0.87 – 0.99)
Elastic-Net	1 (0.97 - 1)	1.00 (0.93 - 1.00)	1.00 (0.95 - 1.00)	1.00 (0.93 - 1.00)	1.00 (0.95 - 1.00)	1.00 (0.93 - 1.00)	0.62 (0.52 – 0.72)
Boruta	0.80 (0.71 - 0.86)	0.53 (0.39 - 0.67)	0.99 (0.93 - 1.00)	0.97 (0.82 - 1.00)	0.74 (0.65 - 0.83)	0.68 (0.58 - 0.81)	0.62 (0.50 – 0.73)
BorutaShap	0.90 (0.83 - 0.94)	0.79 (0.66 - 0.89)	0.97 (0.91 - 1.00)	0.95 (0.85 - 0.99)	0.87 (0.78 - 0.93)	0.87 (0.79 - 0.95)	0.90 (0.85 – 0.96)
Uzlaşmacı (Konsensüs)	0.77 (0.69 - 0.84)	0.62 (0.48 - 0.75)	0.88 (0.78 - 0.94)	0.79 (0.63 - 0.90)	0.76 (0.66 - 0.85)	0.70 (0.60 - 0.82)	0.71 (0.61 – 0.81)

PTD: Pozitif tahmin değeri, NTD: Negatif tahmin değeri, EAKA: Eğri altında kalan alan, GA: Güven aralığı.

Tablo 11: DSA modelinin test performansı

Değişken seçim yöntemi	Performans ölçütleri (%95 GA)						
	Doğruluk	**Duyarlılık**	**Seçicilik**	**PTD**	**NTD**	**F_1 skor**	**EAKA**
Hiçbiri	0.61 (0.42 - 0.78)	0.62 (0.32 - 0.86)	0.61 (0.36 - 0.83)	0.53 (0.27 - 0.79)	0.69 (0.41 - 0.89)	0.57 (0.37 - 0.83)	0.50 (0.29 – 0.72)
LASSO	0.87 (0.70 - 0.96)	0.85 (0.55 - 0.98)	0.89 (0.65 - 0.99)	0.85 (0.55 - 0.98)	0.89 (0.65 - 0.99)	0.85 (0.73 - 1.00)	0.85 (0.66 – 1.00)
Elastic-Net	0.77 (0.59 - 0.90)	0.85 (0.55 - 0.98)	0.72 (0.47 - 0.90)	0.69 (0.41 - 0.89)	0.87 (0.60 - 0.98)	0.76 (0.60 - 0.95)	0.56 (0.35 – 0.78)
Boruta	0.68 (0.49 - 0.83)	0.54 (0.25 - 0.81)	0.78 (0.52 - 0.94)	0.64 (0.31 - 0.89)	0.70 (0.46 - 0.88)	0.58 (0.37 - 0.83)	0.58 (0.34 – 0.82)
BorutaShap	0.68 (0.49 - 0.83)	0.69 (0.39 - 0.91)	0.67 (0.41 - 0.87)	0.60 (0.32 - 0.84)	0.75 (0.48 - 0.93)	0.64 (0.46 - 0.89)	0.69 (0.50 – 0.88)
Uzlaşmacı (Konsensüs)	0.77 (0.59 - 0.90)	0.77 (0.46 - 0.95)	0.78 (0.52 - 0.94)	0.71 (0.42 - 0.92)	0.82 (0.57 - 0.96)	0.74 (0.59 - 0.96)	0.89 (0.78 – 1.00)

PTD: Pozitif tahmin değeri, NTD: Negatif tahmin değeri, EAKA: Eğri altında kalan alan, GA: Güven aralığı.



Şekil 11: DSA modeline ilişkin ROC grafikleri

Tablo 12: LightGBM modelinin eğitim performansı

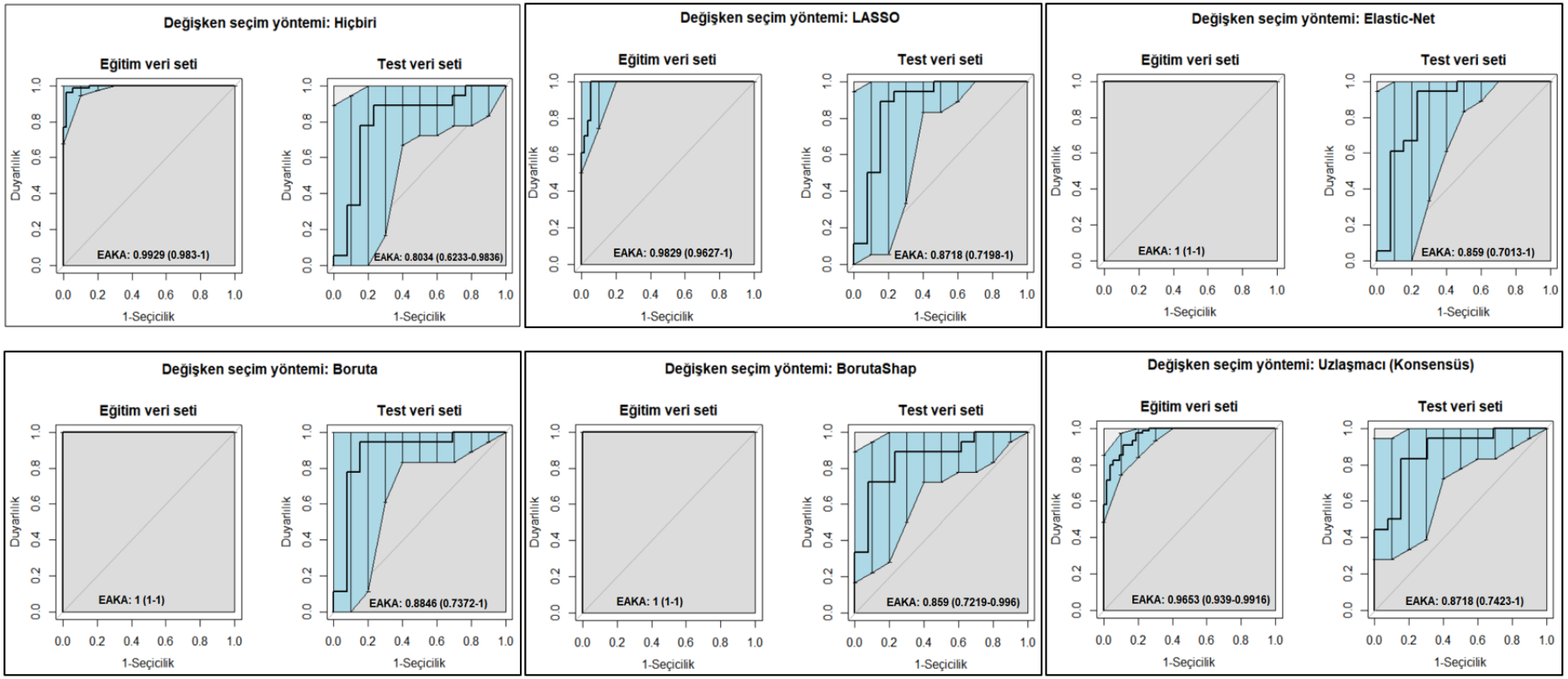
Değişken seçim yöntemi	Performans ölçütleri (%95 GA)						
	Doğruluk	**Duyarlılık**	**Seçicilik**	**PTD**	**NTD**	**F_1 skor**	**EAKA**
Hiçbiri	0.94 (0.88 - 0.97)	0.87 (0.75 - 0.95)	0.99 (0.93 - 1.00)	0.98 (0.89 - 1.00)	0.91 (0.83 - 0.96)	0.92 (0.87 - 0.98)	0.99 (0.98 – 1.00)
LASSO	0.98 (0.93 – 1.00)	0.94 (0.84 - 0.99)	1.00 (0.95 - 1.00)	1.00 (0.93 - 1.00)	0.96 (0.89 - 0.99)	0.97 (0.94 - 1.00)	0.98 (0.96 - 1.00)
Elastic-Net	1.00 (0.97 - 1.00)	1.00 (0.93 - 1.00)	1.00 (0.95 - 1.00)	1.00 (0.93 - 1.00)	1.00 (0.95 - 1.00)	1.00 (0.93 - 1.00)	1.00 (1.00 – 1.00)
Boruta	1.00 (0.97 - 1.00)	1.00 (0.93 - 1.00)	1.00 (0.95 - 1.00)	1.00 (0.93 - 1.00)	1.00 (0.95 - 1.00)	1.00 (0.93 - 1.00)	1.00 (1.00 – 1.00)
BorutaShap	1.00 (0.97 - 1.00)	1.00 (0.93 - 1.00)	1.00 (0.95 - 1.00)	1.00 (0.93 - 1.00)	1.00 (0.95 - 1.00)	1.00 (0.93 - 1.00)	1.00 (1.00 – 1.00)
Uzlaşmacı (Konsensüs)	0.88 (0.81 - 0.93)	0.81 (0.68 - 0.91)	0.93 (0.85 - 0.98)	0.90 (0.77 - 0.97)	0.87 (0.78 - 0.94)	0.85 (0.79 - 0.93)	0.97 (0.94 – 0.99)

PTD: Pozitif tahmin değeri, NTD: Negatif tahmin değeri, EAKA: Eğri altında kalan alan, GA: Güven aralığı.

Tablo 13: LightGBM modelinin test performansı

Değişken seçim yöntemi	Performans ölçütleri (%95 GA)						
	Doğruluk	**Duyarlılık**	**Seçicilik**	**PTD**	**NTD**	**F_1 skor**	**EAKA**
Hiçbiri	0.84 (0.66 - 0.95)	0.77 (0.46 - 0.95)	0.89 (0.65 - 0.99)	0.83 (0.52 - 0.98)	0.84 (0.60 - 0.97)	0.80 (0.64-1.00)	0.80 (0.62 – 0.98)
LASSO	0.87 (0.70 - 0.96)	0.85 (0.55 - 0.98)	0.89 (0.65 - 0.99)	0.85 (0.55 - 0.98)	0.89 (0.65 - 0.99)	0.85 (0.73-1.00)	0.87 (0.72 – 1.00)
Elastic-Net	0.87 (0.70 - 0.96)	0.77 (0.46 - 0.95)	0.94 (0.73 - 1.00)	0.91 (0.59 - 1.00)	0.85 (0.62 - 0.97)	0.83 (0.70-1.00)	0.86 (0.70 – 1.00)
Boruta	0.90 (0.74 - 0.98)	0.85 (0.55 - 0.98)	0.94 (0.73 - 1.00)	0.92 (0.62 - 1.00)	0.89 (0.67 - 0.99)	0.88 (0.76-1.00)	0.88 (0.73 – 1.00)
BorutaShap	0.84 (0.66 - 0.95)	0.77 (0.46 - 0.95)	0.89 (0.65 - 0.99)	0.83 (0.52 - 0.98)	0.84 (0.60 - 0.97)	0.80 (0.64-1.00)	0.86 (0.72 – 1.00)
Uzlaşmacı (Konsensüs)	0.84 (0.66 - 0.95)	0.69 (0.39 - 0.91)	0.94 (0.73 - 1.00)	0.90 (0.55 - 1.00)	0.81 (0.58 - 0.95)	0.78 (0.63-1.00)	0.87 (0.74 – 1.00)

PTD: Pozitif tahmin değeri, NTD: Negatif tahmin değeri, EAKA: Eğri altında kalan alan, GA: Güven aralığı.



