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

OMİK VERİLERİNDE OTOMATİK MAKİNE ÖĞRENİMİ ALGORİTMALARININ PERFORMANSININ DEĞERLENDİRİLMESİ

Hazırlayan
Meltem ÜNLÜSAVURAN

Danışman
Doç.Dr. Gökmen ZARARSIZ

Yüksek Lisans Tezi

Aralık 2019
KAYSERİ

T.C.
ERCİYES ÜNİVERSİTESİ
SAĞLIK BİLİMLERİ ENSTİTÜSÜ
BİYOİSTATİSTİK BİLİM DALI

OMİK VERİLERİNDE OTOMATİK MAKİNE ÖĞRENİMİ ALGORİTMALARININ PERFORMANSININ DEĞERLENDİRİLMESİ

(Yüksek Lisans Tezi)

Hazırlayan
Meltem ÜNLÜSAVURAN

Danışman
Doç.Dr. Gökmen ZARARSIZ

Aralık 2019
KAYSERİ

BİLİMSEL ETİĞE UYGUNLUK

Bu tezin kendi çalışmam olduğunu, tüm bilgilerin akademik ve etik kurallara uygun bir şekilde elde edildiğini beyan ederim. Aynı zamanda akademik ve etik kuralların gerektirdiği gibi tüm materyal ve sonuçları tam olarak aktardığımı, başkalarının eserlerinden yararlanılması durumunda ilgili eserlere bilimsel kurallara uygun olarak atıfta bulunduğumu ve kaynaklar listesinde gösterdiğimi belirtirim.

Adı Soyadı: Meltem ÜNLÜSAVURAN

İmza: 

YÖNERGEYE UYGUNLUK ONAYI

"Omik Verilerinde Otomatik Makine Öğrenimi Algoritmalarının Performansının Değerlendirilmesi" adlı Yüksek Lisans tezi, Erciyes Üniversitesi Lisansüstü Tez Önerisi ve Tez Yazma Yönergesi'ne uygun olarak hazırlanmıştır.

Tezi Hazırlayan

Meltem ÜNLÜSAVURAN

Danışman

Doç.Dr. Gökmen ZARARSIZ

Halk Sağlığı Anabilim Dalı / Biyoistatistik Bilim Dalı Başkanı

Prof.Dr. Ahmet ÖZTÜRK

Doç.Dr. Gökmen ZARARSIZ danışmanlığında **Meltem ÜNLÜSAVURAN** tarafından hazırlanan "Omik Verilerinde Otomatik Makine Öğrenimi Algoritmalarının Performansının Değerlendirilmesi" adlı bu çalışma jürimiz tarafından Erciyes Üniversitesi Sağlık Bilimleri Enstitüsü Halk Sağlığı Anabilim Dalı / Biyoistatistik Bilim Dalında yüksek lisans tezi olarak kabul edilmiştir.

27 / 12/ 2019

JÜRİ:

Danışman :Doç.Dr. Gökmen ZARARSIZ

(Erciyes Üniversitesi Biyoistatistik AD)

Üye :Prof.Dr. Ahmet ÖZTÜRK

(Erciyes Üniversitesi Biyoistatistik AD)

Üye : Dr.Öğr.Üyesi Vahap ELDEM

(İstanbul Üniversitesi Fen Fakültesi Biyoloji Bölümü)

ONAY :

Bu tezin kabulü Enstitü Yönetim Kurulunun ..... tarih ve ..... sayılı kararı ile onaylanmıştır.

...../...../.....

Prof.Dr. Bilal AKYÜZ

Enstitü Müdürü

TEŞEKKÜR

Bana her konuda her zaman destek olan danışmanım olmanın da ötesinde yaşam koçluğu yapan, kişiliği ve başarılarıyla ile rol model ve ilham kaynağı olan danışman hocam Doç.Dr. Gökmen ZARARSIZ'a yönlendirmeleri ve destekleri için her zaman minnet duyacak olup teşekkür ederim.

Tecrübeleri ile bana her zaman yol gösterici olan, bölümde aile ortamında hissettirecek içtenlik ve samimiyetle, bir baba yaklaşımıyla hep arkamda olan bölüm başkanımız Prof.Dr. Ahmet ÖZTÜRK'e üzerimde ki emeklerinden dolayı teşekkürü borç bilirim.

Yüksek lisans eğitimim boyunca yanımda olan, öneri ve katkılarda bulunan Öğr.Gör. Gözde E. ZARARSIZ'a, desteğini esirgemeyen Ar.Gör.Dr Dinçer GÖKSÜLÜK'e çok teşekkür ederim. Yüksek lisansa başladığım ilk zamanlarda birlikte yol katettiğim, ortak çalışmaktan çok keyif aldığım Meryem ÇAVUŞOĞLU'na, ince düşünceleriyle mutlu etmeyi bilen, dostluğunu esirgemeyen M. Yasemin A. SEYFELİ'ye, tez yazım aşamam da desteğini esirgemeyen Ahu DURMUŞÇELEBİ'ye, yol arkadaşım olan Funda İPEKTEN'e, karşılaştığım her türlü problemle ilgili bana her zaman destek olan Cem SÖNMEZ'e katkılarından dolayı teşekkür ederim.

Beni bu günlere getiren, her kararımda arkamda durup bana her zaman destek olan kocaman aileme sonsuz teşekkürler.

OMİK VERİLERİNDE OTOMATİK MAKİNE ÖĞRENİMİ ALGORİTMALARININ PERFORMANSININ DEĞERLENDİRİLMESİ

Meltem ÜNLÜSAVURAN

Erciyes Üniversitesi, Sağlık Bilimleri Enstitüsü

Biyoistatistik Anabilim Dalı

Yüksek Lisans Tezi, Aralık 2019

ÖZET

Omik çalışmalar sağlık alanında birçok problemi çözmeye yardımcı olmuştur. Omik teknolojiler yüksek miktarda veri üretmektedirler. Verilerden anlamlı bilginin elde edilmesi için istatistiksel analiz yöntemlerine başvurulmaktadır. Makine öğrenmesi, omik verilerden anlamlı bilgilerin elde edilmesi için sıklıkla kullanılmaktadır.

Bu çalışma kapsamında sınıflandırma problemleri ele alınmıştır. Sınıflandırma problemlerini çözmek için çok sayıda makine öğrenmesi yöntemi bulunmaktadır. Araştırmacıların bu yöntemler arasından hangisini uygulayacaklarına karar vermeleri gerekmektedir. Uygulanacak yöntemlerin, ayarlanması gereken çeşitli parametreleri bulunmaktadır. Bu yöntem ve parametrelerden en uygun seçimin yapılması için otomatik makine öğrenmesi yöntemleri geliştirilmiştir.

Otomatik makine öğrenmesi ile veri analiz sürecinin otomatik gerçekleştirilmesi hedeflenmiştir. Mevcut geliştirilmiş otomatik makine öğrenmesi yöntemleri, makine öğrenme sürecinin ön işleme, değişken seçimi, model seçimi, parametre seçimi gibi bazı basamaklarını otomatikleştirmiştir. Biz de bu çalışmada otomatik makine öğrenmesi yöntemlerinin omik veriler için performansını araştırdık.

Kullanılan toplam 29 omik verisinden 16'sı mikrodizi verisi, 6'sı RNA-dizileme verisi ve 7'si metabolomik verisidir. Verilerin tamamı gerçek veri setleridir. Bu verileri sınıflandırmak için otomatik makine öğrenmesi yöntemlerinden H2O ve TPOT kullanılmıştır. Makine öğrenmesi yöntemlerinden ise RF, DVM ve NSC yöntemleri kullanılmıştır. Her bir yöntemin birbirlerine karşı avantaj ve dezavantajlarının araştırılması amaçlanmıştır.

Analiz sonuçlarına göre mikrodizi verilerinden Alizadeh-V1, Armstrong-V1, Armstrong-V2, Bittner, Chen, Chowdary, Garber, Gordon, Laiho, Lapointe-V1 ve West; RNA dizileme verilerinden Alzheimer, Fare Embriyo, Fare Kök Hücre, Rahim Ağzı Kanseri ve Lemfoblastoid; metabolomik verilerinden ST000369, ST000389, ST000388, ST000390, ST000356 ve ST000391 ile H2O ve TPOT yöntemiyle en iyi sınıflandırma performansları elde edilmiştir. Toplam 11 veride ise RF, DVM ve NSC yöntemleriyle en iyi sınıflandırma performansları elde edilmiştir. Sonuç olarak her veri için kullanılabilecek en iyi performansı gösteren bir yaklaşım bulunamamıştır. Daha kesin değerlendirmelerin yapılabilmesi için gelecek çalışmalarda simülasyon çalışmaları ve daha farklı veri setleri ile denemelerin yapılması gerekmektedir.

Anahtar Kelimeler: AutoML, Gen ifade verisi, Makine öğrenmesi, Omik veri, Sınıflandırma

PERFORMANCE EVALUATION OF AUTOMATED MACHINE LEARNING ALGORITHMS IN OMICS DATA

Meltem ÜNLÜSAVURAN

Erciyes University, Graduate School of Health Sciences

Department of Biostatistics

Master of Thesis, December 2019

ABSTRACT

Omics studies have helped solve many health problems. Omics technologies create large amounts of data. In order to obtain meaningful information from the data, statistical analysis methods are used. Machine learning is often used to obtain meaningful information from omics data.

In this study, classification problems are discussed. Many machine learning methods exist to solve classification problems. Researchers must decide which of these methods to use. Methods to be applied have several parameters that need to be set. Automated machine learning methods have been developed in order to make the most appropriate selection of these methods and parameters.

Automated machine learning is aimed at automating the data analysis process. The current developed automated machine learning methods have automated some of the steps of the machine learning process such as pre-processing, feature selection, model selection, parameter selection. In this study, we researched the performance of automated machine learning methods for omics data.

Of the 29 omics data used, 16 were microarray data, 6 were RNA-sequencing data and 7 were metabolomics data. All data are real datasets. Automatic machine learning methods H2O and TPOT have been used to classify these data. Machine learning methods RF, SVM and NSC have been used to classify these data. It is aimed to research the advantages and disadvantages of each method against each other.

According to the analysis results, the best classification performances for Alizadeh-V1, Armstrong-V1, Armstrong-V2, Bittner, Chen, Chowdary, Garber, Gordon, Laiho, Lapointe-V1 and West from microarray data, Alzheimer, Mouse Embryo, Mouse Stem

Cell, Cervical and Lymphoblastoid from RNA sequencing datasets, ST000369, ST000389, ST000388, ST000390, ST000356 and ST000391 from metabolomics data were obtained by H2O and TPOT methods. The best classification performance in 11 datasets were obtained with RF, DVM and NSC methods. As a result, the best performing approach for each datasets could not be found. In order to make more accurate assessments, in the future studies should be conducted with simulation studies and experiments with different data sets.

Keywords: AutoML, Gene expression data, Machine learning, Omics data, Classification

İÇİNDEKİLER

KISALTMALAR VE SİMGELER

AID: Automatic Interaction Detection

ALL: Akut Lenfoblastik Lösemi

AML: Akut Miyoleid Lösemi

API: Uygulama Kullanıcı Arayüzü

ASR: Otomatik Konuşma Tanıma

AutoML: Otomatik Makine Öğrenmesi

bACC: Dengeli Doğruluk Oranı

BOHB: Çoklu Uygunluk Optimizasyonunu ve Bayesian Optimizasyonunu

CART: Sınıflandırma ve Regresyon Ağaçları

CASH: Kombine Algoritma Seçimi ve Hiperparametre Optimizasyonu

cDNA: Tamamlayıcı Deoksiribo Nükleik Asik

CGS: Kartezyen Izgara Arama

CHAID: Chi-Squared Automatic Interacton Detection

Da: Dalton (Atamik Kütle Birimi)

DBN: Derin İnanç Ağları

DBSCAN : Gürültülü Uygulamaların Yoğunluk Tabanlı Uzaysal Kümelenmesi

DEA: Anlamlılık Analizi

DESeq: Differential Gene Expression Analysis

DLBCL: Yaygın Büyük B Hücreli Lenfoma

DNA: Deoksiribo Nükleik Asik

DNN: Derin Sinir Ağları

DRF: Varsayılan Random Forest

DRF: Yoğun Rastgele Ormanlar

DVM: Destek Vektör Makineleri

F1 skoru: Kesinlik ve Duyarlılık Harmonik Ortalaması

FDA: Esnek Ayırma Analizi

FL: Foliküler Lenfoma

GA: Genetik Algoritma

GBM: Gradiyent Artırma Makineleri

GC/QTOF/MS: Gaz Kromatografi – Kuadratik Kutuplu Uçuş Zamanlı Kütle Spektrometresi

GC/TOF/MS: Gaz Kromatografi – Uçuş Zamanlı Kütle Spektrometresi

GC-MS: Gaz Kromatografi-Kütle Spektrofotometri

GLM: Genelleştirilmiş Doğrusal Model

HILIC/QTOF/MS: Hidrofilik Etkileşim Sıvı Kromatografisi- Kuadratik Kutuplu Uçuş Zamanlı Kütle Spektrometresi