Şekil 12: LightGBM modeline ilişkin ROC grafikleri

Tablo 14: YOK modelinin eğitim performansı

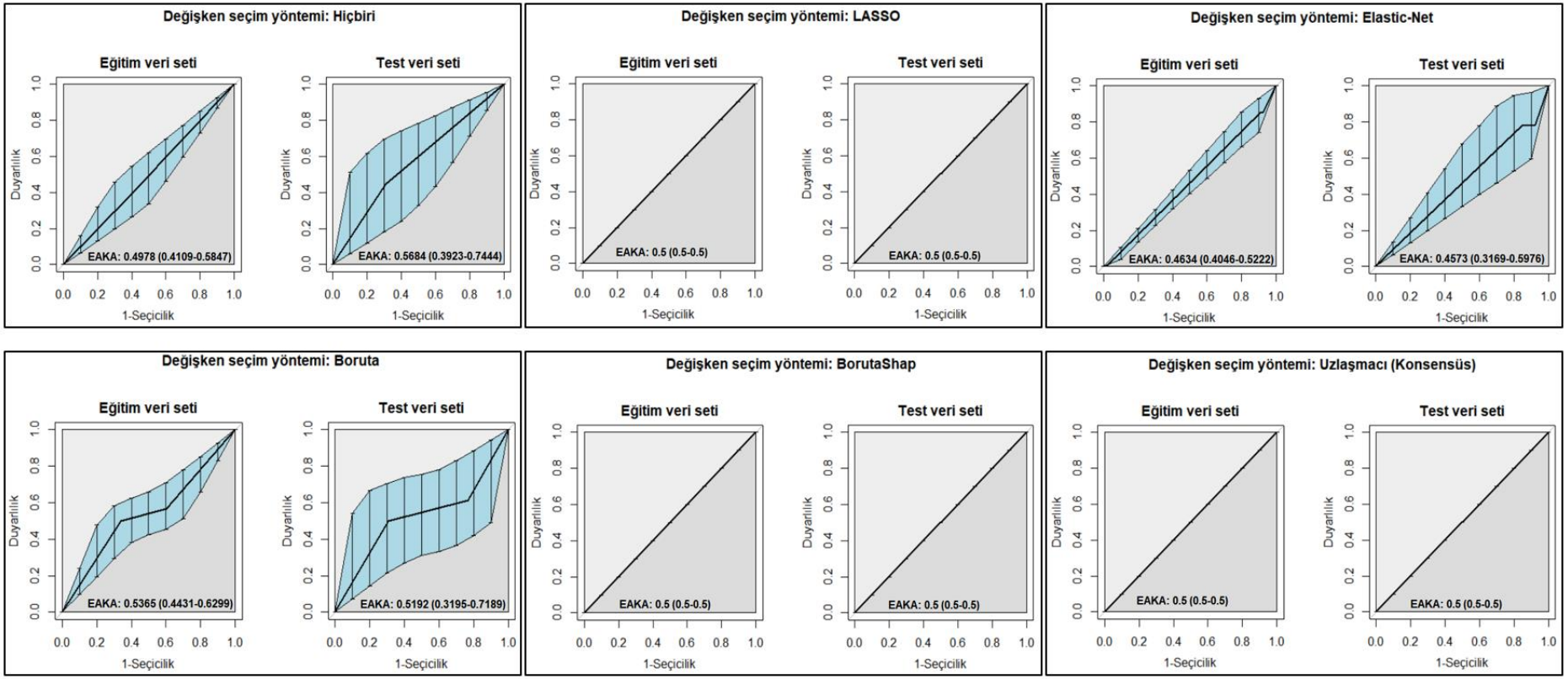
Değişken seçim yöntemi	Performans ölçütleri (%95 GA)						
	Doğruluk	Duyarlılık	Seçicilik	PTD	NTD	F_1 skor	EAKA
Hiçbiri	0.58 (0.49 - 0.67)	0.00 (0.00 - 0.07)	1.00 (0.95 - 1.00)	-	0.58 (0.49 - 0.67)	-	0.49 (0.41 – 0.58)
LASSO	0.58 (0.49 - 0.67)	0.00 (0.00 - 0.07)	1.00 (0.95 - 1.00)	-	0.58 (0.49 - 0.67)	-	0.50 (0.50 – 0.50)
Elastic-Net	0.58 (0.49 - 0.67)	0.00 (0.00 - 0.07)	1.00 (0.95 - 1.00)	-	0.58 (0.49 - 0.67)	-	0.46 (0.40 – 0.52)
Boruta	0.58 (0.49 - 0.67)	0.00 (0.00 - 0.07)	1.00 (0.95 - 1.00)	-	0.58 (0.49 - 0.67)	-	0.46 (0.32 – 0.60)
BorutaShap	0.58 (0.49 - 0.67)	0.00 (0.00 - 0.07)	1.00 (0.95 - 1.00)	-	0.58 (0.49 - 0.67)	-	0.50 (0.50 – 0.50)
Uzlaşmacı (Konsensüs)	0.58 (0.49 - 0.67)	0.00 (0.00 - 0.07)	1.00 (0.95 - 1.00)	-	0.58 (0.49 - 0.67)	-	0.50 (0.50 – 0.50)

PTD: Pozitif tahmin değeri, NTD: Negatif tahmin değeri, EAKA: Eğri altında kalan alan, GA: Güven aralığı.

Tablo 15: YOK modelinin test performansı

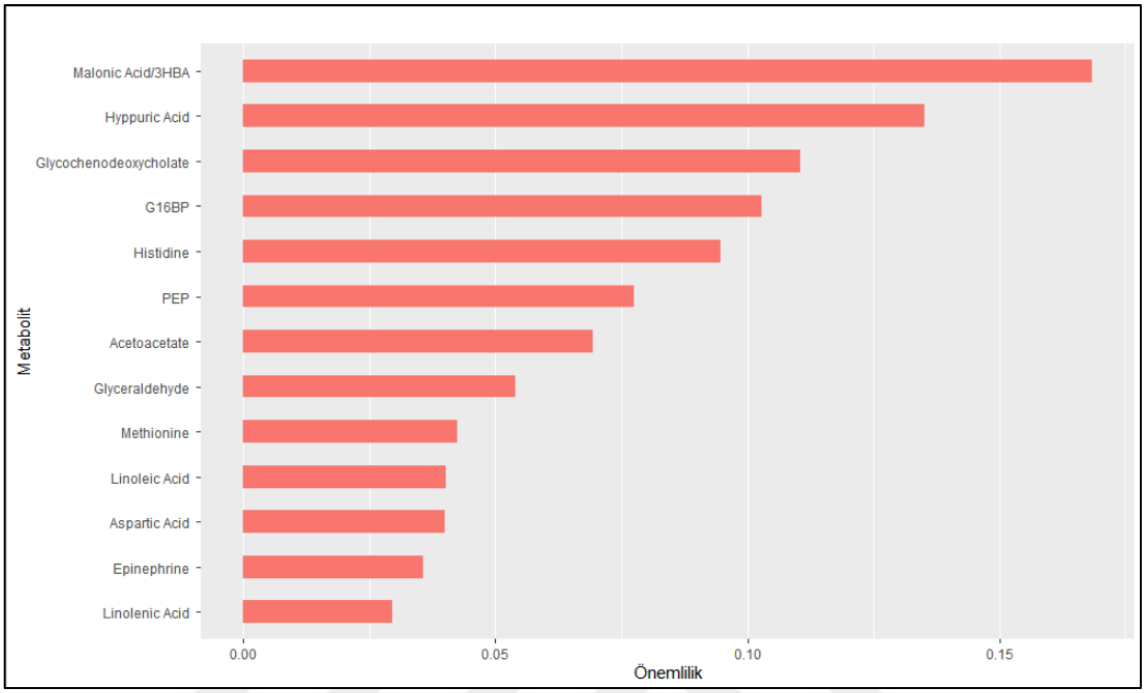
Değişken seçim yöntemi	Performans ölçütleri (%95 GA)						
	Doğruluk	**Duyarlılık**	**Seçicilik**	**PTD**	**NTD**	**F_1 skor**	**EAKA**
Hiçbiri	0.58 (0.39 - 0.75)	0.00 (0.00 - 0.25)	1.00 (0.81 - 1.00)	-	0.58 (0.39 - 0.75)	-	0.57 (0.39 - 0.74)
LASSO	0.58 (0.39 - 0.75)	0.00 (0.00 - 0.25)	1.00 (0.81 - 1.00)	-	0.58 (0.39 - 0.75)	-	0.50 (0.50 – 0.50)
Elastic-Net	0.58 (0.39 - 0.75)	0.00 (0.00 - 0.25)	1.00 (0.81 - 1.00)	-	0.58 (0.39 - 0.75)	-	0.46 (0.32 – 0.60)
Boruta	0.58 (0.39 - 0.75)	0.00 (0.00 - 0.25)	1.00 (0.81 - 1.00)	-	0.58 (0.39 - 0.75)	-	0.52 (0.32 – 0.72)
BorutaShap	0.58 (0.39 - 0.75)	0.00 (0.00 - 0.25)	1.00 (0.81 - 1.00)	-	0.58 (0.39 - 0.75)	-	0.50 (0.50 – 0.50)
Uzlaşmacı (Konsensüs)	0.58 (0.39 - 0.75)	0.00 (0.00 - 0.25)	1.00 (0.81 - 1.00)	-	0.58 (0.39 - 0.75)	-	0.50 (0.50 – 0.50)

PTD: Pozitif tahmin değeri, NTD: Negatif tahmin değeri, EAKA: Eğri altında kalan alan, GA: Güven aralığı.



Şekil 13: YOK modeline ilişkin ROC grafikleri

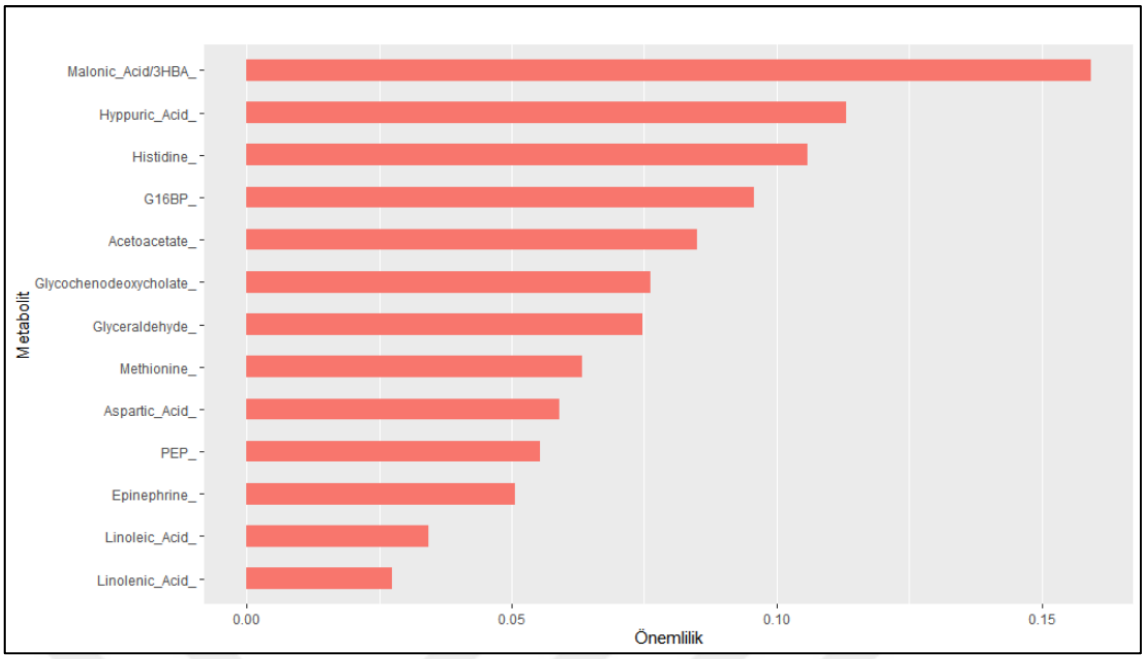
En yüksek sınıflandırma performansı veren XGBoost ve LightGBM modellerinden, Boruta değişken seçim yöntemi sonrası elde edilen değişken önemlilik tabloları Tablo 16-17'de, grafikleri ise Şekil 14-15'de verilmiştir.



Şekil 14: Boruta değişken seçim yöntemi sonrası XGBoost modelinden elde edilen değişken önemlilikleri

Tablo 16: Boruta değişken seçim yöntemi sonrası XGBoost modelinden elde edilen değişken önemlilikleri

Metabolit	**Önemlilik**
Malonic Acid/3HBA	0.168
Hippuric Acid	0.135
Glycochenodeoxycholate	0.11
G16BP	0.103
Histidine	0.095
PEP	0.077
Acetoacetate	0.069
Glyceraldehyde	0.054
Methionine	0.042
Linoleic Acid	0.04
Aspartic Acid	0.04
Epinephrine	0.036
Linolenic Acid	0.03



Şekil 15: Boruta değişken seçim yöntemi sonrası LightGBM modelinden elde edilen değişken önemlilikleri

Tablo 17: Boruta değişken seçim yöntemi sonrası LightGBM modelinden elde edilen değişken önemlilikleri

Metabolit	**Önemlilik**
Malonic Acid/3HBA	0.159
Hippuric Acid	0.113
Histidine	0.106
G16BP	0.096
Acetoacetate	0.085
Glycochenodeoxycholate	0.076
Glyceraldehyde	0.075
Methionine	0.063
Aspartic Acid	0.059
PEP	0.055
Epinephrine	0.051
Linoleic Acid	0.034
Linolenic Acid	0.027

5. TARTIŞMA

KRK, büyük bir halk sağlığı sorunudur ve tedavinin sonuçları erken teşhise bağlıdır. Metabolomikler erken tanı, evreleme, prognoz ve takip için biyobelirteçler sağlayabilir. KRK dünyada en sık görülen kanser türlerindendir. KRK'nın çoğunluğu sporadik olarak meydana gelir ve çevresel etkilerin hastalığın baskın nedeni olduğunu gösterir. İn vivo ve in vitro çalışmalar, protein alımının KRK riski üzerindeki etkisini araştırmış ve aşırı protein tüketiminin DNA hasarına yol açabileceğini ve kolonosit bütünlüğünün korunmasını etkileyebileceğini öne sürmüştür. Bağırsak mikrobiyomunda diyete bağlı değişikliklerin son zamanlarda KRK salgınlarına katkıda bulunduğu varsayılmaktadır. Bu duruma göre, çalışmalar, tümörler tercihen konak mikrobiyomlarının yaklaşık %70'i tarafından kolonize edilen distal kolon ve rektumda geliştiğinden, bağırsak mikrobiyomunun KRK'nın başlatılması ve ilerlemesi için önemli olabileceğini öne sürmüştür. Mikrobiyom, muhtemelen interlökinler, tümör nekroz faktörü-alfa ve reaktif oksijen türleri gibi aracıları kullanarak, KRK gelişimini destekleyen bir mikro çevre oluşturma potansiyeline sahiptir. Ayrıca, bağırsak mikrobiyotasının metabolik ürünleri KRK geliştirme riskini artırabilir (76).

Tıbbi onkolojide yeni sitostatiklerin geliştirilmesine bağlı tüm ilerlemelere rağmen, kemoterapi, radyoterapi ve cerrahiyi içeren yeni multidisipliner protokoller, tümör evresine bağlı olarak farklı sekans ve kombinasyonlarda bir araya getirilerek, biyolojik tedavi için yeni moleküller (monoklonal antikorlar) geliştirilmektedir. Çoğu metastatik hasta sadece palyatif tedavi alacak olsa da, tedavi amaçlı agresif cerrahiden fayda görebilecek vakalar vardır. Tıp dünyasını ilgilendiren sorunlardan biri de cerrahi tedavi ile tedavi edilebilecek vakalar ile gizli metastazlı vakaları ayırt edebilmek için biyobelirteç bulmaktır. Son on yıllık literatürün sunduğu olası çözümlerden biri "metabolomik" olarak bildirilmektedir. Metabolomik, bir hücre, organ veya organizmadaki çeşitli hücresel metabolik süreçlerle ilgili küçük moleküllerin veya metabolit profillerinin benzersiz kimyasal parmak izlerinin sistematik olarak incelenmesidir (77).

Moleküler biyoloji alanı, özellikle onkolojik olanlar olmak üzere cerrahi hastaların tanı, prognoz ve takibinde büyük etkileri olan son on yılda hızlı bir ilerleme kaydetmiştir. Hasta evreleme üzerinde büyük potansiyele sahip moleküler biyolojinin yeni doğan dallarından biri, son on yılda büyük bir gelişme gösteren metabolomiktir.

Cerrahi örneklerden veya biyolojik sıvılardan (serum, idrar, vb.) elde edilen biyolojik numunelerde, metabolitlerin varlığını ortaya çıkarmak için kütle spektrofotometrisi tespiti ile birlikte MRI veya kromatografi kullanır. Yüksek hassasiyeti ve hızlı veri toplaması nedeniyle, kütle spektrometrisi, kromatografi ile birlikte metabolomik alanında giderek artan önemli bir rol oynamaktadır. Metabolomik yöntemlerine yönelik teknoloji, deneysel modeller, yazılım ve veri tabanları konusunda büyük ilerleme kaydedilmiştir (78-80).

Bu veriler ışığında, bu tez çalışmasında; metabolomik teknolojisini kullanarak kolorektal kanseri hastalığının çeşitli topluluk ve derin öğrenme yöntemleri ile sınıflandırılması ve bir karar destek sistemi niteliğinde kullanılabilecek ardışık kod sistemi (AKD, pipeline) oluşturulması amaçlanmıştır. AKD oluşturulurken XGBoost, LightGBM, DSA ve YOK gibi ağaç tabanlı topluluk öğrenme ile derin öğrenme modelleri kullanılmıştır. Ayrıca omik çalışmalarında değişken seçim aşamasının önemli yer tutması nedeniyle modelleme öncesi LASSO, Elastic-Net, Boruta, BorutaShap ve Uzlaşmacı (Konsensüs) değişken seçimi yaklaşımları kullanılarak ilgili modellerin sınıflandırma performansları üzerindeki etkileri araştırılmıştır. Modellerin sınıflandırma performansları çeşitli performans ölçütleri ile değerlendirilmiştir. Geliştirilen AKD, omik verileri üzerinde kapsamlı veri analizi yapacak olan araştırmacılar için çalışma ekinde (Ek-3) paylaşılmıştır.

XGBoost modelinin test veri seti üzerindeki performansı tüm performans metrikleri açısından incelendiğinde doğruluk, duyarlılık, seçicilik, PTD, NTD, F_1-skoru ve EAKA değerleri sırasıyla, 0.87 (0.70 - 0.96), 0.77 (0.46 - 0.95), 0.94 (0.73 - 1.00), 0.91 (0.59 - 1.00), 0.85 (0.62 - 0.97), 0.83 (0.70 – 1.00) ve 0.91 (0.80 - 1.00) olarak hesaplanmıştır. Bu bulgular, Boruta değişken seçim yöntemi uygulandıktan sonra elde edilen değerlerdir. Hesaplanan bu değerler hiçbir değişken seçim yöntemi uygulanmadığı durumda elde edilen sınıflandırma performans değerlerine yakın olmakla birlikte, daha az sayıda değişken ile elde edilen başarım daha öncelikli olarak ele alınmıştır. Dolayısıyla Boruta+XGBoost kombinasyonunun diğer kombinasyonlara göre daha başarılı olduğu söylenebilir. Elde edilen bu bulgu doğrultusunda yapılan literatür taramalarında, güncel bir çalışmada (81), hedefsiz (untargeted) metabolomik veriler yardımıyla KRK'nın evrelerini sınıflandırmada XGBoost modeli kullanılmış ve ilgili modelin PLS-DA (Partial Least Squares-Discriminant Analysis), SVM (Support Vector Machines) ve Random Forest modellerine göre sınıflandırma performansı olarak bariz üstünlük sağladığı rapor

edilmiştir. Yine bir radyomik verileri tabanlı in situ karsinoma hastalığına yönelik yapılan bir biyobelirteç tahmin çalışmasında (82), minimum redundancy maximum relevance ensemble (mRMRe) değişken seçim yöntemine göre, Boruta tekniği ile seçilen değişkenlerin daha iyi sınıflandırma sonuçları verdiği rapor edilmiştir. Yine çeşitli omik teknolojileri ile yapılan çalışmalarda da Boruta yönteminin omik çalışmalarında daha sıklıkla kullanılmaya başlandığı görülmektedir (83-85). Veri boyutlarının artması sonucu, en uygun değişken alt kümesinin bulma süresindeki uzama, Boruta algoritmasının önündeki en büyük engeldir. Nitekim mevcut tez çalışmasında Boruta algoritmasının çalışma süresi LASSO ve Elastic-Net yaklaşımlarına göre uzundu.

DSA modeline ilişkin bulgular incelendiğinde, hem eğitim hem de test veri seti üzerinde en iyi sınıflandırma performansı LASSO değişken seçim yöntemi uygulandıktan sonra elde edilmiştir. Söz konusu model için test verisi üzerinden elde edilen performans metrikleri, doğruluk, duyarlılık, seçicilik, PTD, NTD, F_1-skoru ve EAKA sırasıyla 0.87 (0.70 - 0.96), 0.85 (0.55 - 0.98), 0.89 (0.65 - 0.99), 0.85 (0.55 - 0.98), 0.89 (0.65 - 0.99), 0.85 (0.73 - 1.00) ve 0.85 (0.66 - 1.00) olarak hesaplanmıştır. Derin öğrenme ve omik teknolojilerinin bir araya getirilmesi görece yeni olmasına rağmen omik çalışmalarındaki biyobelirteç keşif görevlerini görevleri yerine getirmek için derin öğrenme modellerinin uygun oldukları belirtilmiştir (86). Fakat, kayıp fonksiyonu için global bir optimum bulmanın zorluğu, girdi değişkenlerinin sayısı arttıkça bir sinir ağında öğrenilecek ağırlıkların sayısının katlanarak artması ve bunun da aşırı öğrenme (overfitting) olasılığını artırması ve uygun bir derin sinir ağının tasarımının uzmanlık gerektirmesi gibi nedenlerden dolayı derin öğrenme teknolojilerinin omik teknolojilerindeki kullanışlılığını sorgulatmaktadır (87). LASSO tekniği, özellikle değişken sayısının örnek sayısından oldukça fazla olduğu ($n<<p$) omik verilerinde uygun değişken alt kümesinin seçiminde, hızlı sonuç verme özelliği nedeniyle sıklıkla kullanılmaktadır. Fakat değişkenler arası yüksek korelasyon yapısı durumunda ortaya çıkan çoklu bağlantı (multicollinearity) varlığında, LASSO yönteminin uygulanması önerilmemektedir (88).

YOK modeli, tüm modeller arasında en düşük sınıflandırma performansı sergileyen model olmuştur. Tüm değişken seçim yöntemleri uygulandığında dahi, doğruluk, duyarlılık, seçicilik, pozitif tahmin değeri, negatif tahmin değeri F1-skoru ve EAKA değerleri sırasıyla en yüksek 0.58 (0.39 - 0.75), 0.00 (0.00 - 0.25), 1.00 (0.81 - 1.00), 0.58 (0.39 - 0.75) ve 0.52 (0.32 – 0.72) olarak hesaplanmıştır.

LightGBM modelinin eğitim ve test veri seti üzerindeki sınıflandırma performansı birlikte ele alındığında modeller içindeki en başarılı sınıflandırma performansını sergilediği görülmüştür. İlgili modelin test veri seti üzerindeki performansı doğruluk, duyarlılık, seçicilik, PTD, NTD, F1-skoru ve EAKA ölçütleri açısından sırasıyla 0.90 (0.74 - 0.98), 0.85 (0.55 - 0.98), 0.94 (0.73 - 1.00), 0.92 (0.62 - 1.00), 0.89 (0.67 - 0.99), 0.88 (0.76 - 1.00) ve 0.88 (0.73 - 1.00) olarak hesaplanmıştır. Bu bulgular, Boruta değişken seçim yöntemi uygulandıktan sonra elde edilmiştir. Diğer değişken seçim stratejileri açısından ele alındığında da yine LightGBM modelinin diğer modellere göre başarılı sınıflandırma sonuçları verdiği görülmektedir. İlgili modelin başarılı sınıflandırma bulgularının yanı sıra, hesaplama hızı anlamında da diğer modellerden daha başarılı olduğu görülmüştür. Güncel sayılabilecek bir model olan LightGBM modelinin, omik çalışmalarında yeni yeni kullanılmaya başlandığı görülmektedir. Meme kanserinin poligenik risk skorunu tahminine yönelik yapılan bir güncel çoklu-omik çalışmasında (89), LightGBM modelinin, ilgili çalışmada kullanılan diğer geleneksel doğrusal modeller ile makine öğrenmesi yöntemlerine göre daha başarılı sonuçlar verdiği rapor edilmiştir.

XGBoost ve LightGBM modellerinin değişken önemlilik bulguları birlikte ele alındığında, ilk 5 metabolik değişken içinde 4 tanesinin (Malonic Acid/3HBA, Hippuric Acid, G16BP ve Histidine) ortak olarak her iki modelde de yer aldığı görülmektedir. Yapılan araştırmalarda, KRK hastalığının varlığı durumunda, Hippuric acid ve G16BP seviyesinin yükseldiği buna karşın Malonic Acid/3HBA ve Histidine seviyelerinin düştüğü rapor edilmiştir (44, 90, 91). Değişken önemlilik tablolarında XGBoost ve LightGBM modelleri tarafından, ortak olarak ilk 5 metabolit içinde gösterilen bu 4 metabolitin, KRK hastalığının erken teşhisinde biyobelirteç olarak araştırıcılara önerilebilir.

Bu tez çalışmasında metabolomik teknolojisi ve çeşitli derin öğrenme ve topluluk öğrenme yöntemleri kullanılarak KRK hastalığının sınıflandırılması ve bir ardışık kod sistemi oluşturularak karar destek sistemi oluşturulması amaçlanmıştır. Bulgular incelendiğinde, topluluk öğrenme yöntemlerinden olan XGBoost ve LightGBM modellerinin diğer derin öğrenme modellerine kıyasla daha başarılı sınıflandırma sonuçlar verdiği görülmüştür. Özellikle LightGBM modeli hem eğitim hem de test veri setlerinin sınıflandırmasında başarılı sonuçlar verdiği söylenebilir. Ayrıca LightGBM modeli bu sınıflandırma başarımını tüm değişken seçim yöntemleri bazında elde etmiştir

ve ilgili model, veriyi işleme ve hesaplama hızı anlamında en üstün model olmuştur. XGBoost modeli de LightGBM modelinden sonra ikinci en iyi sınıflandırma performansını veren model olmuştur. Bu kapsamda, ağaç tabanlı gradiyent artırma modellerinin, popüler derin öğrenme modelleri ile başa çıkacak yüksek sınıflandırma başarımı ve düşük hesaplama maliyetlerine sahip modeller olduğu sonucu çıkarılabilir.

6. SONUÇ VE ÖNERİLER

Bu tez çalışmasında, çeşitli derin öğrenme ve topluluk öğrenme yöntemleri kullanılarak kolorektal kanseri metabolomik veri seti üzerinden sınıflandırma ve tahmin işlemleri yapılmıştır. Modellerin sınıflandırma performansları doğruluk, duyarlılık, seçicilik, PTD, NTD, F-skoru ve ROC EAKA ölçütleri ile değerlendirilmiştir.

Bulgular incelendiğinde, en zayıf sınıflandırma performansı gösteren modelin YOK olduğu görülmektedir. Modelin hiçbir değişken seçimi senaryosunda istenilen sınıflandırma performansını başarısını gösteremediği görülmektedir. Diğer derin öğrenme modeli olan DSA'nın ise YOK modeline göre daha iyi sınıflandırma performansı göstermiştir. DSA modelinin en iyi sınıflandırma performansını LASSO değişken seçimi yöntemi sonrası verdiği görülmektedir. Genel olarak derin öğrenme yaklaşımlarının, uygun ağ mimarisinin oluşturulmasında ve hiperparametre optimizasyonununda uzmanlık gerektirmesi, hatalı/eksik oluşturulabilecek ağ mimarisinin ve/veya yanlış hiperparametre aralığının seçimi, yanlış sonuçlar elde etme ve/veya aşırı uyum (overfitting) gibi olumsuzluklara neden olabileceğinden dolayı omik teknolojileri ile birlikte kullanımında dikkatli olunması gerekmektedir.

Topluluk öğrenme yöntemleri ele alındığında, XGBoost ve LightGBM modellerinin diğer derin öğrenme modellerine kıyasla daha başarılı sınıflandırma sonuçları vermiştir. Özellikle LightGBM modeli hem eğitim hem de test veri setlerinin sınıflandırmasında iyi sonuçlar verdiği söylenebilir. Ayrıca LightGBM modeli bu sınıflandırma başarımını tüm değişken seçim yöntemleri bazında elde etmiştir. LightGBM modeli, veriyi işleme ve hesaplama hızı anlamında en üstün model olmuştur. Bu özelliğinin LightGBM modelinin daha büyük verilerde çalışıldığında, XGBoost modeline göre daha tercih edilebilir hale getirmektedir.