ID3: Iterative Dichotomiser 3

IG: Bilgi Kazancı

KLL: Kronik Lenfositik Lösemi

KNN: K-En Yakın Komşu Algoritması

LC-MS: Sıvı Kromatografi-Kütle Spektrofotometri

LDA: Doğrusal Ayırma Analizi

LLE: Yerel-Doğrusal Gömme

MCC: Matthew Korelasyon Katsayısı

miRNA: Mikro Ribonükleik Asit

MRMR: Minimum Kalabalık Maksimum İlgililik (Minimum Redundancy Maximal Relevancy)

mRNA: Mesajcı Ribonükleik Asit

NGS: Yeni Nesil Dizileme

NLU: Doğal Dil Anlayışı

NMR: Nükleer Manyetik Rezonans

NSC: En Yakın Küçültülmüş Merkezler

PCA: Temel Bileşenler Analizi

QDA: Karesel Ayırma Analizi

RBM: Sınırlandırılmış Boltzmann Makineleri

RDA: Düzenlenmiş Ayırma Analizi

RF: Random Forest

RFE: Yinelemeli Değişken Seçimi

RGS: Rastgele Izgara Araması

RNA: Ribonükleik Asit

RoBO: Sağlam Bayes Optimizasyonu

SCRDA: Küçültülmüş Merkezlerin Düzenlenmesiyle Ayırma Analizi

scRNA: Tek Hücreli Rinonükleik Asit

SMAC: Sıralı Model Tabanlı Algoritma Konfigürasyonu

TPOT: Ağaç Tabanlı İş Akışı Optimizasyon Aracı

TRUBA: Türk Ulusal e-Bilim e-Altyapısı

t-SNE: T-dağıtık Stokastik Komşu Gömme

UI: Kullanıcı Arayüzü

ULAKBİM: Ulusal Akademik Ağ ve Bilgi Merkezi

vst: Varyans-Dengeleyici Dönüşüm

XRT: Extremely Randomized Forest

YSA: Yapay Sinir Ağları

TABLOLAR LİSTESİ

ŞEKİLLER LİSTESİ

1.GİRİŞ VE AMAÇ

Canlının biyolojik yapısını ve sistematik olarak işleyişini anlamaya yönelik yapılan çalışmalar neticesinde zamanla genomik, transkriptomik, proteomik, epigenomik, metagenomik ve metabolomik gibi çalışma alanları oluşmuştur. Bu alanlarda yapılan çalışmalar yardımıyla sağlıklı olma-olmama durumu, çeşitli hastalıkların moleküler mekanizmasının aydınlatılması, canlı-çevre ilişkisi gibi konular açıklanmaya çalışılmaktadır. Bu çalışmaların kapsamlılığı ve açıklayıcılığı bakımından tek bir omik bilimi ile değerlendirme yapmaktansa birden çok omik bilimi yaklaşımı ile konuyu ele almak daha faydalı olacaktır (Ge ve ark., 2013). Literatürde omik bilimleri ile yapılan çalışmalara bakıldığı zamanda çoklu omik çalışmaların artışı göze çarpmaktadır. Bunun nedeni tek bir omik bilimi; tüm genetik bilginin bilinmesi ile canlıyı tamamen anlamanın mümkün olmayacağının anlaşılmasıdır. Yani canlıyı etkileyebilecek genetik özellikleri ile birlikte beslenme durumu, çevresel etkenler gibi birçok etken birlikte değerlendirilmelidir (Manzoni ve ark., 2018). Bu nedenle bu tez kapsamında tek bir omik bilimi açısından konuya odaklanmaktansa büyük resmin anlaşılabilmesi için birbirinin bütünleyicisi olan birçok omik bilimi açısından konuyu ele almayı hedefledik. Burada amaç her bir omik biliminden elde edilen verileri ayrı ayrı en iyi şekilde sentez edebilmek ve elde edilen nihai sonuçlardan yola çıkarak araştırmacıya bütünsel bir bakış açısı kazandırabilmektir.

Omik teknolojiler binlerce ya da milyonlarca makro ve mikro molekülün tespit edilmesini sağlamaktadır. Omik teknolojilerin kullanımıyla birlikte genomik, proteomik, metabolomik, transkriptomik ve diğer omik yaklaşımlarının kullanımı, araştırılmakta olan örneğe dair küresel bir resim sağlar. Sistem biyolojisi, biyolojik sistemlerdeki çoklu etkileşimlerin araştırılması ve açıklanmasına odaklanan biyoloji, mühendislik, bilgisayar bilimi, biyoinformatik, fizik gibi alanlardan faydalanan çok disiplinli bir çalışma alanıdır. Sistem biyoloji yaklaşımı, organizmayı bir bütün olarak ele alır ve omik araştırmasıyla temel biyolojik bilgiyi birleştirir (Buriani ve ark., 2012).

Bir biyolojik sistemdeki problemleri anlayıp çözebilmek için geleneksel yöntem ve teknolojiler yerini yüksek performanslı yeni yaklaşımlara bırakmıştır. Sistem biyolojisi geleneksel bir yaklaşımdan ziyade tüm sürecin daha iyi anlaşılması için bütünsel ve bütünleştirici yeni yaklaşımları kullanır. Omik teknolojileri kapsamlı analizler için çeşitli avantajlar ve dezavantajlara sahip önemli araçlardır. Genomik, transkriptomik, proteomik ve metabolomik teknoloji platformu, yüksek verimli teknolojilerdir (Jerez, 2008).

Omik teknolojilerin temel amacı spesifik bir biyolojik örnekte bulunan tüm gen, transkript, protein ve metabolitlerin tanımlanmasıdır. Güçlü omik teknolojileri, biyobelirteç keşfi, hücresel metabolizma, kanserin erken teşhisi, hücre gelişimi ve hücre ölümü gibi durumlarla ilişkili sinyal moleküllerinin belirlenmesini sağlayabilecek yeni bir bakış açısı kazandırmıştır. Omikler yalnız biyolojik süreçlerin anlaşılmasına katkı sağlamakla kalmayıp aynı zamanda hastalıkların erken teşhisi, daha doğru bir şekilde teşhis edilmesi ve tedavi edilmesi olasılığını bir gerçeklik haline getirmiştir. Biyolojik sistemi anlamaya dayalı hastalıkların tanısını belirleme, ilerleyişini öngörme ve önleme, ilaç geliştirme, gibi birçok problem için omik biyobelirteçlerinin belirlenmesi ve bu belirteçlere dayalı sınıflandırma işlemine ihtiyaç duyulmaktadır. Omik alanlarında çeşitli cihazlar ile örneklerin analiz edilmesinden sonra elde edilen verinin çeşitli ön işleme aşamaları ile sınıflandırma işlemine hazır hale getirilmesi gerekmektedir. Yüksek boyutlu, gürültülü ve ilişkili yapıya sahip olarak elde edilen bu verilerden biyolojik bilginin sınıflandırma yöntemleri ile elde edilmesi zorlu bir süreçtir (Başaran ve ark., 2010). Bu nedenle bu tür büyük verilerde bu işleri otomatikleştirmek hem akademiye hem endüstriye ivme kazandıracaktır.

Omik bilimlerini çalışılabilir hale getiren teknolojik gelişmeler ve bu alanda yapılan çalışmalardan elde edilen kapsamlı ve aydınlatıcı bilgiler araştırmacıların bu alana yönelik çalışma motivasyonlarını arttırmaktadır. Yapılan çalışmaların artmasıyla birlikte bu çalışmalardan elde edilen veri miktarı da aynı oranda artış göstermektedir. Elde edilen bu verilere bakılarak çalışmanın doğru bir şekilde yapılıp yapılmadığına yönelik fikir sahibi olabildiğimiz gibi çalışmanın nihai sonucunu da yine bu veri setlerine bakarak söyleyebiliriz. Bu nedenle elde edilen verilerin doğru ve profesyonel yaklaşımlarla kapsamlı bir şekilde değerlendirilmesi önemlidir. Elde edilen veriler çalışmadan çalışmaya farklılık göstermekle birlikte genelde oldukça büyük boyutlarda

elde edilmektedir. Çok yüksek boyutlu bir ya da birkaç veriden anlamlı bilginin elde edilmesi oldukça titiz ve zaman alıcı çalışmalar gerektirmektedir. Ayrıca bu verilerin yapısı gereği bilinen klasik istatistiksel yöntemlerden ziyade farklı istatistiksel yaklaşımlar gerektirmektedir. Her omik biliminde elde edilen verinin yapısına göre de ayrı ayrı dikkat edilmesi gereken hususlar bulunmaktadır. Bu hususlardan bazıları kullanılan cihazlara göre elde edilen verilerin farklı uzantılarda ve farklı yapılarda bulunmasıdır. Elde edilen ham verilerin işlenebilir sayısal veriye dönüştürülmesinde farklı yazılımlara başvurularak verilerin ön işlemeye hazır hale getirilmesi ve yine bazı farklılıklar içeren ön işleme basamaklarının veriye uygulanarak istatistiksel olarak analizlere hazır hale getirilmesi gerekmektedir (Karahalil, 2016). Bu noktada omik çalışmaların gerçekleştirilme amacına bağlı olarak genellikle; elde edilen verilerden edinilen deneyimle yeni bir gözlemin hangi kategoride olacağını tahmin etmeye yönelik olarak sınıflandırma, verilerin belirli gruplar halinde ifade edilip edilmediğini belirlemek amacıyla kümeleme, çalışma grupları arasında farklılık olup olmadığını belirlemek amacıyla karşılaştırma şeklinde üçe ayrılmaktadır. Bahsedilen bu amaçlara yönelik olarak yüzlerce istatistiksel yöntem bulunmaktadır. Bu tez kapsamında çalışmalardaki amaçların büyük çoğunluğunu oluşturan sınıflandırma yaklaşımları üzerinde çalışılacaktır. Bu çalışmalar esnasında literatürde sınıflandırma yapmak amacıyla yapılmış olan omik çalışmaların verileri kullanılacaktır.

Omik çalışmalarının genel hatları (i) verilerin hazırlanan örneklerden çeşitli teknolojiler kullanılarak elde edilmesi (ii) her omik bilimine ve teknolojiye göre farklılık gösteren ham veri formatlarının işlenebilir hale getirilmesi (iii) işlenebilir hale gelen bu verilerin eksik değer ataması, filtreleme, ölçeklendirme gibi ön işleme aşamalarından geçmesi (iv) istatistiksel yaklaşımlarla verilerin analizi (v) analiz sonuçlarına göre elde edilen bilginin yorumlanması ve görselleştirilmesi şeklindedir (Fondi ve Liò, 2015). Her bir aşama ayrı bir uzmanlık gerektirmekte olup birbirine alternatif olabilecek birçok farklı yöntem içermektedir. Bu aşamalardaki uygulamalar çalışmadan çalışmaya farklılık göstermekte, standart bir yöntem kullanımı söz konusu olmamaktadır. Literatürde yapılmış olan bir çalışmada uygulanan yönteme alternatif olabilecek bir başka yöntem uygulansaydı daha iyi sonuçlar elde edilebilir miydi? Her bir aşamada birçok alternatif söz konusuyken kaç farklı yöntemi deneyerek bu araştırmayı ne kadar detaylandırmamız gerekir ya da ne kadar detaylandırabiliriz? Omik çalışmaları için bütün bir iş akışında bu soruları cevaplamak için henüz net bir cevap verememekteyiz.

Fakat gerçekleştirecek olduğumuz bu çalışma kapsamında istatistiksel yaklaşımlarla verilerin analizi aşaması için sınıflandırma analizi boyutunda birden fazla yöntemin denendiği ve ayarlamaların yapıldığı otomatik makine öğrenimi yöntemleri denenecektir. Diğer aşamalara yönelik olarak da yapacak olduğumuz bu çalışma ve devam çalışmaları ile otomatik sistemler geliştirilerek ivme kazandırılabilir. Böylece oldukça zaman alan, deneme yanılma yoluyla araştırılması gereken birçok nokta otomatik olarak araştırılacaktır. Araştırmacı vaktinin çoğunu harcadığı bu deneme yanılma kısımlarından diğer kısımlara kaydırabilecektir verimliliğini artırabilecektir (Feurer, 2015).