Bu tez çalışmasında, topluluk öğrenme yöntemleri tüm değişken seçim senaryolarında derin öğrenme yöntemlerine kıyasla çok daha iyi sınıflandırma sonuçları verdiği görülmüştür. Söz konusu topluluk öğrenme yöntemlerinin, daha büyük veri setlerinde daha hızlı sonuç verebilmeleri için grafik işlem birimi (GPU) destekli sürümlerinin kullanılmasının işlem ve zaman maliyetleri açısından daha verimli olacağı önerilebilir. Ayrıca bu tez çalışmasında kullanılan topluluk öğrenme yöntemleri ile DSA modelinin, açıklanabilir/yorumlanabilir yapay zekâ çalışmaları kapsamında geliştirilen,

SHAP (SHapley Additive exPlanation), LIME (Local Interpretable Model-Agnostic Explanations), vb. tekniklerle entegre edilerek klinik yorumlamaya katkı sağlayabilecek daha yüksek çıktılı analiz raporları elde edilebilir.

Sonuç olarak bu tez çalışmasında kullanılan ve başarılı sonuçlar veren XGBoost ve LightGBM modellerinin omik ile diğer biyolojik ve tıbbi veri setlerinde kullanılabileceği ve derin öğrenme modellerine kıyasla daha az uzmanlık gerektiren yaklaşımlar olması nedeniyle araştırıcılara önerilmektedir.

KAYNAKLAR

1. El-Meniawy MM, Mohamed SI, Morsy MM, Mohamed HA-A. Colorectal Cancer Management: An Overview. *Ann Romanian Soc Cell Biol* 2021, 25(4): 6217-28.

2. Chiang T-H, Lee Y-C. Options of Colorectal Cancer Screening: An Overview. In: Chiu H-M, Chen H-H, (eds). *Colorectal Cancer Screening: Theory and Practical Application*. Singapore: Springer Singapore, 2021: 29-40.

3. Amir Hashim NA, Ab-Rahim S, Wan Ngah WZ, Nathan S, Ab Mutalib NS, Sagap I, A Jamal AR, Mazlan M. Global metabolomics profiling of colorectal cancer in Malaysian patients. *BioImpacts : BI* 2021, 11(1): 33-43.

4. Hofmann M, Klinkenberg R. *RapidMiner: Data mining use cases and business analytics applications*. Florida, CRC Press. 2016.

5. d'Alché-Buc F, Ralaivola L, editors. Incremental learning algorithms for classification and regression: Local strategies. AIP Conference Proceedings; 2002: American Institute of Physics.

6. Fayyad U, Piatetsky-Shapiro G, Smyth P. From data mining to knowledge discovery in databases. *AI magazine* 1996, 17(3): 37-54.

7. Maimon O, Rokach L. Introduction to Knowledge Discovery and Data Mining. In: Maimon O, Rokach L, (eds). *Data Mining and Knowledge Discovery Handbook*. Boston, MA: Springer US, 2010: 1-15.

8. Wei R, Wang J, Su M, Jia E, Chen S, Chen T, Ni Y. Missing Value Imputation Approach for Mass Spectrometry-based Metabolomics Data. *Sci Rep* 2018, 8(1): 1-10.

9. Smiti A. A critical overview of outlier detection methods. *Comput Sci Rev* 2020, 38(1): 1-11.

10. Li J, Cheng K, Wang S, Morstatter F, Trevino RP, Tang J, Liu H. Feature selection: A data perspective. *ACM Comput Surv* 2017, 50(6): 1-45.

11. Chen CW, Tsai YH, Chang FR, Lin WC. Ensemble feature selection in medical datasets: Combining filter, wrapper, and embedded feature selection results. *Expert Systems* 2020, 37(5): 1-10.

12. Liang J, Hou L, Luan Z, Huang W. Feature Selection with Conditional Mutual Information Considering Feature Interaction. *Symmetry* 2019, 11(7): 1-17.

13. Hameed SS, Petinrin OO, Osman A, Hashi FS. Filter-wrapper combination and embedded feature selection for gene expression data. *Int J Advance Soft Compu Appl* 2018, 10(1): 90-105.

14. Hira ZM, Gillies DF. A Review of Feature Selection and Feature Extraction Methods Applied on Microarray Data. *Adv Bioinform* 2015, 2015(1): 1-13.

15. Zou H, Xue L. A selective overview of sparse principal component analysis. *Proc IEEE* 2018, 106(8): 1311-20.

16. Supratak A, Li L, Guo Y. Feature extraction with stacked autoencoders for epileptic seizure detection. *Annu Int Conf IEEE Eng Med Biol Soc* 2014, 4(1): 4184-7.

17. Smith MT, Vermeulen R, Li G, Zhang L, Lan Q, Hubbard AE, Forrest MS, McHale C, Zhao X, Gunn L. Use of 'Omic'technologies to study humans exposed to benzene. *Chem-Biol Interact* 2005, 153(1): 123-7.

18. Vlaanderen J, Moore LE, Smith MT, Lan Q, Zhang L, Skibola CF, Rothman N, Vermeulen R. Application of OMICS technologies in occupational and environmental health research; current status and projections. *Occup Environ Med* 2010, 67(2): 136-43.

19. Fitzmaurice GM, Laird NM, Ware JH. *Applied longitudinal analysis*. Boston, MA, John Wiley & Sons. 2012.

20. Ordovas JM, Corella D. Nutritional genomics. *Annu Rev Genomics Hum Genet* 2004, 5(1): 71-118.

21. Budak ŞÖ, Dönmez S. Gıda Biliminde Yeni Omik Teknolojileri. *GIDA* 2012, 37(3): 173-9.

22. Ağırbaşlı D, Ülman YI. Genomik risk skorlaması perspektifinden koroner arter hastalığı, etik yaklaşım ve öneriler. *Anadolu Kardiyol Derg* 2012, 12(1): 171-7.

23. Organization WH. *Genomics and world health: Report of the Advisory Committee on Health Research*, World Health Organization. 2002.

24. Kadakkuzha BM, Puthanveettil SV. Genomics and proteomics in solving brain complexity. *Mol Biosyst* 2013, 9(7): 1807-21.

25. Del Boccio P, Urbani A. Homo sapiens proteomics: clinical perspectives. *Ann Ist Super Sanità* 2005, 41(4): 479-82.

26. Piétu G, Mariage-Samson R, Fayein N-A, Matingou C, Eveno E, Houlgatte R, Decraene C, Vandenbrouck Y, Tahi F, Devignes M-D. The Genexpress IMAGE

knowledge base of the human brain transcriptome: a prototype integrated resource for functional and computational genomics. *Genome Res* 1999, 9(2): 195-209.

27. McGettigan PA. Transcriptomics in the RNA-seq era. *Curr Opin Chem Biol* 2013, 17(1): 4-11.

28. Dong Z, Chen Y. Transcriptomics: Advances and approaches. *Sci China Life Sci* 2013, 56(10): 960-7.

29. Milward EA, Shahandeh A, Heidari M, Johnstone DM, Daneshi N, Hondermarck H. Transcriptomics. In: Bradshaw RA, Stahl PD, (eds). *Encyclopedia of Cell Biology*. Waltham: Academic Press, 2016: 160-5.

30. Twyman RM. Proteomics. In: Chadwick R, editor. *Encyclopedia of Applied Ethics*. San Diego: Academic Press, 2012: 642-9.

31. Noor Z, Ahn SB, Baker MS, Ranganathan S, Mohamedali A. Mass spectrometry–based protein identification in proteomics—a review. *Brief Bioinformatics* 2021, 22(2): 1620-38.

32. Martorell-Marugán J, Tabik S, Benhammou Y, del Val C, Zwir I, Herrera F, Carmona-Sáez P. Deep learning in omics data analysis and precision medicine. In: Husi H, editor. *Computational Biology*. Brisbane, AU: Codon Publications, 2019: 37-53.

33. de Los Campos G, Hickey JM, Pong-Wong R, Daetwyler HD, Calus MP. Whole-genome regression and prediction methods applied to plant and animal breeding. *Genetics* 2013, 193(2): 327-45.

34. Vignoli A, Tenori L, Giusti B, Takis PG, Valente S, Carrabba N, Balzi D, Barchielli A, Marchionni N, Gensini GF. NMR-based metabolomics identifies patients at high risk of death within two years after acute myocardial infarction in the AMI-Florence II cohort. *BMC medicine* 2019, 17(1): 1-13.

35. Nalbantoglu S. Metabolomics: basic principles and strategies. *Mol Med* 2019, 10(3): 1-16.

36. Sanchez S, Demain AL. Metabolic regulation and overproduction of primary metabolites. *Microb Biotechnol* 2008, 1(4): 283-319.

37. Patti GJ, Yanes O, Siuzdak G. Innovation: Metabolomics: the apogee of the omics trilogy. *Nat Rev Mol Cell Biol* 2012, 13(4): 263-9.

38. Liu X, Locasale JW. Metabolomics: A Primer. *Trends Biochem Sci* 2017, 42(4): 274-84.

39. Emwas AH, Roy R, McKay RT, Tenori L, Saccenti E, Gowda GAN, Raftery D, Alahmari F, Jaremko L, Jaremko M, Wishart DS. NMR Spectroscopy for Metabolomics Research. *Metabolites* 2019, 9(7): 1-39.

40. Cambiaghi A, Ferrario M, Masseroli M. Analysis of metabolomic data: tools, current strategies and future challenges for omics data integration. *Brief Bioinformatics* 2016, 18(3): 498-510.

41. Loke MF, Chua EG, Gan HM, Thulasi K, Wanyiri JW, Thevambiga I, Goh KL, Wong WF, Vadivelu J. Metabolomics and 16S rRNA sequencing of human colorectal cancers and adjacent mucosa. *PLoS One* 2018, 13(12): 1-20.

42. Martín-Blázquez A, Díaz C, González-Flores E, Franco-Rivas D, Jiménez-Luna C, Melguizo C, Prados J, Genilloud O, Vicente F, Caba O. Untargeted LC-HRMS-based metabolomics to identify novel biomarkers of metastatic colorectal cancer. *Sci Rep* 2019, 9(1): 1-9.

43. Dalal N, Jalandra R, Sharma M, Prakash H, Makharia GK, Solanki PR, Singh R, Kumar A. Omics technologies for improved diagnosis and treatment of colorectal cancer: Technical advancement and major perspectives. *Biomed Pharmacother* 2020, 131(1): 1-15.

44. Zhu J, Djukovic D, Deng L, Gu H, Himmati F, Chiorean EG, Raftery D. Colorectal cancer detection using targeted serum metabolic profiling. *J Proteome Res* 2014, 13(9): 4120-30.

45. Metabolomics Workbench. https://www.metabolomicsworkbench.org/. Son Erişim Tarihi: 09.04.2021.

46. Rubin DB. *Multiple imputation for nonresponse in surveys*. Boston, MA, John Wiley & Sons. 2004.

47. Rao AR, Reimherr M. Modern multiple imputation with functional data. *Stat* 2021, 10(1): 1-14.

48. Stekhoven DJ, Bühlmann P. MissForest—non-parametric missing value imputation for mixed-type data. *Bioinformatics* 2012, 28(1): 112-8.

49. Liaw A, Wiener M. Classification and regression by randomForest. *R news* 2002, 2(3): 18-22.

50. Tibshirani R. Regression shrinkage and selection via the lasso. *J R Stat Soc Series B Stat Methodol* 1996, 58(1): 267-88.

51. Bulut H. *R Uygulamaları ile Çok Değişkenli İstatistiksel Yöntemler*. Ankara, Nobel Akademik Yayıncılık. 2019.

52. Friedman J, Hastie T, Tibshirani R. Regularization Paths for Generalized Linear Models via Coordinate Descent. *J Stat Softw* 2010, 33(1): 1-22.

53. Zou H, Hastie T. Regularization and variable selection via the elastic net. *J R Stat Soc Series B Stat Methodol* 2005, 67(2): 301-20.

54. Kursa MB, Rudnicki WR. Feature selection with the Boruta package. *J Stat Softw* 2010, 36(11): 1-13.

55. Breiman L. Random forests. *Mach Learn* 2001, 45(1): 5-32.

56. Chen L, Li Z, Zeng T, Zhang Y-H, Li H, Huang T, Cai Y-D. Predicting gene phenotype by multi-label multi-class model based on essential functional features. *Mol Genet Genom* 2021, 1-14.

57. Keany E. BorutaShap : A wrapper feature selection method which combines the Boruta feature selection algorithm with Shapley values. https://zenodo.org/record/4247618#.YOn_1ugzaUl. Son Erişim Tarihi: 20.06.2021.

58. Fryer D, Strümke I, Nguyen H. Shapley values for feature selection: the good, the bad, and the axioms. *arXiv* 2021, 30(1): 1-8.

59. Lundberg S, Lee S-I. A unified approach to interpreting model predictions. *arXiv* 2017, 10(1): 1-10.

60. Molnar C. Interpretable Machine Learning. A Guide for Making Black Box Models Explainable. https://christophm.github.io/interpretable-ml-book/. Son Erişim Tarihi: 20.06.2021.

61. Ushey K, Allaire J, Tang Y, Eddelbuettel D, Lewis B, Keydana S, Hafen R, Geelnard M. R Interface to Python. https://rstudio.github.io/reticulate/. Son Erişim Tarihi: 02.03.2021.

62. Alotaibi B, Alotaibi M. Consensus and majority vote feature selection methods and a detection technique for web phishing. *J Ambient Intell Humaniz Comput* 2021, 12(1): 717-27.