Makine öğrenimi yaklaşımlarından biri olarak da bilinen sınıflandırma analizleri için kullanılan birçok algoritma bulunmakta ve yeni algoritmaların geliştirilmesi üzerine çalışmalar devam etmektedir. Bu yöntemlerin diğer birçok alanda olduğu gibi omik verilerde de başarısı kanıtlanmıştır. Her çalışma ve veride bu yöntemlerin performansları ve avantajlı ve dezavantajlı durumları değişmektedir. Bu sebeple elde edilen omik verisinden biyolojik olarak işe yarar bilginin çıkarılması için hangi metodun uygun olabileceğini öngörmek, birden fazla yöntemi denemek ve bu yöntemlerin birbirleriyle eşleştirilmesiyle oluşan kombinasyonlarının kullanımını gerektirmektedir. Kullanılacak olan sınıflandırma algoritmasını uygulayabilmek için bu algoritmanın matematiksel ve istatistiksel yapısını ve mevcut veriye bu yöntemlerin uygulayabilir olup olmadığını bilmek, uygulanabilir ise veriyi uygun bir formata getirmemiz gerekmektedir. Ayrıca her algoritmada optimize edilmesi gereken parametreler ve hiperparametreler en yüksek sınıflandırma performansını verecek şekilde ayarlanmalıdır. Bu aşamalar insan gücüne dayalı olarak bir birini izleyen basamaklar şeklinde gerçekleşmektedir. Çalışma sonucunda hasta olmayan birine hasta ya da hasta olan birine sağlıklı demek gibi yanlış sonuçların önüne geçilmesi için her bir basamağın doğru bir şekilde uygun yöntem ile gerçekleştirilmesi gerekmektedir. Ciddi bir bilgi birikimi ve deneyim gerektiren bu aşamalar bu alanda uzman çalışmacılar için oldukça fazla emek ve zaman gerektirirken bu konunun biraz daha uzağındaki kişiler için içinden çıkılmaz bir hal almaktadır. Bu sorunların üstesinden gelmek için literatürde kullanılmaya yeni başlanan otomatik makine öğrenimi yaklaşımları kullanılarak zaman alıcı, deneyim ve emek gerektiren aşamaların basitleştirilmesi hedeflenmiştir (Feurer, 2015; Thornton ve ark., 2013).

Son yıllarda verimli teknolojilerin yaşam bilimlerinde kullanımının artması omik bilimlerinde yüksek hacimli verilerin, açık kaynaklı biyolojik veri tabanlarının hızla artmasını sağlamıştır. Rahatlıkla erişilebilen büyük boyutlu veriler bu alanda çalışan araştırmacıların tüm gen, protein ve metabolit ağlarını analiz ederek hücrelerdeki moleküler mekanizmaları aydınlatabilme, bütünsel bir bakış açısı kazanabilme umutlarını arttırmıştır. Yüksek verimli deneyler ile ortaya çıkan çok fazla veri çok daha fazla bilgi edinmemizi sağlasa da birçok yeni zorluğu da ortaya çıkarmaktadır. Veri miktarı arttıkça ulaşılacak olan anlamlı bilgi bu veri kalabalığının arasında fark edilemeyebilir. Diğer bazı problemler ise boyutsallık laneti, gürültülü yapı, çoklu test problemi, metodolojik problem, değerlendirme problemi ve bedava yemek olmaz (no-free-lunch) sorunu şeklindedir (Glaab, 2011).

Yüksek hacimli omik verilerin istatistiksel olarak analizlerinde karşılaşılan problemleri aşmak için son yıllarda yeni yöntemler geliştirilmeye çalışılmaktadır. Omik verilerin sınıflandırılabilmesi için birçok makine öğrenmesi yöntemi önerilip uygulanmıştır. Bu yöntemlerden bazıları destek vektör makineleri (DVM), random forest (RF) ve en yakın küçültülmüş merkezler (NSC)'dir.

(Brown ve ark., 2000) mikrodizi sınıflandırması için uyarlanmış destek vektör makineleri algoritmasını kullandı ve maya verileri üzerindeki performansını gösterdi. Bu çalışmada yüksek boyutlu destek vektör makinesi çekirdekleri kullanıldığında ağaç tabanlı sınıflandırıcıdan, Fisher'in doğrusal ayırıcı yönteminden ve doğrusal çekirdekli destek vektör makinesinden daha iyi sonuçlar alınmıştır. (Truong ve ark., 2004) metabolomik verilerinde değişken seçimi için random forest ve destek vektör makinelerini uyguladı ve doğrusal olarak programlanmış destek vektör makinelerinde en iyi sonucu aldı. (Díaz-Uriarte ve Alvarez de Andrés, 2006) gen seçimi ve sınıflandırma amacıyla random forest algoritmasını değerlendirdi. Araştırmacı hem kendi türettiği veri setlerini hem de 9 gerçek veri setini kullandı ve bu yaklaşım sınıflandırma doğruluğunu korurken diğer yaklaşımlardan daha az sayıda gen seçerek bu performansı gösterdi. (Chen ve ark., 2014) en yakın küçültülmüş merkezler yöntemini ve diğer yöntemleri 3 gerçek metabolomik veri setiyle ve türetilmiş veriler ile değerlendirdi. En yakın küçültülmüş merkezler yönteminde diğer yöntemlere göre gürültü daha az oldu ama yinede bilgi içermeyen değişkenler bulundu. Ayrıca bu çalışmada küçültülmüş merkezlerin düzenlenmesiyle ayırma analizi (SCRDA)

yaklaşımında en az önemli değişken sayısını buldu ve bulunanların da literatür ile uyumlu olarak gerçekten önemli olduğu gösterildi.

Standart sınıflandırma yöntemleri kullanılarak omik verilerde iyi bir sınıflandırma performansı elde edilmiş çalışmalar olsa da her bir yöntemin uygulamasında çeşitli zorluklar mevcuttur. Literatürde herhangi bir alanda her veri için her zaman en iyi sonucu veren bir yöntem yoktur, çalışmadan çalışmaya ve veriden veriye yöntemlerin başarısı değişkenlik göstermektedir. Bu nedenle bir veri elde edildiğinde farklı yöntemler ve bu farklı yöntemlerin birlikte kullanımını denemek daha yüksek performanslı sonuçlar elde edilmesini sağlar (Guyon ve ark., 2015).

Literatürde kullanılmaya yeni başlanan otomatik makine öğrenimi yaklaşımları modele alınacak değişkenlerin belirlenmesi, hiperparametre optimizasyonu ve uygun yöntemin belirlenmesi gibi aşamaları otomatize etmeyi başarmıştır.

Otomatik makine öğrenimi yöntemlerinden H2O, çeşitli programlama dilleri ve arayüzler vasıtasıyla erişilebilen bir sunucu kümesinde yürütülmektedir. H2O, Java dilinde geliştirilmiştir. R, Python, Tableau, Javascript ve Flow (web UI) bağlamaları içerir. Makine öğrenmesi modellerinden temel derin öğrenim modellerini, genelleştirilmiş doğrusal modelleri (GLM), yoğun rastgele ormanlar (DRF) ve gradiyent artırma makineleri (GBM) gibi yöntemleri desteklemektedir. Parametre optimizasyonu için kartezyen ızgara arama (CGS) ve rastgele ızgara araması (RGS) yöntemlerini destekler (Balaji ve Allen, 2018).

Bir diğer otomatik makine öğrenimi yaklaşımlarından olan TPOT (Ağaç Tabanlı İş Akışı Optimizasyon Aracı) genetik programlama kullanarak makine öğrenim iş akışını otomatik olarak oluşturan ve optimize eden bir Python aracıdır. TPOT Auto-sklearn gibi, scikit-learn ile birlikte çalıştığı için bir scikit-learn sarımı(wrapper) olarak tanımlanır. Fakat Bayesçi optimizasyonu kullanan Auto-sklearn'in aksine TPOT yaklaşımı genetik programlamaya dayanır. Yaklaşım farklı olsa da, amaç aynıdır: otomatik parametre seçimi, çeşitli algoritmalarla modellemeler yapma, değişken seçimlerinin araştırılması, yinelenen model oluşturma ve model değerlendirmesi yapmaktır. TPOT'un önemli özelliklerinden biride, bir scikit-learn iş akışı şeklinde en iyi performansa sahip model için hazır Python kodunun üretilmesidir. Tüm aday modellerin en iyi performansını temsil eden bu kod, daha sonra sadece bir nihai ürün olarak değil, bir başlangıç noktası olarak çalışabilecek şekilde değiştirilebilir veya

incelenebilir (Olson ve ark., 2016a).

Auto-Keras Pythonda yazılan ve TensorFlow, CNTK veya Theano'nun üzerinden çalıştırılabilen yüksek düzeyli bir sinir ağları uygulama programlama ara yüzüdür. Hızlı deneylerin yapılmasına olanak tanımaya odaklanarak geliştirilmiş bir derin öğrenme kütüphanesidir. (keras.io:keras documentation) Auto-Keras otomatik makine öğrenimi için açık kaynaklı bir yazılım kütüphanesidir. Auto-Keras, derin öğrenme modellerinin mimarisini ve parametrelerini otomatik olarak aramak için fonksiyonlar sağlar (Jin ve ark., 2019a). Bu yöntemler dışında auto_ml, auto-sklearn, auto-WEKA gibi makine öğrenimi iş akışının tamamını ya da belirli kısımlarını otomatik olarak ayarlayan makine öğrenimi yaklaşımları da bulunmaktadır.

Omik verilerde başarılı bir sınıflandırma yapmak için, güçlü istatistiksel algoritmalar gereklidir. Bu algoritmalar, son teknolojilerin yüksek verimli yetenekleriyle başa çıkabilmeli, gen ve gen ürünlerinin en önemlilerinden oluşan alt kümeyi tanımlamalı ve sınıf kategorilerini doğru bir şekilde tahmin edebilmelidir. Omik verilerde sınıflandırma yaklaşımları ile biyobelirteçlerin belirlenmesi kişiselleştirilmiş tıp, ilaç geliştirme, kanser gibi hastalıkların erken teşhisi ve tedavisi gibi önemli alanlara çok büyük katkılar sağlamaktadır (Zararsız, 2015).

Literatürde nispeten yeni ve kullanımı giderek artan otomatik makine öğrenimi yaklaşımları uygun yöntemin belirlenmesi, parametre optimizasyonu ve değişken seçimi gibi zaman alıcı ve deneyim gerektiren aşamaları otomatize etmeyi başarmıştır. Gijsberg ve ark. (2019) yapmış oldukları çalışmada auto-WEKA, auto-sklearn, TPOT ve H2O AutoML yöntemlerin performansını çeşitli veriler kullanarak karşılaştırmıştır. İki sınıflı ve çok sınıflı veriler üzerinde yapılan çalışmada otomatik makine öğrenimi algoritmalarının performansları değerlendirilmiştir (Gijsbers ve ark., 2019).

Otomatik makine öğrenimi yöntemlerinin performansları her ne kadar çeşitli veriler açısından karşılaştırılmış olsa da literatürde bu yeni yaklaşımların omik verilerde performanslarının araştırılmasına rastlanmamıştır. Literatürdeki bu açıktan yola çıkarak bu tez kapsamında otomatik makine öğrenimi algoritmalarını omik verilere uygulamaya karar verildi. Otomatik makine öğrenimi yaklaşımları ile omik verilerde makine öğrenimi algoritmalarından hangisinin kullanılacağı otomatik olarak belirlenecek ve bu algoritmaların parametre optimizasyonu da yine otomatik olarak gerçekleştirilecektir. Ayrıca otomatik makine öğrenimi ile diğer standart yöntemlerin performansları

karşılaştırılacaktır. Bu çalışma kapsamında hedeflenen amaçlar; (i) zaman alıcı analiz süreçlerini hızlandıracak yöntemlerin kullanılabilirliğini denetlemek (ii) deneyim gerektiren aşamaları kolaylaştırmak (iii) verilerden elde edilen bilgi düzeyini artırmak (iv) alanda uzman olmayan kişilerinde analizlerini gerçekleştirebileceği bir yol haritası oluşturmaktır. Özetle bu çalışmanın amacı, omik verilerinde otomatik makine öğrenimi yöntemlerinin performansını değerlendirmek ve mevcut yöntemlere karşı başarısını sınamaktır.

2. GENEL BİLGİLER

Teknolojik gelişmelerin yaşam bilimlerine sağlamış olduğu katkıyla birilikte omik bilimler de hızla gelişmiştir. Bu durum kamuya açık biyolojik veri ve veri tabanlarının katlanarak artmasını sağlamıştır. Omik çalışma alanlarından yüksek hacimli veriler elde edilmeye başlandıkça bu verilerin işlenebilmesi için de yeni yaklaşımlar geliştirilmiştir. Kolayca ulaşılabilen büyük miktardaki veriler araştırmacılara anlamlı bilgiyi elde edebilecekleri yöntemleri kullanarak biyolojik bilgiye ulaşma imkanı sağlamaktadır. Omik verilerden anlamlı bilginin elde edilmesinde birçok istatistiksel yöntem kullanılabilmektedir. Bu yöntemlerin birçoğu makine öğrenimi başlığı altında yer almaktadır. Otomatik makine öğrenimi yaklaşımları ise makine öğreniminde deneme yanılma yoluyla gerçekleştirilen kısımları otomatik hale getirmekte olup oldukça yeni bir çalışma alanıdır. Aşağıda omik bilimlerinden, makine öğreniminden, makine öğrenimi yaklaşımlarından, çeşitli algoritmalardan ve otomatik makine öğrenimi yöntemlerinden bahsedilecektir.

2.1. OMİK BİLİMLERİ

2.1.1. Genomik

Genomik organizmaların tüm genomlarının araştırılmasıdır. İnsan genomu yaklaşık 20000 gen kodlayan 3 milyar DNA baz çiftinden oluşur. Kodlama bölgeleri tüm genomun yaklaşık %1-2'sini oluştururken geri kalan %98-99'luk kodlayıcı olmayan bölgeler yapısal ve fonksiyonel ilişkiye sahiptir. Genom ve genlerle ilgili bilginin açıklanmasına yönelik yapılan çalışmaların tümü olarak özetlenebilecek olan genomik; bir canlıdaki tüm genlerin ayrı ayrı tanımlanmasına, genlerin birbirleri ve çevre ile etkileşimlerinin araştırılmasına ve genlerin zaman-yer-miktar olarak üretim ve aktivasyonlarının incelenmesini amaçlar (International Human Genome Sequencing, 2004).

Proteinler hücrelerin ve dokuların yapısal bileşenleridir ve biyolojik sistemlerin birçok önemli işlevini yerine getirir. Proteinlerin üretimi DNA ile kodlanan, tüm hücreler için ortak olan ve kişinin ömrü boyunca çoğunlukla statik olan genlerle kontrol edilir. Genlerden protein üretimi transkripsiyon ve translasyon olarak bilinen iki ana aşamayı içerir. Transkripsiyon sırasında, tek bir mesajcı ribonükleik asit (mRNA) dizisi geni kodlayan DNA segmentinden kopyalanır. Transkripsiyondan sonra mRNA, proteini oluşturmak için bir amino asit zincirini bir araya getirmek için bir şablon olarak kullanır. Gen ifade araştırmaları, biyolojik bir sistemde kopyalanan mRNA'nın miktarını inceler. Çoğu protein, translasyondan sonra, fonksiyonel hale gelmeden önce modifikasyona tabi tutulsa da hücrenin durumundaki değişikliklerin çoğu mRNA seviyelerindeki değişikliklerle ilgilidir ve transkriptomu sistematik değerlendirme için ölçülmeye değer hale getirir. (Parmigiani ve ark., 2003).

Hızlı ve uygun maliyetli yüksek verimli DNA ve RNA dizilemenin mevcudiyeti ile biyoloji ve tıp bilimlerini veri biliminin araştırma alanlarına dönüştürmüştür. Mikrodiziler ve yeni nesil dizileme (NGS) gibi son genetik teknolojilerin sağladığı büyük miktarda veri bu alanda makine öğrenimi uygulamalarını artırmıştır. Mikrodizi çalışmalarında, araştırmacılar çoğunlukla ilgili genomik özellikleri tanımlamak ve bu özelliklere dayanarak öngörülerde bulunmak için farklı koşullardan örnekler ile çalışılır. Bu koşullara örnek lösemi tipi (AML ve ALL), tümör büyümesi (büyüyen ve sabit), tedavi yanıtı (evet ve hayır), hayatta kalma durumu (sağ ve ölü), patojenik bakteri tipi (brusella ve helikobakter) verilebilir. Gen, transkript, mikro RNA (miRNA) gibi genomik özelliklerin tanımlanması biyolojik belirteç keşfi ya da değişken seçimi vasıtasıyla bahsetmiş olduğumuz durumların sınıflandırma problemleri kullanılarak öngörülmesini sağlar. Genellikle amaç bu koşulları ayırabilmek için en az değişken alt kümesiyle kesin tahminler elde etmektir. Gen ifadesi teknolojileri ve makine öğrenmesi yaklaşımları bu sınıflandırma işlemini daha kesin ve güvenilir kılar (Zararsız, 2015).

Genom araştırmalarında asıl amaç genom dizilimlerinin kullanılarak genomların fonksiyonlarını anlamaktır. Bu bağlamda aşağıda yer alan üç soru araştırılır:

- Her bir genin fonksiyonel rolleri nelerdir ve hangi hücresel süreçlere katılırlar;
- Genler nasıl düzenlenir, genler ve gen ürünleri nasıl etkileşir, bu etkileşim ağları nasıldır;
- Gen ifade seviyesinin çeşitli hücre tiplerinde ve durumlarında nasıl farklılaştığı,

gen ifadesinin çeşitli hastalıklar veya ilaç tedavileri tarafından nasıl değiştiği. Gen transkriptinin çeşitli dokularda, gelişim aşamalarında ve çeşitli koşullar altında bolluğunu bilmek bu soruları cevaplayabilmek için önemlidir. mRNA, bir genin nihai ürünü olmasa da, transkripsiyon, gen regülasyonundaki ilk adımdır ve gen regülasyon ağlarını anlamak için transkript seviyeleri hakkında bilgi gereklidir. Üstelik, şu anda mRNA seviyelerinin ölçümü oldukça ucuzdur ve protein seviyelerinin doğrudan ölçümlerinden daha verimli bir şekilde yapılabilir. mRNA ve hücrede protein bolluğu arasındaki korelasyonu belirlemek basit olmasa da bir hücrede mRNA'nın bulunmaması, ilgili proteinin çok yüksek düzeyde olmadığı anlamına gelebilir ve böylece transkriptom bilgisi kullanılarak proteoma ilişkin en azından kalitatif tahminlerde bulunulabilir. mRNA ve protein seviyesi arasındaki ilişkiyi araştıran çalışmalar bulunmakta ve halen devam etmektedir (Celis ve ark., 2000).

Gen ifadesini transkript düzeyinde izleme yeteneği, DNA mikrodizi teknolojilerinin ortaya çıkmasıyla mümkün olmuştur (Forecast, 1999). DNA mikrodizi 1990'ların başından itibaren gen ifade çalışmalarında yaygın olarak hücre ve dokulardaki gen ifade profillerindeki global değişikliklerin incelenmesinde kullanılan nispeten yeni ve güçlü bir teknolojidir. Bir mikrodizi, üzerine tek iplikçikli DNA moleküllerinin sabit konumlara bağlandığı bir cam slayttır. Bir dizide, her biri tek bir gene ilişkin on binlerce nokta olabilir ve binlerce genin ifade seviyeleri tek bir deneyde eş zamanlı olarak çalışılabilir (Heller, 2002).

DNA mikrodizi teknolojisi, birçok genin aktivasyonunun aynı zamanda izlenebilmesi, hızlı bir yöntem olması, hasta ve sağlıklı hücrelerdeki genlerin aktivitelerinin karşılaştırılmasını sağlaması ve hastalıkları alt gruplar halinde kategorize edebilme gibi avantajlarının yanı sıra, tek bir seferde çok fazla veri analizi yapıldığından, tüm sonuçların analizinin zaman alması, sonuçların yorumlamak için çok kompleks olabilmesi, sonuçların yeterince kantitatif olmaması ve oldukça pahalı bir teknoloji olması gibi bazı dezavantajlara da sahiptir (Bal ve BUDAK, 2012; Brazma ve Vilo, 2000)

2.1.2. Transkriptomik

Transkriptomik protein ve gen ifadesi ile ilgilidir. Protein seviyeleri ve aktivitelerindeki değişiklikleri öngörmek için kullanılır. Transkriptom bir hücrede mRNA dahil olmak üzere transkriptlerin tam kümesi ve miktarıdır (Karahalil, 2016). mRNA'lar genomdaki

belirli genlerle eşleştiği için bir genotip ile ifade edilen fenotipi arasında bağlantı kurmak mümkündür. RNA profili, doku ve hücre tipleri arasındaki fonksiyonel farklılıklar, genler arasındaki etkileşim, ifade edilen diziler, gen regülasyonu ve düzenleyici diziler ile herhangi bir hastalık veya hastalık için aday genlerin tanımlanması için ipuçları sağlar. Transkriptomik tüm transkript türlerini belirlemeyi, transkripsiyonel yapıları değerlendirmeyi, büyüme sırasında ve hastalık gibi farklı durumlar altında değişen ifade seviyelerinin nicelleştirilmesini amaçlamaktadır (Dong ve Chen, 2013; Wang ve ark., 2009)

RNA ifade seviyeleri çok dinamiktir ve hem genetik hem de epigenetik bilgiyi bütünleştirir. Böylece hücrenin işlevsel durumunu yansıtmada oldukça iyidir. Bu gibi birçok avantajlı özellikleri nedeniyle RNA-dizileme giderek gen ifadesinin ölçülmesinde standart bir teknik haline gelmiştir (Zararsız, 2015).

Mikrodizi yönteminde DNA çözeltilerinin sentezlenmesi, saflaştırılması ve depolanması esnasında yoğun işçilik gerektirmesi, genlerin saptamasındaki başarısızlıklar, mikrodizi çalışmaları yapılırken aynı gen ailesinin birbiriyle yakından ilişkili farklı üyelerini temsil eden klonlar arasındaki dizi benzeşmeleri, gürültülü veri üretmesi, farklı deney sonuçları ile karşılaştırılmasındaki problemler ve yeni transkriptleri belirleyememe gibi dezavantajlardan dolayı gen ifadesi profillemede mikrodiziler yerine yeni nesil dizileme yöntemleri tercih edilmeye başlanmıştır (Durmuşçelebi, 2019; Mortazavi ve ark., 2008).

RNA-dizileme, yeni nesil dizileme teknolojilerinin yeteneklerini kullanan mikrodizi teknolojilerindeki kısıtlılıkların üstesinden gelebilme yeteneğine sahip ve yüksek çıktılı dizileme prensibine dayanarak işlemleri daha hızlı ve ucuz yapabilen yeni bir araçtır (Nagalakshmi ve ark., 2008; Wang ve ark., 2009). RNA-dizileme verileri, mRNA transkriptlerinin bolluğuna karşılık gelen kesikli sayma verilerinden oluşmaktadır. RNA dizileme teknolojisi, aynı anda binlerce genin ifade düzeylerini sayısallaştırabilmektedir.

RNA-dizileme verileri bir referans genom veya transkriptomla hizalanan dizileme okumaları sayısı olan kesikli sayma değerlerini içermektedir. Mikrodizi verileri uygun dönüşümlerle normal dağılıma yaklaştırıldıktan sonra normal dağılımı kullanan algoritmalar kullanılır. RNAdizileme verileri, genellikle küçük değerlere sahip olan sayma değerlerinden oluşmakta ve belirli bir ortalama-varyans ilişkisi olduğu için normal dağılıma uyan yaklaşımların doğrudan uygulanması uygun olmamaktadır

(Witten, 2012) Bu nedenle, RNA-dizileme verilerinin analizi için ya veriyi normal dağılıma yaklaştıran dönüşümler uygulamak ya da Poisson, negatif Binom gibi kesikli olasılık dağılımlarına dayanan yöntemler kullanarak doğrudan sayma verileri ile analiz etmek gerekir (Anders ve Huber, 2010; Zheng ve ark., 2014; Durmuşçelebi, 2019)

2.1.3. Proteomik

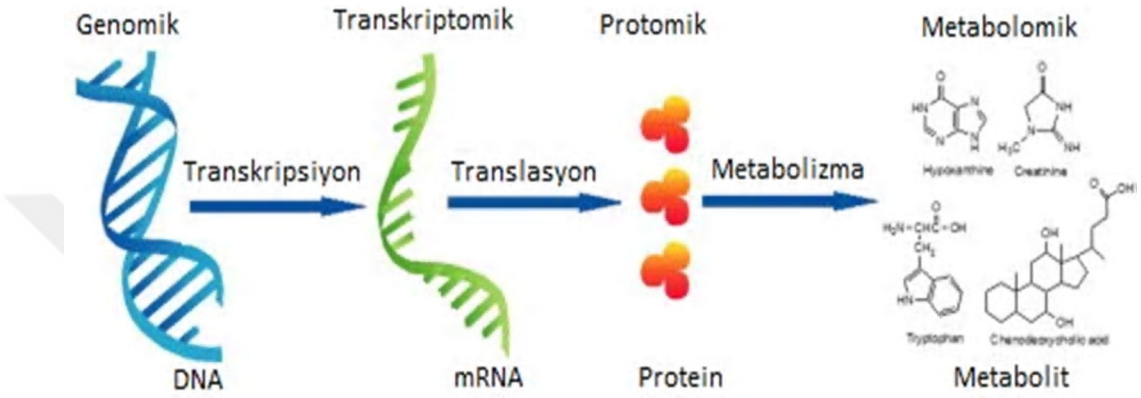
Proteom; belli bir zaman ve mekanda bir organizmanın sahip olduğu ve ifade ettiği tüm proteinlerin toplamıdır. Proteinler sadece genler tarafından kodlanan polipeptid yapıları değil, aynı zamanda sentez sonrası modifikasyonları da içermektedir. Mekan terimi farklı proteinlerin farklı hücre kompartmanlarında ve farklı hücre tiplerinde ifadesini belirtir. Zaman ise farklı gelişim evreleri, çevresel koşullar, çeşitli hastalıklar, yaşlılık gibi süreçlere işaret eder (Marko-Varga, 2004)

Proteomik; belli bir zamanda belli bir yerde bulunan tüm proteinlerin yapılarını, yerleşimlerini, miktarlarını, translasyon sonrası modifikasyonlarını, doku ve hücrelerdeki işlevlerini, diğer proteinler ve makro moleküller ile olan etkileşimini belirlemeyi amaçlamaktadır. Proteomik, dinamik bir terim olup farklı koşullarda hücre, doku veya vücut sıvılarındaki proteinlerin yüksek verimli teknolojiler ile kantitatif analizi olarak tanımlanır (Başaran ve ark., 2010; Marko-Varga, 2004)

2.1.4. Metabolomik

Metabolomik küçük molekül -omiksi olarak tanımlanabilir. Daha kapsamlı bir şekilde ifade edecek olursak metabolomik bir hücrede, dokuda veya biyolojik sıvılarda bulunan tüm metabolitlerin NMR (nükleer manyetik rezonans), GC-MS (gaz kromatografi-kütle spektrofotometri) ve LC-MS (sıvı kromatografi-kütle spektrofotometri) gibi yüksek verimli teknolojilerle kısa sürede, doğru bir şekilde ayrılması, tanımlanması ve ölçülmesidir. Bir canlının genetik özellikleri ve beslenme şekline uygun fenotipik yanıtını karakterize etmek için o canlının metabolik profilinin kantitatif ölçümüdür. Metabolom ise biyolojik örnekte bulunan hormon, sinyal molekülü, metabolik ara ürün ve ikincil metabolit gibi saniyeler içinde değişebilen dinamik bir yapıya sahip küçük moleküllerin (<1500 Da) tümüdür. Metabolomik çalışmaları mutant hücrelerinin genetik ile ilgili kısımlarında, toksikolojide, ilaç keşfinde, beslenme, kanser, diyabet, doğal ürünlerin keşfedilmesi, biyobelirteç tespiti, enzim substrat ilişkisi değerlendirmesi, metabolik yolak analizleri ve bu gibi daha birçok alanda önemli bir rol oynar (Patti ve ark., 2012; Schmidt, 2004; Viant, 2009).