63. Candel A, LeDell E. Deep learning with H2O. https://docs.h2o.ai/h2o/latest-stable/h2o-docs/booklets/DeepLearningBooklet.pdf. Son Erişim Tarihi: 06.06.2021.

64. Bai F, Hong D, Lu Y, Liu H, Xu C, Yao X. Prediction of the Antioxidant Response Elements' Response of Compound by Deep Learning. *Front Chem* 2019, 7(385): 1-10.

65. Chen T, Guestrin C. Xgboost: A scalable tree boosting system. *arXiv* 2016, 5(2): 1-13.

66. Qiu Y, Zhou J, Khandelwal M, Yang H, Yang P, Li C. Performance evaluation of hybrid WOA-XGBoost, GWO-XGBoost and BO-XGBoost models to predict blast-induced ground vibration. *Eng Comput* 2021, 10(1): 1-18.

67. Chen T, He T. xgboost: eXtreme Gradient Boosting. https://cran.r-project.org/web/packages/xgboost/vignettes/xgboost.pdf. Son Erişim Tarihi: 06.05.2021.

68. Ke G, Meng Q, Finley T, Wang T, Chen W, Ma W, Ye Q, Liu T-Y. Lightgbm: A highly efficient gradient boosting decision tree. *Adv Neural Inf Process Syst* 2017, 30(1): 3146-54.

69. Zhang C, Lei X, Liu L. Predicting Metabolite–Disease Associations Based on LightGBM Model. *Front Genet* 2021, 12(1): 1-11.

70. Ke G, Soukhavong D, Lamb J, Meng Q, Finley T, Wang T, Chen W, Ma W, Ye Q, Liu T-Y. lightgbm: Light Gradient Boosting Machine. https://cran.r-project.org/web/packages/lightgbm/index.html. Son Erişim Tarihi: 01.06.2021.

71. Thieme A, Mitulla B, Schulze F, Spiegier AW. Epidemiological data on Werdnig-Hoffmann disease in Germany (West-Thüringen). *J Human genetics* 1993, 91(3): 295-7.

72. Sugarman EA, Nagan N, Zhu H, Akmaev VR, Zhou Z, Rohlfs EM, Flynn K, Hendrickson BC, Scholl T, Sirko-Osadsa DA. Pan-ethnic carrier screening and prenatal diagnosis for spinal muscular atrophy: clinical laboratory analysis of> 72 400 specimens. *Eur J Hum Genet* 2012, 20(1): 27-32.

73. Liu G, Bao H, Han B. A Stacked Autoencoder-Based Deep Neural Network for Achieving Gearbox Fault Diagnosis. *Math Probl Eng* 2018, 5(1): 1-10.

74. Rong X. deepnet: Deep Learning Toolkit in R. https://cran.r-project.org/web/packages/deepnet/index.html. Son Erişim Tarihi: 04.05.2021.

75. Canty A, Ripley B. Package 'boot'. https://cran.r-project.org/web/packages/boot/boot.pdf. Son Erişim Tarihi: 01.07.2021.

76. Yang J, Seo H, Lee WH, Lee DH, Kym S, Park YS, Kim JG, Jang I-J, Kim Y-K, Cho J-Y. Colorectal cancer diagnostic model utilizing metagenomic and metabolomic data of stool microbial extracellular vesicles. *Sci Rep* 2020, 10(1): 1-10.

77. Răchieriu C, Graur F, Moiș E, Socaciu C, Eniu D, Alhajjar N. Metabolomic profile of colorectal cancer patients and its clinical implications. *Rom Biotechnol Lett* 2020, 25(1): 2045-54.

78. Misra BB. New software tools, databases, and resources in metabolomics: updates from 2020. *Metabolomics* 2021, 17(5): 1-24.

79. Belhaj MR, Lawler NG, Hoffman NJ. Metabolomics and Lipidomics: Expanding the Molecular Landscape of Exercise Biology. *Metabolites* 2021, 11(3): 1-34.

80. Zhang A, Sun H, Yan G, Wang P, Wang X. Metabolomics for Biomarker Discovery: Moving to the Clinic. *Biomed Res Int* 2015, 20151-7.

81. Tian M, Lin Z, Wang X, Yang J, Zhao W, Lu H, Zhang Z, Chen Y. Pure Ion Chromatograms Combined with Advanced Machine Learning Methods Improve Accuracy of Discriminant Models in LC–MS-Based Untargeted Metabolomics. *Molecules* 2021, 26(9): 1-16.

82. Song M, Lin J, Song F, Wu D, Qian Z. The value of MR-based radiomics in identifying residual disease in patients with carcinoma in situ after cervical conization. *Sci Rep* 2020, 10(1): 1-8.

83. Degenhardt F, Seifert S, Szymczak S. Evaluation of variable selection methods for random forests and omics data sets. *Brief Bioinformatics* 2019, 20(2): 492-503.

84. Wu P, Chen D, Ding W, Wu P, Hou H, Bai Y, Zhou Y, Li K, Xiang S, Liu P, Ju J, Guo E, Liu J, Yang B, Fan J, He L, Sun Z, Feng L, Wang J, Wu T, Wang H, Cheng J, Xing H, Meng Y, Li Y, Zhang Y, Luo H, Xie G, Lan X, Tao Y, Yuan H, Huang K, Sun W, Qian X, Li Z, Huang M, Ding P, Wang H, Qiu J, Wang F, Wang S, Zhu J, Ding X, Chai C, Liang L, Wang X, Luo L, Sun Y, Yang Y, Zhuang Z, Li T, Tian L, Zhang S, Zhu L, Chen L, Wu Y, Ma X, Chen F, Ren Y, Xu X, Liu S, Wang J, Yang H, Wang L, Sun C, Ma D, Jin X, Chen G. The Trans-omics Landscape of COVID-19. *medRxiv* 2020, 10(2): 1-43.

85. Long NP, Park S, Anh NH, Nghi TD, Yoon SJ, Park JH, Lim J, Kwon SW. High-Throughput Omics and Statistical Learning Integration for the Discovery and Validation of Novel Diagnostic Signatures in Colorectal Cancer. *Int J Mol Sci* 2019, 20(2): 1-15.

86. Zhang Z, Zhao Y, Liao X, Shi W, Li K, Zou Q, Peng S. Deep learning in omics: a survey and guideline. *Brief Funct Genomics* 2018, 18(1): 41-57.

87. Abdollahi-Arpanahi R, Gianola D, Peñagaricano F. Deep learning versus parametric and ensemble methods for genomic prediction of complex phenotypes. *Genet Sel Evol* 2020, 52(1): 1-15.

88. Bravo-Merodio L, Williams JA, Gkoutos GV, Acharjee A. -Omics biomarker identification pipeline for translational medicine. *J Transl Med* 2019, 17(1): 1-10.

89. Ma B, Pan J, Hou X, Li C, Xiong T, Gong Y, Song F. The Construction of Polygenic Risk Scores for Breast Cancer Based on LightGBM and Multiple Omics Data. *Res Sqr* 2021, 12(3): 1-34.

90. Zhu J, Djukovic D, Deng L, Gu H, Himmati F, Zaid MA, Chiorean EG, Raftery D. Targeted serum metabolite profiling and sequential metabolite ratio analysis for colorectal cancer progression monitoring. *Anal Bioanal Chem* 2015, 407(26): 7857-63.

91. Kong C, Gao R, Yan X, Qin H. Research progression of blood and fecal metabolites in colorectal cancer. *IJS Oncology* 2018, 3(1): 1-6.

Doktora Tezi Başlığı ve Danışman(lar)ı:

Derin Öğrenme ve Topluluk Öğrenme Yöntemlerine Dayalı Bilgisayar Destekli Tanı Sisteminin Geliştirilmesi: Omik Teknolojileri Üzerine Uygulaması – Prof. Dr. Cemil ÇOLAK

Görevler:

Görev Unvanı	Görev Yeri	Yıl
Arş. Gör.	İnönü Üniversitesi Sağlık Bilimleri Enstitüsü	2014 - …
Uzm. Yrd.	Türkiye Halk Bankası A.Ş.	2012 - 2014

Projelerde Yaptığı Görevler:

1. İnönü Üniversitesi Turgut Özal Tıp Merkezin De Koroner Arter Bypass Operasyonu Yapılan Hastalarda Gelişen Postoperatif Atriyal Fibrilasyonun Tıbbi Bilgi Keşfi Süreci İle Tahmini, Yükseköğretim Kurumları Tarafından Destekli Bilimsel Araştırma Projesi, Araştırmacı: **Ahmet Kadir Arslan**, , 15/11/2017 - 27/11/2020 (Ulusal)

2. Tip 2 Diyabet Mellitus İle İlgili Risk Faktörlerini Saptamada Çok Değişkenli İstatiksel Yöntemlerinin Karşılaştırılması, Yükseköğretim Kurumları Tarafından Destekli Bilimsel Araştırma Projesi, Araştırmacı: Yaşar Şeyma, Araştırmacı: Balıkçı Çiçek İpek, Araştırmacı: **Arslan Ahmet Kadir**, Yürütücü: Yoloğlu Saim, Araştırmacı: Çolak Cemil, Araştırmacı: Şahin İbrahim, , 13/07/2016 - 26/12/2018 (Ulusal)

3. Koroner Arter Hastalığı Risk Faktörleri Üzerinde Lipit Profillerinin Tekli/Çoklu Ortak Etkilerinin Değerlendirilmesi Ve Açık Kaynak Web Tabanlı Bir Karar Destek Yazılımının Geliştirilmesi, -Tübitak 1001, Araştırmacı: **Arslan Ahmet Kadir**, , 27/10/2019 (Devam Ediyor) (Ulusal)

4. Covid-19 Tanısına Yönelik Tıbbi Görüntülere Dayalı Yapay Zekâ Tabanlı Klinik Karar Destek Sistemlerin İn Geliştirilmesi, Yükseköğretim Kurumları Tarafından Destekli Bilimsel Araştırma Projesi, Araştırmacı: **Ahmet Kadir Arslan**, 23/06/2020 (Devam Ediyor) (Ulusal)

5. Obez Bireylerde Atriyal Fibrilasyonu Etkileyebilecek Faktörlerinin Belirlenmesi, Yükseköğretim Kurumları Tarafından Destekli Bilimsel Araştırma Projesi, Araştırmacı: **Ahmet Kadir Arslan**, 15/02/2016 - 30/12/2016 (Ulusal)

6. Biyoistatistik Ve Tıp Bilişiminde Julia Programlama Dili İle Yapay Zekâ Tabanlı Web Yazılımlarının Geliştirilmesi, Yükseköğretim Kurumları Tarafından Destekli Bilimsel Araştırma Projesi, Araştırmacı: **Ahmet Kadir Arslan**, 05/05/2021 (Devam Ediyor) (Ulusal)

Ödüller:

1. Poster birinciliği ödülü, I. Uluslararası BioTürkiye Kongresi, İstanbul, 2020.

2. Poster birinciliği ödülü, XVIII. Ulusal ve I. Uluslararası Biyoistatistik Kongresi, Afyon Kocatepe Üniversitesi, 2016.

ESERLER

A. Uluslararası hakemli dergilerde yayımlanan makaleler:

1. Cansel Neslihan, Ucuz İlknur, **Arslan Ahmet Kadir**, Kayhan Tetik Burcu, Çolak Cemil, Melez Şahide Nur İpek, Gümüştakım Raziye Şule, Ceylan Sinem, Zeren Öztürk Güzin, Kılıç Öztürk Yasemin, Çadırcı Dursun, Demir Akca Ayşe Semra (2021). Prevalence and Predictors Of Psychological Response During Immediate Covıd-19 Pandemic. International Journal Of Clinical Practice, 75(5), Doi: 10.1111/İjcp.13996 (Yayın No: 7102702)

2. Yaşar Şeyma, **Arslan Ahmet Kadir**, Çolak Cemil, Yoloğlu Saim (2020). A Developed Interactive Web Application for Statistical Analysis: Statistical Analysis Software. Middle Black Sea Journal of Health Science, 226-238. (Yayın No: 6694031)

3. Gurunluoglu Semra,Şamdancı Emine, **Arslan Ahmet Kadir**, Akpolat Nusret,Şahin Nurhan,Gökçe Hasan (2020). Prognostic Significance Of Poorly Differentiated Cluster Grading System İn İntestinal Type Gastric Adenocarcinoma. Annals Of Medical Research, 27(8), 2022-2230. (Yayın No: 6731639)

4. Arslan Ahmet Kadir, Küçükakçalı Zeynep,Çolak Cemil (2020). Normal Dağılıma Uygunluğu Değerlendirmek İçin Açık Kaynak Web Tabanlı Yazılım: Normal Dağılımı İnceleme Yazılımı. Fırat Tıp Dergisi, 25(2), 62-68. (Yayın No: 6825339)

5. Arslan Ahmet Kadir, Küçükakçalı Zeynep,Balıkçı Çiçek İpek,Çolak Cemil (2020). A Novel Interpretable Web-Based Tool On The Associative Classification Methods: An Application On Breast Cancer Dataset. The Journal Of Cognitive Systems (Yayın No: 6733593)

6. Arslan Ahmet Kadir, Balıkçı Çiçek İpek,Çolak Cemil (2019). Open Source Web Based Software On Random Assignment Methods And Usage: Random Assignment Software. Turkiye Klinikleri Journal Of Biostatistics, 11(3), 267-274., Doi: 10.5336/Biostatic.2019-70571 (Yayın No: 5639490)