Omik bilimleri ayrı ayrı ele almış olsak da aslında birbirleriyle oldukça ilişkilidirler. Şekil 2.10'da da görüldüğü gibi genlerin araştırılmasıyla ilgili genomik biliminden sonra, DNA'nın transkripsiyonu sonucunda oluşan transkriptleri inceleyen transkriptomik, bu transkriptler aracılığıyla oluşan proteinleri inceleyen proteomik bilimi gelişmiştir. Daha sonra da vücutta bulunan son ürünleri, metabolitleri inceleyen metabolomik bilimi gelişmiştir.



Şekil 2.1 Genomik transkriptomik proteomik metabolomik (Zhao ve Lin, 2014)

2.1.5. Epigenomik

Epigenom kelimesi, epigenetik kelimesinden türetilmiştir ve Waddington (1942) tarafından tanıtılmıştır. Waddington epigenomu "fenotipi oluşturan genler ve gen ürünleri arasındaki nedensel etkileşimleri araştıran biyoloji dalı" olarak tanımladı ve o zamandan beri, DNA dizilerinde herhangi bir değişiklik içermeyen spesifik genlerin düzenleyici ve bilgi depolama mekanizmalarını tanımlamak için kullanıldı. İnsan genomunun dizilenmesinden bu yana geliştirilen yeni genomik araçlar, kromatin durumları ve DNA modifikasyonlarının genom çapında değerlendirilmesine izin verdi ve epigenomik bellek gibi beklenmedik yeni epigenetik prensiplerin keşfedilmesine yol açtı (Carlberg ve Molnár, 2018; Dirks ve ark., 2016).

Epigenomik araştırmasının amacı, genetik regülasyonu anlamak ve bunun hücresel büyüme ve farklılaşmaya etkisi ile yaşlanma ve hastalık gibi süreçlerdeki rolünü aydınlatmak olarak özetlenebilir (Greally ve Jacobs, 2013; Novik ve ark., 2002).

2.1.6. Metagenomik

Metagenom terimi ilk olarak 1998 yılında J. Handelsman (2004) tarafından çevresel bir örnekteki genetik materyalin toplamını tanımlamak için kullanılmıştır. Kültürsüz organizmaların fizyolojik ve genetik bilgisine erişim sağlamak için tasarlanan yöntemler arasında, bir mikroorganizma popülasyonunun genomik analizi olarak metagenomik güçlü bir yaklaşım olarak ortaya çıkmıştır. Metagenomik kelimesi, analizlerin bir analizi olan bir meta-analizde olduğu gibi benzer fakat aynı olmayan çalışmalardan oluşan bir koleksiyonun analizini elde etmek için koordine edildi (Glass, 1976; Handelsman ve ark., 1998). Topluluk genomikleri, çevresel genomikler ve popülasyon genomikleri aynı yaklaşımın eş anlamlılarıdır. Yeni biyolojik motiflerin keşfi, kısmen metagenomik klonların fonksiyonel analizine bağlıdır (Handelsman, 2004).

Büyük boyutlu omik verileri ilişkili yapıda, gürültülü ve yüksek boyutludur. Bu veriler ile çalışmak diğer verilerden farklı olarak bazı zorluklar içermektedir. Bu zorlukların kaynağı omik çalışmalardan elde edilen verilerin genel özelliklerinden kaynaklanmaktadır (Debnath ve ark., 2010; Glaab, 2011). Aşağıda bu verilerde yaşanan problemlerden bazıları özetlenmiştir.

Boyutsallık sorunu (curse of dimensionality): Omik verileri yüksek boyutludur ve değişken sayısı (p) gözlem sayısından (n) oldukça fazladır. Bu tür verilerde geleneksel istatistiksel yöntemler kullanılamamaktadır. Bu gibi durumlarda nasıl güçlü kümeleme veya denetimli tahmin sonuçları elde edilebilir (Hastie ve ark., 2009)?

Gürültü sorunu: Genlerin, proteinlerin, metabolitlerin büyük bir kısmının ifade değerleri gürültüyle (hem teknik hem de biyolojik gürültü kaynakları ile) maskelenir veya baskın hale getirilirse, farklı biyolojik koşullarda önemli ölçüde farklı şekilde düzenlenmiş genler, proteinler veya metabolitler güvenilir şekilde nasıl tanımlanabilir?

Çoklu test problemi: sıfır hipotezinin yanlış reddedilmeleri, giriş verilerinin yüksek boyutluluğu azaltılarak veya tekrarlanan hipotez testlerini hesaba katıp hipotez testlerini düzelterek nasıl önlenebilir?

Değerlendirme problemi: Makine öğrenimi modellerinin genelleme hatalarını tahmin etme yöntemlerinin, küçük örneklem büyüklüğüne sahip omik veri kümelerindeki varyansı veya yanlılığı doğru olarak tahmin etmede sınırlamalara sahip olma eğiliminde olduğu göz önüne alındığında, doğrulama için güvenilir bir iş akışı nasıl yapılmalıdır

(Braga-Neto ve Dougherty, 2004).

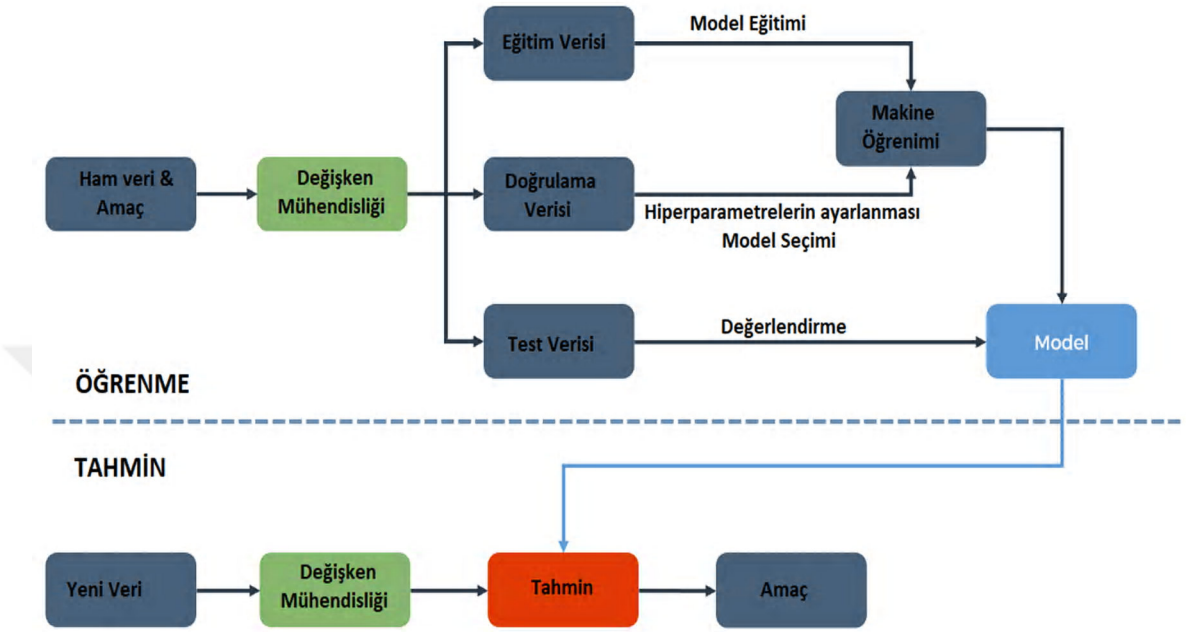
Metodolojik problem ve no-free-lunch (bedava yemek olmaz) sorunu: Belirli bir omik çalışmasından elde edilen veri analizi için herhangi bir "seçim yöntemi" mevcut mu, yoksa belirli bir analiz problemini verimli ve etkili bir şekilde çözmek için hangi algoritmalar karşılaştırılmalı veya birleştirilmelidir? Eğer çeşitli yöntemlerin farklı veri kümeleri üzerinde farklı güçlü ve zayıf yönleri olduğu gösterilmişse (örneğin farklı kanser türleri için veri setleri), güçlü bir analiz sistemi nasıl oluşturulabilir ve birçok farklı veri kümesinde yüksek bir performans nasıl elde edilebilir?

Bu gibi zorluklar omik veriler ile çalışmayı zorlaştırmaktadır. Bu gibi problemlerin üstesinden gelebilmek için çalışmalar halen devam etmektedir.

2.2. MAKİNE ÖĞRENİMİ

Geçtiğimiz yıllarda makine öğrenimi bilişim teknolojilerinin temeli haline gelmiş olup fark edilmese bile yaşamımızın bir parçası haline gelmiştir. Günümüzde veri miktarındaki artış hızı düşünüldüğünde etkili veri analizinin teknolojik ilerleme süreci için gerekli bir bileşen olarak daha da yaygınlaşacağı düşünülmektedir. Öğrenme deneme yanılma yoluyla edinilen tecrübelerden bilgi, beceri ve yetenek kazanmaktır (www.tdk.gov.tr, Erişim tarihi: 10.11.2019). Makine öğrenimi de elde edilen veriden anlamlı kalıpların otomatik olarak elde edilmesi şeklinde ifade edilebilir (Shalev-Shwartz ve Ben-David, 2014). Son birkaç on yıldır büyük veri setlerinden gerekli olan bilginin elde edilmesi için başvurulan en yaygın araçlardan biri haline gelmiştir. Günlük hayatımızda da makine öğrenimi temelli bir teknoloji ile iç içe bulunmaktayız. Arama motorlarının en iyi sonuçları getirmesi, e-postaların spam ve spam değil şeklinde filtrelenmesi, kredi kartı dolandırıcılıklarının tespit edilmesi, dijital kameraların yüz tanıması, akıllı telefonların sesli komutları tanıması, otomobillerin kaza önleme sistemleri bu kullanımlara örnek olarak verilebilir. Ayrıca makine öğrenmesi astronomi, tıp ve biyoinformatik gibi bilimsel uygulamalarda da sıklıkla kullanılmaktadır. İnsan becerilerinin çoğu zeki varlıklardan örnek alıp deneyimleyerek öğrenilir veya geliştirilir. Makine öğrenimi de öğrenme ve uyarlama yeteneğine sahip programlama ile ilgilidir (Shalev-Shwartz ve Ben-David, 2014). Makine öğrenimi için sık kullanılan tanımlamalardan biri bilgisayarlar ile açıkça programlama yapmadan öğrenme imkanı sağlayan çalışma alanı şeklindedir. Makine öğreniminin temel dayanağı, istatistiksel analiz içeren algoritmalar oluşturarak veri alma, bu verileri işleyerek anlamlı çıktı

oluşturma, yeni veriler ortaya çıktıkça güncelleme yaparak sonucu tahmin etme şeklindedir (Samuel, 1969).



Şekil 2.2. Makine öğrenmesi süreci

Şekil 2.1'de genel makine öğrenme süreci gösterilmektedir. Belli bir amaca yönelik olarak veri elde edildikten sonra model oluşturulacak değişkenlerin elde edilmesi (değişken seçimi ve değişken oluşturulması) gerekmekte olup bu aşamada birçok farklı yaklaşım ve yöntemler bulunmaktadır. Şekil 2.1'de değişken mühendisliği olarak gösterilmekte olan bu aşamada en az değişken ile en anlamlı modelin oluşturulması hedeflenmektedir. Kullanılacak değişkenler belirlendikten sonra belirli oranlara göre veri seti gözlem bazında eğitim verisi, doğrulama verisi ve test verisi olarak ayrılmalıdır. Eğitim verisi kullanılarak oluşturulan modelin hiperparametre ayarlaması için doğrulama verisi kullanılmakta olup bu aşamadan sonra elde edilen modelin genel doğruluğu test verisi ile değerlendirilmektedir. Değerlendirme sonucunda modelin performansı belirlenip gelecek verilerde tahminde bulunmak için yeterli bir model olup olmadığına karar verilir.