7. Çağın Yasir Furkan,Bilgiç Yılmaz,Berber İlhami,Yıldırım Oğuzhan,Erdoğan Mehmet Ali,Fırat Feyza, **Arslan Ahmet Kadir**, Çolak Cemil,Seçkin Yüksel,Harputluoğlu Muhsin Murat Muhip (2019). The Risk Factors Of Portal Vein Thrombosis İn Patients With Liver Cirrhosis. Experimental And Therapeutic Medicine, 3189-3194., Doi: 10.3892/Etm.2019.7300 (Yayın No: 5143526)

8. Arslan Ahmet Kadir, Tunç Zeynep,Güldoğan Emek,Çolak Cemil (2019). Performance Comparison Of Some Imputation Methods Used İn Missing Value(S) Analysis: A Simulation Study. Turkiye Klinikleri Journal Of Biostatistics, 11(1), 15-23., Doi: 10.5336/Biostatic.2018-62788 (Yayın No: 5521327)

9. Gürünlüoğlu Kubulay,Ceran Özcan Canan,Yıldırım İsmail Okan,Kutlu Ramazan,Saraç Kaya,Yıldız Turan,Bayrakçı Ercan,Taşçı Aytaç, **Arslan Ahmet Kadir**, Demircan Mehmet (2018). Use Of Angiographic Embolization İn Pediatric Abdominal Trauma-İnduced Solid Organ İnjuries. Ulusal Travma Ve Acil Cerrahi Dergisi-Turkish Journal Of Trauma Emergency Surgery, Doi: 10.5505/Tjtes.2018.00056 (Yayın No: 4559386)

10. Arslan Ahmet Kadir, Yaşar Şeyma,Çolak Cemil,Yoloğlu Saim (2018). Wsspas: An Interactive Web Application For Sample Size And Power Analysis With R Using Shiny. Turkiye Klinikleri Journal Of Biostatistics, 3(10), 224-246., Doi: 10.5336/Biostatic.2018-62787 (Yayın No: 4559371)

11. Gürünlüoğlu Kubulay,Bayrakçı Ercan,Koçbıyık Alper,Gökçe Hasan,Taşkapan Mehmet Çağatay,Taşçı Aytaç,Aksungur Zeynep, **Arslan Ahmet Kadir**, Demircan Mehmet (2018). The Effects Of Total Parenteral Nutrition On Telomerase Expression And İn Rabbit. Journal Of Turgut Ozal Medical Center, 25(2), 193-198., Doi: 10.5455/Jtomc.2018.01.021 (Yayın No: 4208685)

12. Çolak Mehmet Cengiz,Karaaslan Erol,Çolak Cemil, **Arslan Ahmet Kadir**, Erdil Nevzat (2017). Handling İmbalanced Class Problem For The Prediction Of Atrial Fibrillation İn Obese Patient. Bıomedıcal Research-Indıa, 28(7), 3293-3299. (Yayın No: 3563883)

13. Özdemir Ramazan,Yağmur Jülide,Açıkgöz Nusret,Cansel Mehmet,Karıncaoğlu Yelda,Ermiş Necip,Pekdemir Hasan, **Arslan Ahmet Kadir** (2017). Relationship Between Serum Homocysteine Levels And Structural-Functional

Carotid Arterial Abnormalities İn İnactive Behçet's Disease. Kardiologia Polska, Doi: 10.5603/Kp.A2017.0227 (Yayın No: 3729261)

14. Güldoğan Emek, **Arslan Ahmet Kadir**, Çolak Mehmet Cengiz,Çolak Cemil,Erdil Nevzat (2017). An İntelligent System For The Classification Of Postoperative Pleural Effusion Between 4 And 30 Days Using Medical Knowledge Discovery.. Bıomedıcal Research-Indıa, 28(4), 1553-1556. (Yayın No: 3563884)

15. Arslan Ahmet Kadir, Çolak Cemil,Sarıhan Mehmet Ediz (2016). Different Medical Data Mining Approaches Based Prediction Of İschemic Stroke. Computer Methods And Programs İn Biomedicine, 130, 87-92., Doi: 10.1016/J.Cmpb.2016.03.022 (Yayın No: 2613482)

16. Fırat Feyza, **Arslan Ahmet Kadir**, Çolak Cemil,Harputluoğlu Hakan (2016). Estimation Of Risk Factors Associated With Colorectal Cancer An Application Of Knowledge Discovery İn Databases. Kuwait Journal Of Science, 43(2), 151-161. (Yayın No: 2843645)

17. Çolak Mehmet Cengiz,Çolak Cemil,Erdil Nevzat, **Arslan Ahmet Kadir** (2016). Investigating Optimal Number Of Cross Validation On The Prediction Of Postoperative Atrial Fibrillation By Voting Ensemble Strategy. Turkiye Klinikleri Journal Of Biostatistics, 8(1), 30-35., Doi: 10.5336/Biostatic.2016-50382 (Yayın No: 3001680)

18. Çolak Cemil, Aydoğan Mustafa Said, **Arslan Ahmet Kadir**, Yücel Aytaç (2015). Application Of Medical Data Mining On The Prediction Of Apache Iı Score. Medicine Science International Medical Journal, 1, Doi: 10.5455/Medscience.2015.04.8274 (Yayın No: 1394340)

B. Uluslararası bilimsel toplantılarda sunulan ve bildiri kitaplarında (proceedings) basılan bildiriler:

1. Büyükçelebi Hakan,Açak Mahmut,Düz Serkan, **Arslan Ahmet Kadir**, Kurak Kemal (2019). Avrupa''nın Altı Farklı Futbol Liginde Mücadele Eden Takımların Oyun İçi Değişkenlerinin Analizi. 17. Uluslararası Spor Bilimleri Kongresi, 182-189. (Tam Metin Bildiri/Sözlü Sunum)(Yayın No:5820948)

2. Arslan Ahmet Kadir, Tunç Zeynep,Çolak Cemil (2019). Open Source Web-Based Software To Evaluate Normal Distribution: Normality Assessment Software. 2019 3rd International Symposium On Multidisciplinary Studies And Innovative Technologies (Ismsıt) (Tam Metin Bildiri/Sözlü Sunum)(Yayın No:5521400)

3. Ucuzal Hasan,Balıkçı Çiçek İpek, **Arslan Ahmet Kadir,** Çolak Cemil (2019). A Web-Based Application For Identifying Objects İn Images: Object Recognition Software. 2019 3rd International Symposium On Multidisciplinary Studies And Innovative Technologies (Ismsıt) (Tam Metin Bildiri/Sözlü Sunum)(Yayın No:5521410)

4. Arslan Ahmet Kadir, Yaşar Şeyma,Çolak Cemil (2019). Breast Cancer Classification Using A Constructed Convolutional Neural Network On The Basis Of The Histopathological Images By An Interactive Web-Based Software. 2019 3rd International Symposium On Multidisciplinary Studies And Innovative Technologies (Ismsıt) (Tam Metin Bildiri/Sözlü Sunum)(Yayın No:5521404)

5. Yaşar Şeyma, **Arslan Ahmet Kadir**, Çolak Cemil,Yoloğlu Saim (2019). A Developed Web-Based Software Can Easily Fulfill The Assumptions Of Correlation, Classification And Regression Tasks İn Data Processing. 2019 International Artificial Intelligence And Data Processing Symposium (Idap), Doi: 10.1109/Idap.2019.8875914 (Tam Metin Bildiri/Sözlü Sunum)(Yayın No:5521317)

6. Kaplan Ali, Güldoğan Emek, Çolak Cemil, **Arslan Ahmet Kadir** (2019). Prediction Of Melanoma From Dermoscopic Images Using Deep Learning-Based Artificial Intelligence Techniques. 2019 International Artificial Intelligence And Data Processing Symposium (Idap), Doi: 10.1109/Idap.2019.8875970 (Tam Metin Bildiri/Sözlü Sunum)(Yayın No:5521292)

7. Arslan Ahmet Kadir, Balıkçı Çiçek İpek,Çolak Cemil (2019). Open Source Web Based Software For Random Assignment/Allocation Methods İn Data Processing. 2019 International Artificial Intelligence And Data Processing Symposium (Idap), Doi: 10.1109/Idap.2019.8875979 (Tam Metin Bildiri/Sözlü Sunum)(Yayın No:5521297)

8. Akbaş Kübra Elif,Kıvrak Mehmet, **Arslan Ahmet Kadir**, Çolak Cemil (2019). Assessment Of Association Rules Based On Certainty Factor: An Application On Heart Data Set. 2019 International Artificial Intelligence And Data

Processing Symposium (Idap), Doi: 10.1109/Idap.2019.8875977 (Tam Metin Bildiri/Sözlü Sunum)(Yayın No:5521315)

9. Arslan Ahmet Kadir, Tunç Zeynep,Çolak Cemil (2019). An Open Sourced Software For Data Transformation And An Application On Simulated Data. 2019 International Artificial Intelligence And Data Processing Symposium (Idap), Doi: 10.1109/Idap.2019.8875876 (Tam Metin Bildiri/Sözlü Sunum)(Yayın No:5521321)

10. Arslan Ahmet Kadir, Yaşar Şeyma,Çolak Cemil (2019). An Intelligent System For The Classification Of Lung Cancer Based On Deep Learning Strategy. 2019 International Artificial Intelligence And Data Processing Symposium (Idap), Doi: 10.1109/Idap.2019.8875896 (Tam Metin Bildiri/Sözlü Sunum)(Yayın No:5521304)

11. Ucuzal Hasan, **Arslan Ahmet Kadir**, Çolak Cemil (2019). Deep Learning Based-Classification Of Dementia İn Magnetic Resonance İmaging Scans. 2019 International Artificial Intelligence And Data Processing Symposium (Idap), Doi: 10.1109/Idap.2019.8875961 (Tam Metin Bildiri/Sözlü Sunum)(Yayın No:5521300)

12. Tosun Nur Saadet,Tanrıverdi Lokman Hekim,Özhan Onural,Vardı Nigar,Parlakpınar Hakan,Yıldız Azibe,Polat Alaadin, **Arslan Ahmet Kadir**, Acet Hacı Ahmet (2018). Does Thalıdomıde Have Beneficıal Or Harmfull Effects On Cısplatın-Induced Cardıotoxıcıty In Rats. 1. Uluslararası Battalgazi Multi Disipliner Çalışmalar Kongresi (Tam Metin Bildiri/Sözlü Sunum)(Yayın No:4634710)

13. Parlakpınar Hakan,İçen Ervanur,Ummuhan Ataa,Özhan Onural,Günata Mehmet,Aladağ Murat,Vardı Nigar,Çiğremiş Yılmaz, **Arslan Ahmet Kadir**, Acet Hacı Ahmet (2018). Investıgatıon Of The Effects Of Apocynın On The Experımental Colıtıs Model Induced By Acetıc Acıd In Rats. Iğdır 1. Uluslararası Multi Disipliner Çalışmalar Kongresi (Tam Metin Bildiri/Sözlü Sunum)(Yayın No:4605128)

14. Parlakpınar Hakan,Başdaş Firdevs,Özhan Onural,Tanrıverdi Hekim,Beytur Ali,Yıldız Azibe,Vardı Nigar,Türköz Yusuf,Üremiş Muhammed Mehdi, **Arslan Ahmet Kadir**, Acet Hacı Ahmet (2018). Sıçanlarda Embelin'in Renal İskemi Ve Reperfüzyonhasarında Koruyucu Ve Tedavi Edici Etkilerininaraştırılması. Iğdır 1. Uluslararası Multi Disipliner Çalışmalar Kongresi (Tam Metin Bildiri/Sözlü Sunum)(Yayın No:4605156)

15. Arslan Ahmet Kadir, Yaşar Şeyma,Çolak Cemil,Yoloğlu Saim (2018). Wsspas: R Shiny Paketi İle Nicel Değişkenlere İlişkin Örneklem Büyüklüğü Ve Güç Analizi Hesaplaması İnteraktif Web Uygulaması. International Statistics Days Conference (Tam Metin Bildiri/Sözlü Sunum)(Yayın No:4586549)

16. Arslan Ahmet Kadir, Tunç Zeynep,Güldoğan Emek,Çolak Cemil (2018). Performance Comparison Of Some Imputation Methods Used İn Missing Value(S) Analysis: A Simulation Study. Isdc2018 (Tam Metin Bildiri/Sözlü Sunum)(Yayın No:4604561)

17. Arslan Ahmet Kadir, Yaşar Şeyma,Çolak Cemil,Yoloğlu Saim (2018). R Shiny Paketi İle Kruskal Wallis H Testi İçin İnteraktif Bir Web Uygulaması. International Statistics Days Conference (Tam Metin Bildiri/Sözlü Sunum)(Yayın No:4586497)

18. Arslan Ahmet Kadir, Çolak Cemil (2017). Sınıf Dengeleyici: İki Sınıflı Verilerde Sınıf Dengesizliği Problemini Gidermek İçin Web Tabanlı Bir Yazılım. Xıx. Ulusal Ve Iı. Uluslararası Biyoistatistik Kongresi (Özet Bildiri/Sözlü Sunum)(Yayın No:3657590)

19. Özdemir Ramazan,Yağmur Jülide,Açıkgöz Nusret,Cansel Mehmet,Karıncalıoğlu Yelda,Ermiş Necip,Pekdemir Hasan, **Arslan Ahmet Kadir** (2017). Relationship Between Serum Homocysteine Levels And Structural-Functional Carotid Arterial Abnormalities İn İnactive Behcet's Disease.. 33. Turkısh Cardıology Congress (Özet Bildiri/Poster)(Yayın No:3998909)

20. Güldoğan Emek, **Arslan Ahmet Kadir**, Çolak Cemil,Yağmur Jülide (2017). Çeşitli Çekirdek Fonksiyonları İle Oluşturulan Destek Vektör Makinesi Modellerinin Performanslarının İncelenmesi: Bir Klinik Uygulama. Xvııı. Uluslararası Ekonometri, Yöneylem Araştırması Ve İstatistik Sempozyumu (Özet Bildiri/Sözlü Sunum)(Yayın No:3657583)

21. Arslan Ahmet Kadir, Güldoğan Emek,Çolak Cemil (2017). Kday: Kayıp Değer Analizinde Kullanılan Çeşitli Tekniklerin Performansını Karşılaştıran Web Tabanlı Bir Yazılım. Xvııı. Uluslararası Ekonometri, Yöneylem Araştırması Ve İstatistik Sempozyumu (Özet Bildiri/Sözlü Sunum)(Yayın No:3657587)

22. Arslan Ahmet Kadir, Çuğlan Songül,Köse Evren,Çolak Cemil (2017). Kronik Obstruktif Akciğer Hastalığının Çeşitli Makine Öğrenmesi Yöntemleri Kullanılarak Sınıflandırılması. Xıx. Ulusal Ve Iı. Uluslararası Biyoistatistik Kongresi (Özet Bildiri/Sözlü Sunum)(Yayın No:3657607)

23. Güldoğan Emek, **Arslan Ahmet Kadir**, Çolak Mehmet Cengiz,Çolak Cemil,Erdil Nevzat (2016). An Intelligent System For The Classification Of Postoperative Pleural Effusion Between 4 And 30 Days Using Medical Knowledge Discovery. I. International Biostatistics Congress (Özet Bildiri/Poster)(Yayın No:3021890)

24. Arslan Ahmet Kadir, Güldoğan Emek,Çolak Cemil (2016). Makine Öğrenmesi Yaklaşımlarından Aşırı Öğrenme Makinesinin Sınıflandırma Performansının Değerlendirilmesi Bir Simülasyon Çalışması. I. International Biostatistics Congress (Özet Bildiri/Poster)(Yayın No:3021767)

25. Çolak Mehmet Cengiz,Karaaslan Erol,Çolak Cemil, **Arslan Ahmet Kadir**, Erdil Nevzat (2016). Handling Imbalanced Class Problem For The Prediction Of Atrial Fibrillation İn Obese Patients. I. International Biostatistics Congress (Özet Bildiri/Poster)(Yayın No:3021997)

26. Çolak Mehmet Cengiz,Çolak Cemil,Erdil Nevzat, **Arslan Ahmet Kadir** (2016). Investigatıng Optimal Number Of Cross Validation On The Predictıon Of Postoperatıve Atrial Fibrillation By Voting Ensemble Strategy. 17. International Symposium On Econometrics, Operations Research And Statistics (Özet Bildiri/Poster)(Yayın No:3021631)

27. Arslan Ahmet Kadir, Çolak Cemil,Sarıhan Mehmet Ediz (2016). Different Medical Data Mining Approaches Based Prediction Of İschemic Stroke. 17. International Symposium On Econometrics, Operations Research And Statistics (Özet Bildiri/Poster)(Yayın No:3020494)

28. Fırat Feyza, **Arslan Ahmet Kadir**, Çolak Cemil,Harputluoğlu Hakan (2016). Estimation Of Risk Factors Associated With Colorectal Cancer An Application Of Knowledge Discovery in Databases. 17. International Symposium On Econometrics, Operations Research And Statistics (Özet Bildiri/Poster)(Yayın No:3020504)

D. Ulusal hakemli dergilerde yayımlanan makaleler:

1. Arslan Ahmet Kadir, Tunç Zeynep,Çolak Cemil (2019). Veri Dönüşümü İçin Açık Kaynak Erişimli Web Tabanlı Yazılım: Veri Dönüşüm Yazılımı. Fırat Üniversitesi Sağlık Bilimleri Tıp Dergisi (Kontrol No: 5521562)

2. Arslan Ahmet Kadir, Tunç Zeynep,Çolak Cemil (2019). Normal Dağılıma Uygunluğu Değerlendirmek İçin Açık Kaynak Web-Tabanlı Yazılım: Normal Dağılımı İnceleme Yazılımı. Fırat Tıp Dergisi (Kontrol No: 5522504)

3. Canbolat Mustafa, **Arslan Ahmet Kadir**, Durmuş Mahmut,Vardı Nigar,Yakıncı Mehmet Cengiz (2018). Tıp Eğitim Müfredatında Koruyucu Sağlık: İnönü Üniversitesi tıp Fakültesi Örneği. İnönü Üniversitesi Sağlık Bilimleri Dergisi, 7(2), 10-13. (Kontrol No: 4437584)

4. Arslan Ahmet Kadir, Yaşar Şeyma,Çolak Cemil,Yoloğlu Saim (2018). R Shiny Paketi İle Kruskal Wallis H Testi İçin İnteraktif Bir web Uygulaması. İnönü Üniversitesi Sağlık Bilimleri Dergisi, 2(7), 49-55. (Kontrol No: 4559405)

5. Güldoğan Emek, **Arslan Ahmet Kadir**, Yağmur Jülide (2017). Çeşitli Çekirdek Fonksiyonları İle Oluşturulan Destek Vektör Makinesi Modellerinin Performanslarının İncelenmesi: Bir Klinik Uygulama. Fırat Tıp Dergisi, 22(3), 136-142. (Kontrol No: 3657612)

E. Ulusal bilimsel toplantılarda sunulan ve bildiri kitaplarında basılan bildiriler:

1. Canbolat Mustafa, **Arslan Ahmet Kadir**, Durmuş Mahmut,Vardı Nigar,Yakıncı Mehmet Cengiz (2018). Koruyucu Sağlığın Tıp Eğitim Müfredatındaki Yeri: İnönü Tıp Örneği. X. Ulusal Tıp Eğitimi Kongresi (Tam Metin Bildiri/Sözlü Sunum)(Yayın No:4421614)

2. Caferoğlu Gürünlüoğlu Semra,Gürünlüoğlu Kubulay,Taşçı Aytaç,Demircan Mehmet,Üremiş Muhammed Mehdi, **Arslan Ahmet Kadir** (2018). Total Parenteral Nütrisyonun Tavşanlarda Telomeraz Seviyesine Etkileri. 1. Gastrointestinal Araştırma Kongresi (Özet Bildiri/Sözlü Sunum)(Yayın No:4271769)

3. Kayhan Tetik Burcu,İnci Coşkun Ebru,Baltacı Hilal,Sertkaya Serap,Gedik Işılay,Selçuk Engin Burak, **Arslan Ahmet Kadir** (2017). Anemisi Olan Ve Olmayan Gebelerde Uyku Ve Yaşam Kalitesinin Değerlendirilmesi. 11. Güz Okulu (Tam Metin Bildiri/Poster)(Yayın No:4379746)

F. Geliştirilen Yazılımlar

1. Örneklem Büyüklüğü ve Güç Analizi Yazılımı (https://biostatapps.inonu.edu.tr/WSSPAS)
2. Bilgi Keşfi Süreci Yazılımı (https://biostatapps.inonu.edu.tr/BKSY)
3. Akciğer Kanseri Sınıflandırma Yazılımı (https://biostatapps.inonu.edu.tr/AKSY)
4. Meme Kanseri Sınıflandırma Yazılımı (https://biostatapps.inonu.edu.tr/MKSY)
5. Sınıf Dengeleyici Analizi Yazılımı (https://biostatapps.inonu.edu.tr/twoclsbalancer)
6. İlişkisel Sınıflandırma Yazılımı (https://biostatapps.inonu.edu.tr/ACS)
7. Rasgele Atama Yazılımı (https://biostatapps.inonu.edu.tr/RAY)
8. Dağılım İnceleme Yazılımı (https://biostatapps.inonu.edu.tr/NDY)
9. İstatistiksel Analiz Yazılımı (https://biostatapps.inonu.edu.tr/IAY)
10. Kruskal-Wallis Yazılımı (https://biostatapps.inonu.edu.tr/kruskalwallis)
11. POAF Risk Tahmin Yazılımı (https://biostatapps.inonu.edu.tr/POAF)
12. Kayıp Değer Analizi ve Atama Yazılımı (https://biostatapps.inonu.edu.tr/KDAY)