Makine öğrenimi yaklaşımı bir çok avantajlı özelliklere sahiptir. Uzun kurallar listesi ve manuel olarak ayarlamaların gerektiği problemlerde makine öğrenimi algoritması bu işi

basitleştirip daha iyi performans gösterebilir. Geleneksel yaklaşımlarla iyi bir çözümü bulunamayan karmaşık problemler için güçlü bir makine öğrenimi yaklaşımı ile bir çözüm bulunabilir. Sabit bir yapısı olmayan, farklı dağılımlar gösteren durumlarda makine öğrenme algoritması yeni verilere göre adapte edilebilir. Karmaşık problemler ve büyük veriler içeren durumlarda veri hakkında fikir edinebilmeyi sağlar. Bununla beraber geleneksel yöntemlere göre işleri basitleştirdiği için daha hızlı çözüme ulaşabilmekte ve çok farklı alanlardan problemlere adapte edilebilmektedir.

Makine öğrenimine geniş bir açıdan bakacak olursak birçok farklı makine öğrenimi sistemi bulunmaktadır (Géron, 2019). Bunlar;

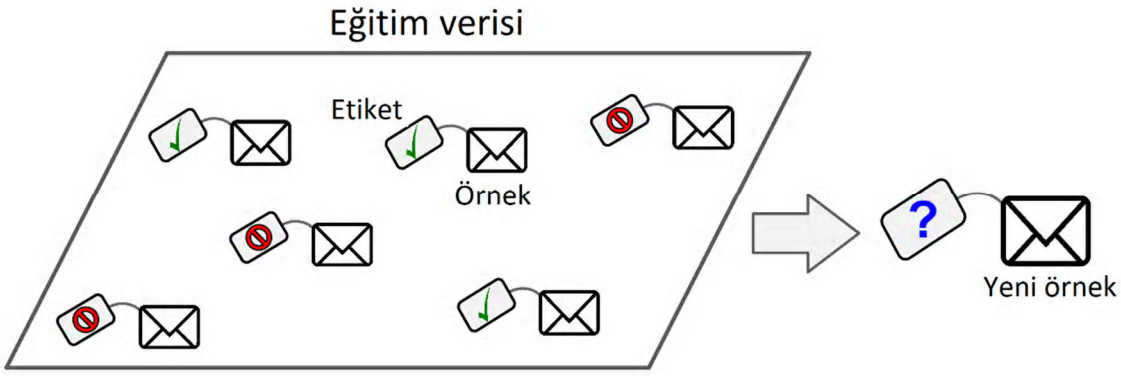
- İnsan denetimi ile verinin eğitilip eğitilemeyeceğine göre denetimli, denetimsiz, yarı denetimli ve pekiştirmeli öğrenme
- Toplu veya çevrim içi (anında ya da kademeli) öğrenme
- Bilinen veri değerleri ile yeni veri değerlerini karşılaştırma ya da bir çok araştırmacının yaptığı gibi eğitim verisinde bulunan kalıpları belirleyerek tahmin modeli oluşturma (örnek tabanlı ve model tabanlı öğrenme)

Bu sistemler tek başına kullanılabileceği gibi birleştirilerek de kullanılabilir. Örneğin güncel bir spam filtreleme algoritması spam olan ve olmayan örnekleri kullanarak eğitilmiş derin bir sinir ağı modeli kullanarak çevrim içi öğrenebilir; böylece bu algoritma çevrim içi, model tabanlı ve denetimli bir öğrenme sistemi haline gelir.

Makine öğrenimi genelde veri ile model eğitimi sırasında sınıf etiketlerinin olup olmamasına ve türüne göre sınıflandırılır. Denetimli, denetimsiz, yarı denetimli ve pekiştirmeli öğrenme şeklinde dört ana kategoriden oluşmaktadır (Géron, 2019).

2.2.1. Denetimli öğrenme

Denetimli öğrenmede algoritmaya işlemesi için verilen veri seti etiket olarak isimlendirilen sınıf bilgisini içermektedir.



Şekil 2.3. Denetimli öğrenme için etiketli eğitim veri seti (spam sınıflandırma)

Bu çalışma kapsamında da üzerinde durulacak olan tipik bir denetimli öğrenme işi sınıflandırmadır. Spam filtreleme işi de sınıflandırmaya uygun bir örnektir. Biyoinformatik alanından bir örnek verecek olursak hasta ve sağlıklı olduğu bilinen bireylerden alınan örneklerden yola çıkarak gruplar arasında gen, protein ve metabolit düzeyine göre modeli eğitip yeni bir gözlemin bu düzeylerine göre hasta ya da sağlıklı olduğu tahmin edilmeye çalışılabilir.

Denetimli öğrenme sadece sınıf değişkeninin kategorik olduğu durumlarda değil nümerik bir değişkenin olduğu durumlarda da çalışabilir. Tahmin edilmeye çalışılan değişkenin nümerik olduğu durumda regresyon olarak isimlendirilen öğrenme işi gerçekleştirilmeye çalışılır. Bir arabanın özeliklerine göre fiyatını tahmin etmeye çalışırsak bir regresyon modeli kullanmamız gerekecektir. Biyoinformatik alanından bir örnek verecek olursak bireylerin farklı metabolit ve gen ifadesi düzeylerine bakarak kandaki glikoz düzeyini tahmin etmeye dayalı bir çalışmayı örnek verebiliriz. Regresyon analizi lojistik resresyon yönteminde olduğu gibi sınıflandırmaya yönelik, yani sınıf etiketini belirlemeye yönelik olarakda gerçekleştirilebilmektedir. Bu yöntemde gözlemlerin hangi sınıfa ait olabileceğine dair olasılıkları verilerek tahminlemeye gidilir (Géron, 2019).

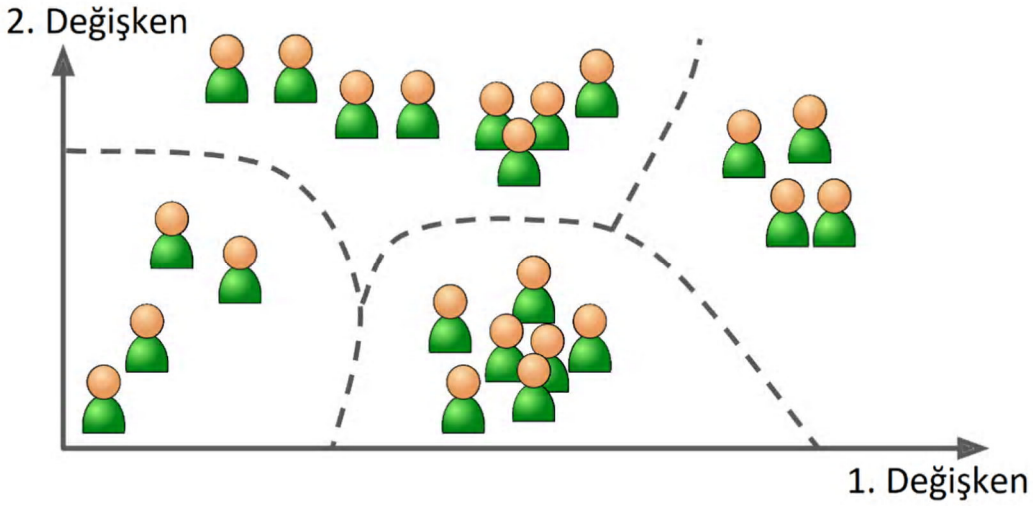
Sıklıkla kullanılan oldukça önemli denetimli öğrenme algoritmalarına örnek verecek olursak;

- K-en yakın komşu algoritması (KNN)
- Doğrusal regresyon
- Lojistik regresyon

- Destek vektör makineleri (DVM)
- Karar ağaçlarına dayalı algoritmalar
- Yapay sinir ağları yöntemlerini sayabiliriz.

2.2.2. Denetimsiz öğrenme

Denetimsiz öğrenmede veri setini tanımlayan bir etiket bulunmamaktadır. Bu sistemde Şekil 2.3'de de görüldüğü gibi eğitici olamadan öğrenilmeye çalışılmaktadır.



Şekil 2.4. Denetimsiz öğrenme için etiketsiz eğitim verisi (kümeleme analizi)

Denetimsiz öğrenme yaklaşımı kümeleme, anomali ve yenilik tespiti, görselleştirme ve boyut indirgeme, birliktelik kuralı öğrenme gibi farklı amaçlar için kullanılabilmektedir.

Sıklıkla kullanılan oldukça önemli denetimsiz öğrenme algoritmalarına örnek verecek olursak;

- Kümeleme amacıyla kullanılan
    - k-ortalamalar
    - DBSCAN
    - Hiyerarşik kümeleme analizi
- Anomali ve yenilik tespiti için kullanılan
    - Tek sınıflı DVM
    - İzolasyon ormanı
- Görselleştirme ve boyut indirgeme için kullanılan

    - Temel bileşenler analizi (PCA)
    - Çekirdek PCA
    - Yerel-doğrusal gömme (LLE)
    - T-dağıtık Stokastik Komşu Gömme (t-SNE)
- Birliktelik kuralı belirlemeye yönelik olarak kullanılan
    - Apriori (olası)
    - Eclat yöntemlerini sayabiliriz.

2.2.3. Yarı denetimli öğrenme

Yarı denetimli öğrenmede verilerin bir kısmının sınıf etiketi bulunurken bir kısmının ise hangi sınıfa ait olduğuna dair bilgi verilmemektedir. Yarı denetimli algoritmaların çoğu denetimli ve denetimsiz öğrenme algoritmalarının kombinasyonudur. Örneğin derin inanç ağları (DBN) sınırlandırılmış Boltzmann makineleri (RBM) olarak isimlendirilen denetimsiz öğrenme bileşenlerinin birleştirilmesine dayanır. Daha sonra birleştirilen tüm sistem denetimli öğrenme yaklaşımı ile değerlendirilir.

2.2.4. Pekiştirmeli öğrenme

Pekiştirmeli öğrenme yaklaşımında sistem biraz daha farklı çalışmakta olup tahminin doğruya yaklaşması için ödüllendirme ya da cezalandırma yapıp iteratif olarak çalışmaktadır. Nihai olarak sistem en iyi ödülü almak için en iyi stratejiyi kendi kendine öğrenmelidir (Géron, 2019).

2.3. SINIFLANDIRMA ANALİZLERİ

Makine öğreniminde denetimli öğrenme yaklaşımlarından biri olarak bahsetmiş olduğumuz sınıflandırma analizi kategorik sınıf değişkeninin tahmin edilmesine yönelik bir süreci ifade etmektedir. Bir gözlemin hangi sınıfa ait olduğunu tahmin etmek o gözlemi bir sınıfa ya da kategoriye atamayı ifade ettiği için sınıflandırma olarak adlandırılmaktadır. Her ne kadar yapılan bu işlem sınıflandırma olsa da birçok yöntem her bir gözlemin hangi sınıfa ait olabileceğine dair olasılıkları ön görür ve bu bağlamda regresyon metodu gibi davranır. Sınıflandırma problemleri pratikte regresyon problemlerinden daha sık meydana gelmektedir. Sınıflandırma probleminin daha iyi anlaşılabilmesi için çeşitli örnekler;

Çeşitli semptomlar ile acil servise gelen hastaların 3 tıbbi durumdan hangisine sahip olduğunu doğru şekilde kestirmeye çalışmak,

Çevrim içi bankacılık hizmetinde gerçekleştirilen bir işlemin dolandırıcılık amaçlı olup olmadığını işlem geçmişi gibi özelliklerden yola çıkarak kestirmeye çalışmak,

Belirli bir hastalığa sahip ve sağlıklı olan birçok kişinin deoksiribonükleik asit (DNA) dizileme verisine bakarak hangi DNA mutasyonlarının hastalığa neden olduğunu belirlemeye çalışmak şeklindedir.

Birçok sınıflandırma tekniği bulunmakta olup halen geliştirilmekte olan yöntemlerde bulunmaktadır. Çok geniş kullanım alanlarına ve güçlü özelliklere sahip olan bazı sınıflandırma yöntemlerine lojistik regresyon, ağaç tabanlı sınıflandırıcılar, destek vektör makineleri, en yakın küçültülmüş merkezler (NSC), K en yakın komşular (KNN), doğrusal ayırma analizi örnek verilebilir (James ve ark., 2014). Bu tez kapsamında uygulanacak olan sınıflandırma yöntemleri random forest (RF), destek vektör makineleri (DVM) ve en yakın küçültülmüş merkezler (NSC) yöntemleridir. Makine öğrenimi nispeten yeni bir çalışma alanı olsa da çok sayıda sınıflandırma yöntemi mevcuttur. Bu nedenle bütün sınıflandırma yöntemlerinden bahsedilemeyecek olup literatürde sıklıkla kullanılmakta olan bazı geleneksel yöntemlerden bahsedilecektir.