Ek-3. Ardışık Kod Dizini (Pipeline)

```
library(foreach)
library(glmnet)
library(reticulate)
library(caret)
library(foreign)
library(xgboost)


numpy=import("numpy",convert = T)
BS=import("BorutaShap",convert = T)
sklearn=import("sklearn",convert = T)

veri=foreign::read.spss("C:/Users/ahmetkadirarslan/Desktop/
tez/ST000284/veri_binomial.sav",
                        use.value.labels = T,to.data.frame
= T)

veri=veri[,c(1:114)]
zero_vars=sapply(2:114, function(x)
length(caret::nearZeroVar(veri[,x])))
one_vars=colnames(veri[-1])[which(sapply(2:114,function(x)
length(which(veri[,x]==1)))>0)]  ### 1 değeri içeren
değişkenler
# "V35" 27 tane 1 değeri içeriyor diğerleri 1 tane
veri=veri[,!colnames(veri)%in%one_vars[2]]

for(i in one_vars[-2]){

  row_index=which(veri[,i]==1)
  veri[row_index,i]=NA

}

set.seed(19880203)

imputed=missForest::missForest(veri[-1],maxiter = 100)$ximp
veri=cbind.data.frame("Group"=veri$Group,imputed)

#for (i in 2:108) {
#  veri[,i]=clusterSim::data.Normalization(veri[,i],"n4")
#}

### lasso ###

X=data.matrix(veri[-1])
y=veri$Group
cv=cv.glmnet(X, y, family = "binomial", nfold = 5, parallel
= TRUE, alpha = 1)
```

```
model=glmnet::glmnet(X, y, alpha = 1, family =
"binomial",lambda = cv$lambda.1se)
coefs=data.matrix(model$beta)
selected_features_lasso=rownames(coefs)[which(coefs[,1]!=0)
]

### elastic-net ###

set.seed(19880203)

X=data.matrix(veri[-1])
y=veri$Group
alphas=seq(0.1, 0.9, 0.05)
search=foreach(i = alphas, .combine = rbind) %dopar% {
  cv=cv.glmnet(X, y, family = "binomial", nfold = 5,
parallel = TRUE, alpha = i)
  data.frame(cvm = cv$cvm[cv$lambda == cv$lambda.1se],
lambda.1se = cv$lambda.1se, alpha = i)
}
cv3=search[search$cvm == min(search$cvm),]

model=glmnet::glmnet(X, y, alpha = cv3$alpha, family =
"binomial",lambda = cv3$lambda.1se)
coefs=sapply(model$beta, as.vector)
remove=which(coefs==0)
selected_features_enet=colnames(X[,-remove])

### Boruta ###

set.seed(19880203)
boruta_output=Boruta::Boruta(Group~., data=veri,
doTrace=2,getImp = Boruta::getImpRfZ)
BO=list(boruta_output,boruta_output$finalDecision)
selected_features_boruta=names(BO[[2]])[which(BO[[2]]=="Con
firmed")]

## BorutaShap ##

set.seed(19880203)
y=numpy$asarray(as.numeric(y))
X=as.data.frame(X)
Feature_Selector =
BS$BorutaShap(importance_measure="shap",#model = model,
                                  classification=TRUE)
Feature_Selector$fit(X=X, y=y, n_trials=100L, sample=FALSE,
                     normalize=FALSE,train_or_test =
"train",verbose=TRUE,
                     random_state = 19880203L)

Feature_Selector$plot(which_features="accepted")
```

```
final_selected=Reduce(intersect,
list(selected_features_boruta,

selected_features_lasso,

selected_features_enet,

Feature_Selector$accepted_columns[[100]]))

var_list=list(

  "all"=colnames(veri[-1]),
  "lasso"=selected_features_lasso,
  "enet"=selected_features_enet,
  "boruta"=selected_features_boruta,
  "borutaShap"=Feature_Selector$accepted_columns[[100]],
  "consensus"=final_selected

)

library(RColorBrewer)
library(VennDiagram)
myCol=brewer.pal(4, "Accent")
venn.diagram(x = list("LASSO"=var_list[[2]],"Elastic-
Net"=var_list[[3]],

"Boruta"=var_list[[4]],"BorutaShap"=var_list[[5]]),
             filename = "Venn.tiff",fill =
c(scales::alpha(myCol[1],0.3), scales::alpha(myCol[2],0.3),

scales::alpha(myCol[3],0.3),scales::alpha(myCol[4],0.3)),
             margin = 0.1,cex = 1,cat.cex=1)


set.seed(19880203)
train_index=caret::createDataPartition(veri$Group,p =
0.8)[[1]]
test_index=c(1:dim(veri)[1])[-train_index]

## xgboost ##

model_xgb=lapply(1:6, function(x) {


data=cbind.data.frame("Group"=veri$Group,veri[,var_list[[x]
]])
  train_data=as.matrix(data[-1][train_index,])
  test_data=as.matrix(data[-1][test_index,])
  xgb.train =
xgb.DMatrix(data=train_data,label=as.integer(data[,1][train
_index])-1)
```

```
  xgb.test =
xgb.DMatrix(data=test_data,label=as.integer(data[,1][test_i
ndex])-1)
  watchlist = list(train=xgb.train, test=xgb.test)
  model=xgb.train(data=xgb.train, booster = "gbtree",
                  #tree_method = 'gpu_hist',
                  nthread = 2, nrounds=1000,
watchlist=watchlist,
                  eval.metric = "error",nfold = 5,
eval.metric = "logloss",
                  objective = "binary:logistic")

  ## Train ##


pred_xgb_raw_train=as.data.frame(predict(model,xgb.train,re
shape=T))

pred_xgb_train=as.factor(ifelse(pred_xgb_raw_train>0.5,1,0)
)
  levels(pred_xgb_train)=c("C", "H")

cm_xgb_train=caret::confusionMatrix(pred_xgb_train,data[,1]
[train_index],positive="C")

  ## test ##


pred_xgb_raw_test=as.data.frame(predict(model,xgb.test,resh
ape=T))

pred_xgb_test=as.factor(ifelse(pred_xgb_raw_test>0.5,1,0))
  levels(pred_xgb_test)=c("C", "H")

cm_xgb_test=caret::confusionMatrix(pred_xgb_test,data[,1][t
est_index],positive="C")

  list(

"train"=list("raw_pred"=pred_xgb_raw_train,"label_pred"=pre
d_xgb_train,"cm"=cm_xgb_train),

"test"=list("raw_pred"=pred_xgb_raw_test,"label_pred"=pred_
xgb_test,"cm"=cm_xgb_test)
  )

})

names(model_xgb)=c("none","lasso","enet","boruta","borutaSh
ap","consensus")
```

```
lapply(1:6, function(i) plotROC(labels =
as.numeric(veri[,1][train_index])-1,predictions =
model_xgb[[i]][[1]]$raw_pred[,1]))


## H2o ##

library(h2o)
h2o.init(port=8888)  ###
https://stackoverflow.com/questions/37779076/oserror-
version-mismatch-while-installing-h2o

model_h2o=lapply(1:6, function(x) {


data=cbind.data.frame("Group"=veri$Group,veri[,var_list[[x]
]])
  train_data=data[train_index,]
  test_data=data[test_index,]
  train_data=as.h2o(train_data)
  test_data=as.h2o(test_data)
  model_h2o=h2o.deeplearning(x = 2:dim(data)[2], y = 1,
training_frame = train_data,nfolds = 5,seed = 19880203)
  #perf=h2o.performance(model_h2o)

  ## train ##

  pred_train=as.data.frame(h2o.predict(model_h2o, newdata =
train_data[-1]))
  pred_h2o_raw_train=sapply(1:dim(pred_train)[1],
function(i)
pred_train[i,which(pred_train[i,1]==colnames(pred_train))])
  pred_h2o_train=as.factor(pred_train[,1])

cm_h2o_train=caret::confusionMatrix(pred_h2o_train,data[,1]
[train_index],positive="C")

  ## test ##

  pred_test=as.data.frame(h2o.predict(model_h2o, newdata =
test_data[-1]))
  pred_h2o_raw_test=sapply(1:dim(pred_test)[1], function(i)
pred_test[i,which(pred_test[i,1]==colnames(pred_test))])
  pred_h2o_test=as.factor(pred_test[,1])

cm_h2o_test=caret::confusionMatrix(pred_h2o_test,data[,1][t
est_index],positive="C")

  list(
    "model"=model_h2o,
```

```
"train"=list("raw_pred"=pred_h2o_raw_train,"label_pred"=pre
d_h2o_train,"cm"=cm_h2o_train),

"test"=list("raw_pred"=pred_h2o_raw_test,"label_pred"=pred_
h2o_test,"cm"=cm_h2o_test)
  )

})


### lightgbm ##

set.seed(19880203)

library(treesnip)
model=parsnip::decision_tree()
parsnip::set_engine(model, "tree")
# boost_tree
model=parsnip::boost_tree(mtry = 1, trees = 1000)
#parsnip::set_engine(model, "catboost")

parsnip::set_engine(model, "lightgbm")

library(lightgbm)

params=list(objective = "binary", metric = "binary_error")

model_lgbm=lapply(1:6, function(x) {


data=cbind.data.frame("Group"=veri$Group,veri[,var_list[[x]
]])
  train_data=data[train_index,]
  train_data=data[train_index,]
  test_data=data[test_index,]
  train_data$Group=as.numeric(train_data$Group)-1
  test_data$Group=as.numeric(test_data$Group)-1
  dtrain=lightgbm::lgb.Dataset(data =
as.matrix(train_data[-1]),label=train_data[,1])
  dtest = lightgbm::lgb.Dataset(data =
as.matrix(test_data[-1]),label=test_data[,1])
  lgb.model = lgb.train(
    params=params,
    data=dtrain,
    nrounds=1000,nfold = 5L,
    #num_threads=8,
    valids=list(test=dtest)
    #early_stopping_rounds=5
    #eval=F1_metric,
  )
```

```
  ## Train ##


pred_lgbm_raw_train=as.data.frame(predict(lgb.model,as.matr
ix(train_data[-1]),reshape=T))

pred_lgbm_train=as.factor(ifelse(pred_lgbm_raw_train>0.5,1,
0))
  levels(pred_lgbm_train)=c("C", "H")

cm_lgbm_train=caret::confusionMatrix(pred_lgbm_train,data[,
1][train_index],positive="C")


## test ##


pred_lgbm_raw_test=as.data.frame(predict(lgb.model,as.matri
x(test_data[-1]),reshape=T))

pred_lgbm_test=as.factor(ifelse(pred_lgbm_raw_test>0.5,1,0)
)
  levels(pred_lgbm_test)=c("C", "H")

cm_lgbm_test=caret::confusionMatrix(pred_lgbm_test,data[,1]
[test_index],positive="C")

  list(

"train"=list("raw_pred"=pred_lgbm_raw_train[,1],"label_pred
"=pred_lgbm_train,"cm"=cm_lgbm_train),

"test"=list("raw_pred"=pred_lgbm_raw_test[,1],"label_pred"=
pred_lgbm_test,"cm"=cm_lgbm_test)
  )


})


### deepnet ###

model_sae=lapply(1:6, function(x) {


data=cbind.data.frame("Group"=veri$Group,veri[,var_list[[x]
]])
  train_data=data[train_index,]
  test_data=data[test_index,]
```

```
model_sae=caret::train(Group~.,data=train_data,method="dnn"
,tuneLength=100L,

trControl=caret::trainControl("cv",5,search = "random"))
  pred_sae=predict(model_sae,test_data[-1])

cm_sae=caret::confusionMatrix(pred_sae,test_data[,1],positi
ve="C")

  ## train ##


pred_train=cbind.data.frame(predict(model_sae,train_data[-
1]),predict(model_sae,train_data[-1],type = "prob"))
  pred_sae_raw_train=sapply(1:NROW(pred_train), function(i)
pred_train[i,which(pred_train[i,1]==colnames(pred_train))])
  pred_sae_train=predict(model_sae,train_data[-1])

cm_sae_train=caret::confusionMatrix(pred_sae_train,data[,1]
[train_index],positive="C")

  ## test ##

  pred_test=cbind.data.frame(predict(model_sae,test_data[-
1]),predict(model_sae,test_data[-1],type = "prob"))
  pred_sae_raw_test=sapply(1:NROW(pred_test), function(i)
pred_test[i,which(pred_test[i,1]==colnames(pred_test))])
  pred_sae_test=predict(model_sae,test_data[-1])

cm_sae_test=caret::confusionMatrix(pred_sae_test,data[,1][t
est_index],positive="C")

  list(

"train"=list("raw_pred"=pred_sae_raw_train,"label_pred"=pre
d_sae_train,"cm"=cm_sae_train),

"test"=list("raw_pred"=pred_sae_raw_test,"label_pred"=pred_
sae_test,"cm"=cm_sae_test)
  )

})

data_train=lapply(1:6, function(i)
cbind.data.frame("pred"=model_xgb[[i]][[1]]$label_pred,"rea
l"=veri$Group[train_index]))
data_test=lapply(1:6, function(i)
cbind.data.frame("pred"=model_xgb[[i]][[2]]$label_pred,"rea
l"=veri$Group[test_index]))
```

```
res=lapply(1:6, function(i) {


tab=xtabs(paste0("~",paste0(colnames(data_test[[i]]),collap
se = "+")),data = data_test[[i]])
  epiR::epi.tests(tab,conf.level = 0.95)

})


## Alternatif güven aralığı hesaplama fonksiyonları ##

acc=function(data,R=1000,conf = 0.95,type = "basic"){

  f=function(data,indices){
    data=data[indices,]
    tab=xtabs(paste0("~",paste0(colnames(data),collapse =
"+")),data = data)
    sum(diag(tab))/sum(tab)
  }

  r1=boot(data=data, statistic=f,R=R)
  trycatch=is.null(try(boot.ci(r1,conf = conf,type =
type)))
  if(trycatch==T){
    f(data)
  }else{
    r2=boot.ci(r1,conf = conf,type = type)
    paste0(round(r2$t0,3)," (",round(r2$basic[4],3),"-
",round(r2$basic[5],3),")")
  }

}

sens=function(data,R=1000,conf = 0.95,type =
"basic",positive="C"){

  f=function(data,indices){
    data=data[indices,]
    tab=xtabs(paste0("~",paste0(colnames(data),collapse =
"+")),data = data)
    caret::sensitivity(tab,positive = positive)
  }

  r1=boot(data=data, statistic=f,R=R)
  trycatch=is.null(try(boot.ci(r1,conf = conf,type =
type)))
  if(trycatch==T){
    f(data)
  }else{
    r2=boot.ci(r1,conf = conf,type = type)
```

```
    paste0(round(r2$t0,3)," (",round(r2$basic[4],3),"-
",round(r2$basic[5],3),")")
  }

}

spec=function(data,R=1000,conf = 0.95,type =
"basic",positive="C"){

  f=function(data,indices){
    data=data[indices,]
    tab=xtabs(paste0("~",paste0(colnames(data),collapse =
"+")),data = data)
    caret::specificity(tab,negative =
dimnames(tab)[[1]][which(dimnames(tab)[[1]]!=positive)],ref
erence=names(dimnames(tab)[2]))
  }

  r1=boot(data=data, statistic=f,R=R)
  trycatch=is.null(try(boot.ci(r1,conf = conf,type =
type)))
  if(trycatch==T){
    f(data)
  }else{
    r2=boot.ci(r1,conf = conf,type = type)
    paste0(round(r2$t0,3)," (",round(r2$basic[4],3),"-
",round(r2$basic[5],3),")")
  }

}

PPV=function(data,R=1000,conf = 0.95,type =
"basic",positive="C"){

  f=function(data,indices){
    data=data[indices,]
    tab=xtabs(paste0("~",paste0(colnames(data),collapse =
"+")),data = data)
    caret::posPredValue(tab,positive = positive)
  }

  r1=boot(data=data, statistic=f,R=R)
  trycatch=is.null(try(boot.ci(r1,conf = conf,type =
type)))
  if(trycatch==T){
    f(data)
  }else{
    r2=boot.ci(r1,conf = conf,type = type)
    paste0(round(r2$t0,3)," (",round(r2$basic[4],3),"-
",round(r2$basic[5],3),")")
  }
```

```
}

NPV=function(data,R=1000,conf = 0.95,type =
"basic",positive="C"){

  f=function(data,indices){
    data=data[indices,]
    tab=xtabs(paste0("~",paste0(colnames(data),collapse =
"+")),data = data)

caret::negPredValue(tab,negative=dimnames(tab)[[1]][which(d
imnames(tab)[[1]]!=positive)])
  }

  r1=boot(data=data, statistic=f,R=R)
  trycatch=is.null(try(boot.ci(r1,conf = conf,type =
type)))
  if(trycatch==T){
    f(data)
  }else{
    r2=boot.ci(r1,conf = conf,type = type)
    paste0(round(r2$t0,3)," (",round(r2$basic[4],3),"-
",round(r2$basic[5],3),")")
  }

}

F_meas=function(data,R=1000,conf = 0.95,type =
"basic",positive="C"){

  f=function(data,indices){
    data=data[indices,]
    tab=xtabs(paste0("~",paste0(colnames(data),collapse =
"+")),data = data)
    caret::F_meas(tab,relevant=positive)
  }

  r1=boot(data=data, statistic=f,R=R)
  trycatch=is.null(try(boot.ci(r1,conf = conf,type =
type)))
  if(trycatch==T){
    f(data)
  }else{
    r2=boot.ci(r1,conf = conf,type = type)
    paste0(round(r2$t0,2)," (",round(r2$basic[4],2)," -
",round(r2$basic[5],2),")")
  }

}
```

```
## plotROC ##

par(pty="s",mfrow=c(1,2))
axis(1, at = seq(0, 1, by = 0.2),outer = F)

plotROC=function(labels,predictions,title){

  library(pROC)
  set.seed(19880203)
  pROC_obj=roc(labels,predictions,
               smoothed = T,legacy.axes = T,
               ci=T, ci.alpha=0.95, stratified=T,
               plot=T, auc.polygon=T, max.auc.polygon=T,
grid=F,
               print.auc=F, show.thres=T,xlab="1-
Seçicilik",ylab="Duyarlılık",
               main=title
               )


  sens.ci=ci.se(pROC_obj,conf.level=0.95, boot.n=1000)
  plot(sens.ci, type="shape", col="lightblue")
  plot(sens.ci)
  print(pROC_obj)

}

split=list(
  "train"=train_index,
  "test"=test_index
)
title=c("Eğitim veri seti","Test veri seti")

method=c(
  "Hiçbiri","LASSO","Elastic-
Net","Boruta","BorutaShap","Uzlaşmacı (Konsensüs)"
)

res=lapply(1:6, function(x){

  lapply(1:2, function(i) plotROC(labels =
as.numeric(veri[,1][split[[i]]])-1,
                                  predictions =
model_h2o[[x]][[i]]$raw_pred,title = title[i]))

  title(paste("Değişken seçim yöntemi:", method[x]),line=-
2, side=2, outer=TRUE, cex=2)

})
```