2.3.1. Geleneksel Sınıflandırma Teknikleri

2.3.1.1. K En Yakın Komşu Algoritması (KNN)

K en yakın komşu sınıflandırma yöntemi kolay anlaşılır bir yöntem olup özellikle küçük örneklem hacmine sahip verilerde iyi çalışmaktadır. Benzer gözlemlerin birbirlerine yakın mesafede olduğu varsayılarak geliştirilmiş olan bu yöntemde “K” bakılacak olan en yakın komşu sayısı olup modelin veriyi iyi sınıflandırabilmesi için optimize edilmesi gereken bir değerdir. Bu yöntem de sınıflandırılmamış bir örnek sınıf etiketi belli olan gruplardan hangisine en yakın ise o sınıf grubuna atanır. K en yakın komşu algoritması uygulama basamakları:

- Başlangıç K parametresi belirlenir.
- Hangi sınıfa ait olduğu belirlenecek olan yeni gözlemin mevcut her bir gözleme olan uzaklığı çeşitli uzaklık-benzerlik fonksiyonları kullanılarak hesaplanır.
- En yakın K tane komşu alınır.
- K tane komşudan en çok tekrar eden sınıfa ilgili gözlem atanır.

Yakın olan grubu belirlemek için Euclidean, Manhatttan, Minkowski gibi uzaklık-

benzerlik fonksiyonlarından faydalanılmaktadır. K en yakın komşu algoritması parametrik olmayan sınıflandırma yöntemi olarak uzun yıllardır kullanılmaktadır. Basitliği göz önüne alındığında düşük sınıflandırma hatasına sahiptir fakat hesaplama açısından masraflı ve eğitim verileri için büyük bir hafıza gerektirmektedir (Cover ve Hart, 1967; Rätsch, 2004).

2.3.1.2. Naive Bayes Sınıflandırıcı

Naive Bayes sınıflandırma yönteminin temelini Bayes teoremi oluşturmaktadır. Bu yöntemde gözlemlerin hangi sınıfa ait olabileceklerine dair olasılıklar hesaplanır ve gözlem en yüksek olasılığa sahip olan sınıfa atanır. Dengesiz verilerin sınıflandırılmasında da kullanılabilen bir yöntemdir. Bu yöntemde her bir değişkenin hangi sınıf kategorisine ait olacağı birbirinden bağımsızdır. Bunun anlamı bir değişkenin olasılığı diğer bir değişkenin olasılığını etkilemediğidir (Mukherjee ve Sharma, 2012).

Bayes Teoremi:

$$P(S_i|X) = \frac{P(S_i)P(X|S_i)}{P(X)} \tag{2.1}$$

$X = (x_1, x_2, \ldots, x_n)$ örnek kümesi

$S = (S_1, S_2, \ldots, S_m)$ sınıf kümesi

$P(S_i|X)$ = X olan bir örneğin i sınıfından olma olasılığı

$P(S_i)$ = i sınıfının ilk olasılığı

P(X) = herhangi bir örneğin X olma olasılığı

$P(X|S_i)$ = i sınıfından bir örneğin X olma olasılığı

Naive Bayes yönteminde amaç $P(S_i|X)$'i maksimize etmektir.

2.3.1.3. Doğrusal Ayırma Analizi (LDA)

Doğrusal ayırma analizi girdi uzayında sınıf içi varyansı en az sınıflar arası uzaklığı en yüksek olan bir hiper düzlem hesaplar (Fisher, 1936). Doğrusallığın sağlandığı durumlarda büyük veri setleri de dahil olmak üzere genelikle oldukça iyi performans gösterir. Bu yöntemin dezavantajı her zaman doğrusal bir ayrım ile iyi bir sınıflamanın yapılamayacağıdır. Çeşitli çekirdek fonksiyonları kullanılarak doğrusal olmayan uzantıları geliştirilmiştir. Sıklıkla kullanılan doğrusal ayırma analizi uzantıları karesel

ayırma analizi (QDA), esnek ayırma analizi (FDA), düzenlenmiş ayırma analizi (RDA)'dir.

Doğrusal ayırma analizinin iki temel varsayımı bulunmaktadır;

- verilerin normal dağılım olarak bilinen Gauss tipi dağılıma uyması uyması
- verideki her bir değişkenin ortalama etrafında değişiminin benzer olup varyansların homojenliği varsayımını sağlaması

Doğrusal ayırma analizi yeni gözlemlerin hangi sınıfa ait olduğuna dair olasılıkları hesaplar ve en yüksek olasılığa sahip olan gruba ilgili gözlem atanır. Olasılık hesaplamada Bayes teoremini kullanmaktadır (denklem 2.1) (Tharwat ve ark., 2017).

2.3.1.4. Lojistik Regresyon

Gözlemleri niteleyen çeşitli bağımsız değişkenler kullanılarak kategorik sınıf değişkeni arasındaki ilişki belirlenmek isteniyorsa lojistik regresyon analizine başvurulabilir. Sınıf değişkeni iki kategorili ise ikili lojistik regresyon, ikiden fazla sıralı kategori varsa sıralı lojistik regresyon ve ikiden fazla sırasız kategori varsa isimsel lojistik regresyon olarak isimlendirilir (Karabulut, 2017).

İkili lojistik regresyonda negatif sınıf 0 pozitif sınıf 1 etiketi ile kodlanır. Gözlemlerin hangi sınıfa ait olduğunu belirlemeye yönelik olarak olasılık değeri hesaplanır. Genellikle bu olasılık değeri için kesim noktası 0.5 alınır. Elde edilen değer 0.5'den küçükse gözlem 0 kategorisine, 0.5'den büyükse 1 kategorisine atanır. Lojistik regresyon yönteminin temelini lojistik fonksiyon olarak da isimlendirilen sigmoid fonksiyon oluşturur.

Sigmoid fonksiyon;

$$f(x) = \frac{1}{1+e^{-x}} \tag{2.2}$$

Bu fonksiyonda x değeri oluşturmuş olduğumuz Y= β_0 + β_1X gibi bir modele karşılık gelmektedir. Lojistik eşitliği ise şu şekildedir ;

$$y = \frac{e^{-(B0 + B1X)}}{1+e^{-(B0 + B1X)}} \tag{2.3}$$

Bu denklemde x tek bir bağımsız değişkene, β_0 oluşturulan model sabitine, β_1 x bağımsız değişkeninin katsayısına karşılık gelmektedir. y ise tahmin etmeye çalıştığımız sınıf değişkenine karşılık gelmektedir ve 0 ile 1 arasında değerler alır (Karabulut, 2017).

2.3.1.5. Karar Ağaçları

Karar ağaçları hem regresyon hem de sınıflandırma problemleri için kullanılabilmektedir. Karar ağaçları ile ilgili araştırmalar uygulamalı istatistik alanında altmışlı yılların başlarında başlamış olup ilk olarak AID (Automatic Interaction Detection) geliştirilmiştir (Morgan ve Sonquist, 1963). Daha sonra CART, C4.5, C5.0, ID3, CHAID ve QUEST gibi birçok karar ağacı algoritması geliştirilmiştir. Genellikle sınıflandırma problemleri için kullanılan karar ağacı yöntemlerinde kökten yapraklara doğru ilerleyen yönlü bir ağaç yapısı kullanılmaktadır. Gürültülü, eksik ve aşırı değerlere sahip verilerle başa çıkma konusunda oldukça iyi bir yöntemdir. Sonuçların anlaşılması basit ve yorumlaması da kolaydır. Hem kategorik hem sayısal verilere uygulanabilen yöntemin maliyeti düşük ve uygulaması kolaydır. Parametrik olmayan bir sınıflandırma yöntemi olan karar ağaçları çoklu bağlantı, heterojenlik ve dağılım ile ilgili hatalardan çok etkilenmez. Bahsetmiş olduğumuz bu avantajlarına rağmen bazı dezavantajları da bulunmaktadır. Bu dezavantajlardan biri bu yöntemin olasılıksal bir modele dayanmıyor olmasıdır. Yani elde edilen sonuçların güven aralığı ve olasılık düzeyi yoktur. Sonuçların güvenilirliği eğitim verisinin doğruluğuna bağlıdır. Ağaç dallanmasının takip edilemeyeceği düzeyde karmaşık ağaçlar üretebilir ve veriyi ezberleme durumu yaşanabilir. Veri ezberleme birçok sınıflandırma yönteminde yaşanan bir problemdir. Bu durum eğitim verisinde çok iyi performans gösteren bir model elde edip, oluşturulan bu model ile yeni gözlemlerin hangi sınıfa ait olduğunu belirleme konusunda başarısız olunması şeklinde açıklanabilir. Karar ağaçlarında bu problem ile başa çıkabilmek için budama ve çeşitli model parametrelerinde kısıtlamalara gidilir. Budama olarak bahsedilen terim ise az sayıda gözlemin bulunduğu dal ve yaprakların karar ağacından çıkarılmasını ifade etmektedir (Kuzey, 2012; Shalev-Shwartz ve Ben-David, 2014).

Karar ağacı oluştururken hangi değişkenlerin hangi sırada kullanılacağının belirlenmesi gerekmektedir. Bu amaçla kullanılmakta olan Entropi ölçüsü bulunmaktadır. Entropi değeri arttıkça kullanılan değişken ile yapılan sınıflandırma sonucu o oranda belirsizdir. Bu nedenle karar ağacı oluştururken ağacın kökünden yapraklarına doğru yapılan sıralama entropi değeri en düşük olandan en yüksek olana doğru olmalıdır (Çalış ve ark., 2014). Entropi değerini hesaplayan formül ise aşağıdaki gibidir (Shannon, 1948):

$$\text{Entropi}(X_1) = -\sum_{i=1}^{m} p_i \log_2 p_i \qquad (2.4)$$

m: entropisi hesaplanacak durum sayısı

p_i: i durumunun olasılığı

X: İlgilendiğimiz olay (bir paranın havaya atılması gibi)

Karar ağacını oluştururken entropiye göre sıralanan değişkenleri dallara ayırırken de bölünme sonucunda en düşük entropiyi veren ayırma seçilmelidir. En iyi ayırmayı yapabilmek için bilgi kazancından faydalanılmaktadır. Gini indeksi olarak da bilinen bilgi kazancı şu şekilde hesaplanmaktadır:

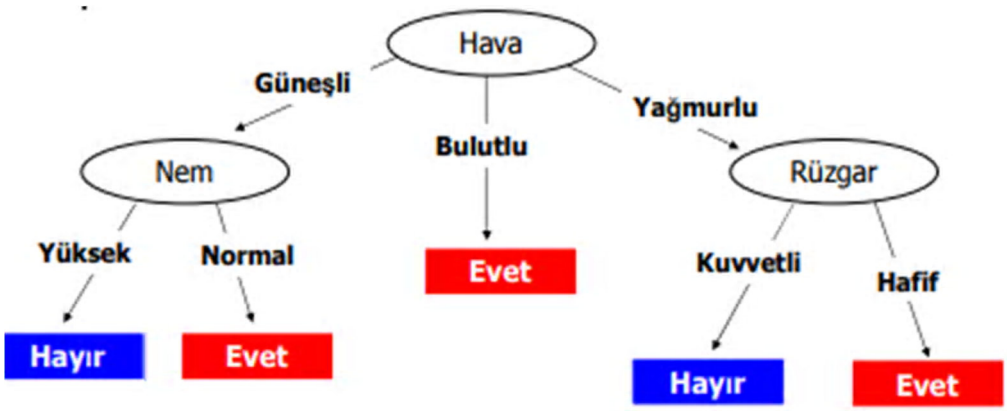
— (2.5)

S: orijinal veri seti,

D: orijinal veri setinin bölünmüş bir parçası,

V: S'nin bir alt kümesi, V'lerin birleşimi S'yi oluşturmaktadır.

Bu formülden yola çıkarak S sistemine göre D değişkeninin bilgi kazancı hesaplanır. Yani bölünmeden önceki orijinal verinin entropisinden her bir değişkene göre hesaplanan entropinin çıkarılmasıyla bilgi kazancı elde edilmiş olur. Böylelikle değişkenlerimiz entropiye göre kökten dal ve yapraklara doğru sıralanır kazanca göre de en iyi ayırma yapılır.



Şekil 2.5. Karar Ağacı

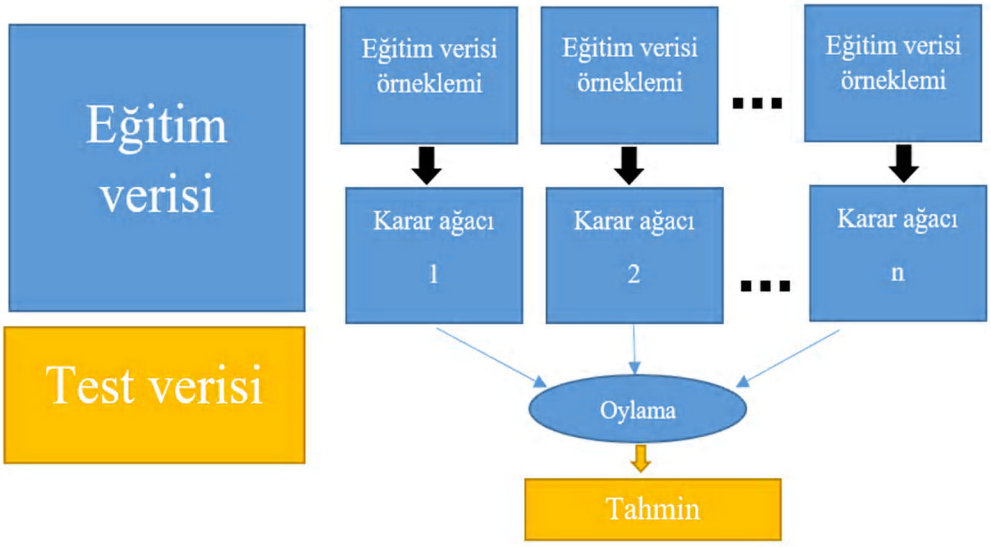
Şekil 2.4'de karar ağaçlarının anlaşılması için sıklıkla kullanılan bir örnek gösterilmektedir. Burada hava dediğimiz durum kök düğümdür. Ağaç bu değişken ile başlayıp havanın güneşli, bulutlu ya da yağmurlu olmasına göre dallara ayrılmaktadır. Bu örnekte tahmin edilmeye çalışılan durum dışarı çıkılacak mı çıkılmayacak mı

sorusunun cevabıdır. Şekil 2.4'de gösterilen ağaca göre eğer hava bulutlu ve nem normal ise dışarı çıkılacak, hava bulutlu ve nem yüksek ise dışarı çıkılmayacak şeklinde bir karar verilir. Hava yağmurlu ve rüzgar şiddetli ise dışarı çıkılmayacak, hava yağmurlu ve rüzgar hafif ise dışarı çıkılacak kararı verilir. Hava bulutlu ise başka bir koşul değerlendirilmeden dışarı çıkılacak şeklinde karar verilir (Dalkılıç ve Dalkılıç, 2015).

2.3.1.6. Random Forest (RF)

Karar ağacı algoritmalarından yola çıkılarak geliştirilen random forest algoritması topluluk öğrenme yöntemi arasında yer alır. Sınıflandırma işlemi sırasında çok sayıda zayıf sınıflandırma kabiliyeti gösteren birbirinden bağımsız karar ağaçlarını kullanarak yeni ve güçlü bir sınıflandırıcı inşa edip sınıflandırma performansını arttırmayı hedefleyen bir algoritmadır. Kullanılan eğitim verisinden faydalanarak meydana getirilen çok sayıda sınıflandırıcı karar ağaçlarının birleşmesiyle karar ormanı oluşur. Oluşturulan karar ağaçlarının bir araya gelmesi rastlantısal olarak gerçekleşir (Breiman, 2001).

Çok sayıda nitel değişken içeren, eksik gözlemlerin bulunduğu ve dağılımı dengesiz veriler kullanarak bu yöntemle başarılı sınıflandırmalar yapılabilir. Topluluktaki ağaç sayısı arttıkça test verisi ile yapılan sınıflandırmada hata tahmini için yanlılığı düşük sonuçlar alınır. Gürültülü veriler ile başa çıkabilir. Random forest algoritması makine öğrenme yöntemleri içerisinde oldukça popülerdir çünkü oldukça iyi tahmin geçerliliğine sahiptir ve modeli yorumlaması kolaydır. Karar ağaçlarını rastgele örnekleyerek seçer ve topluluk sınıflandırıma yöntemlerinin iyileştirilmiş özelliklerini kullanması sebebiyle random forest algoritması daha iyi genellemeler sağlar, geçerli tahminler sunar. Standart sapması düşük tahmin sonuçları vermesi ve ağaçlar arasındaki düşük korelasyon sayesinde kestirim düzeyi yüksek sonuçlar verir. Büyük ağaçlar oluşturularak düşük standart sapmalı sonuçlar elde edilir. Ağaçlar birbirinden ne kadar farklı olursa o denli düşük korelasyona sahip bir topluluk elde edilir.

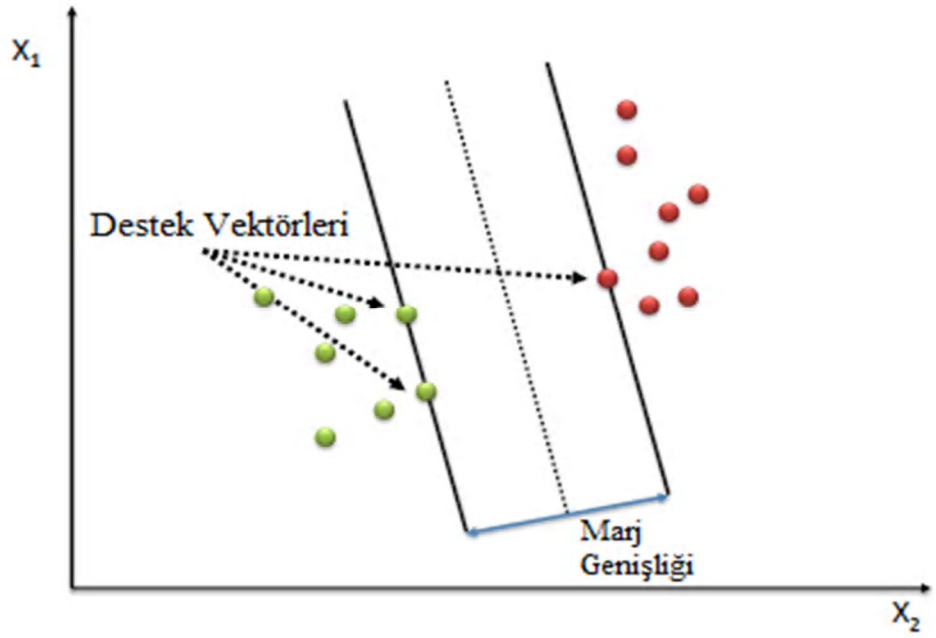


Şekil 2.6. Random forest algoritması

Random forest algoritmasında ağaçlar sürekli olarak verinin farklı alt grupları kullanılarak üretilir ve her bir düğüm dallara ayrılırken rastgele seçilen m tane değişken ile oluşturulur. m sayısının toplam ağaç sayısından çok daha küçük olması önerilir. Bu yöntem için oluşturulan karar ağaçlarında budama yapılmaz. RF algoritmasında kullanıcı tarafından belirlenmesi gereken iki parametre bulunur. Bu parametreler; m : en iyi bölünmeyi sağlamak amacıyla her bir düğümde kullanılan değişkenlerin sayısı, N : oluşturulacak ağaçların sayısı. m sayısı belirlendikten sonra, m tane değişken bütün değişkenler içinden rastgele seçilerek en iyi bölünmeyi sağlayacak olan değişkenler belirlenir. Toplam değişken sayısının karekökü ile orantılı m değerinin alınması genellikle optimum değere yakın kabul edilir. Sınıfların homojenliğini ölçmek için gini indeksi denklem 2.5 yardımıyla hesaplanır. Daha sonra test verisi kullanılarak oluşturulan modelin yeni gözlemler için sınıflandırma başarısı ölçülür. RF algoritması sadece sınıflandırma ve regresyon için kullanılmakla kalmayıp değişken seçimi içinde kullanılan bir yöntem haline gelmiştir. Dahası sınıf etiketinin bilinmediği durumlar için kümeleme analizi olarak kullanılabilen bir uzantısı da geliştirilmiştir (Afanador ve ark., 2016). Biyoinformatik alanında elde edilen verilerdeki problemlerle de başa çıkma konusunda oldukça başarılı performans gösteren bu yöntem kemometrik toplumda kabul görmüş ve sıklıkla kullanılmaktadır.

2.3.1.7. Destek Vektör Makineleri (DVM)

Destek vektör makineleri (DVM) sınıflandırıcı destek vektör makineleri ve regresyon destek vektör makinelerinden oluşur. Makine öğrenimi algoritmaları içinde en dayanıklı ve en iyi performansı gösteren yöntemlerden biridir. 1990'lı yıllarda bilgisayar bilimleri topluluğunca geliştirilmiştir (James ve ark., 2014). Destek vektör makinelerinin biyoinformatik gibi çeşitli alanlarda iyi performans göstermektedir. Sınıflar arasındaki en yüksek mesafe olarak tanımlanan marjı belirleyerek sınıflandırma yapılır. İyi performans göstermesi ve basit uygulanabilirliği olmasına rağmen her veri seti için uygulanamaz çünkü doğrusal bir ayırma yapmaktadır ve bu şekilde bir doğrusal ayırma birçok veride sağlanamayabilir. Destek vektör makineleri basitliği ve kolay uygulanabilirliği sayesinde hızla popülerlik kazanmış ve bu popülerlik sayesinde hızla yenilikler de kazanmıştır.



Şekil 2. 7.Doğrusal Destek Vektör Makinesi

Şekil 2.6'da iki sınıflı bir veride destek vektör makinesi ile maksimum marjın belirlendiği doğrusal bir sınıflandırma gösterilmiştir. Destek vektör makinelerinde en iyi sınıflandırmayı yapacak bir hiperdüzlem belirlenmeye çalışılır. Hiperdüzlem belirlenirken en yüksek marjın belirlenmesi gerekir ve tam orta noktası hiperdüzlem olarak kabul edilir. Matematiksel olarak iki boyutlu uzayda hiperdüzlem şu şekilde

yazılabilir:

$$\beta_0 + \beta_1 X_1 + \beta_2 X_2 = 0 \quad (2.6)$$

β_0, β_1 ve β_2 hiperdüzlemin parametreleri X= $(X_1, X_2)^T$ ise veride bir gözleme karşılık gelen noktadır. X_1 ve X_2 değişkenlerdir.

Eğitim verileri ile en iyi ayırma olan hiperdüzlem belirlendikten sonra yeni bir gözlemin X değişkenlerine ait değerleri denklem 2.6'da yerine yazıldığında elde edilen sonucun 0'dan büyük ya da küçük olmasına göre hiperdüzlemin sağında veya solunda yer alır ve böylelikle yeni gözlemin hangi sınıfta yer aldığı belirlenmiş olur (James ve ark., 2014).

Doğrusallığın sağlanamadığı durumlarda da kullanılabilen çeşitli destek vektör makineleri çekirdek fonksiyonları bulunmaktadır. DVM yöntemi uygulanırken çekirdek hilesi (kernel trick) olarak isimlendirilen bir durum söz konusudur. Bu doğrusal olarak ayrılamayan verilerde kullanılan bir yaklaşımdır ve verileri ifade ederken 2 boyutlu düzlemden 3 boyutlu uzaya taşıyarak yeni bir doğrusal düzlem ile ayrılabilen yapı oluşturulmaya çalışılır.

DVM'larında kullanılan bazı çekirdek fonksiyonlarına polinom, radyal, sigmoid, Gaussian, hiperbolik tanjant çekirdek fonksiyonları kullanan destek vektör makineleri örnek verilebilir. Hangi çekirdek fonksiyonu hangi durumlarda kullanmamız daha uygun olur bununla ilgili olarak karar verebilmek için verinin dağılımı, veri ile kullanılacak fonksiyonun uyumluluğu önemlidir. Aksi takdirde oluşturulan model ile veriyi sınıflandırma konusunda başarı elde edilemez (Heinemann ve ark., 2014).

2.3.1.8. En Yakın Küçültülmüş Merkezler (NSC)

Destek vektör makineleri, karar ağaçları, lojistik regresyon, doğrusal ayırma analizi gibi sınıflandırma metotları veri setinde bulunan tüm değişkenleri kullanarak model oluşturur. Bu yöntemler her ne kadar güçlü sınıflandırma yöntemleri olsa da omik verilerinde olduğu gibi yüksek boyutlu veriler ile kullanıldıkları zaman oluşturulan modellerin kompleksliği artar. En yakın küçültülmüş merkezler (NSC) yönteminde veride bulunan değişkenlerin tamamını kullanmak yerine sınıflandırma için en çok katkı sağlayacak olan değişkenler ile oluşturulan bir alt küme ile model oluşturulur (Tibshirani ve ark., 2003). Böylelikle verinin boyutu düştüğü için oluşturulan modelin varyansıda küçülür, daha kolay yorumlanabilen kompleksliği düşük modeller elde edilir. Bu yöntem köşegenel ayırma analizinin model karmaşıklığının üstesinden gelmek

için onun bir uzantısı olarak geliştirilmiştir. NSC yöntemi, değişkenlerin standartlaştırılmış genel ortalamaları ile standartlaştırılmış sınıf ortalamalarını ya da ortancalarını küçültür. Küçülmüş ortalamaya sahip olan değişkenler veriden elenir. Son olarak elemeden sonra kalan değişkenlerle köşegenel doğrusal ayırma analizi sınıflandırma modeli oluşturulur (Tibshirani ve ark., 2002; Zararsız, 2015). Oluşturulan model ile aşağıdaki ayırma fonksiyonunun (denklem 2.7) değeri maksimize edilmeye çalışılmaktadır; yani yeni bir gözlemin sınıfı belirlenirken bu fonksiyonun değerini hangi sınıf için en yüksek ise gözlem o sınıfa atanır.

$$\delta_k^{NSC}(x_*) = -\sum_{j=1}^{p} \frac{(x_{g*} - \bar{x}_{gk})^2}{(s_g + s_0)^2} + 2\log(\hat{\pi}_k) \tag{2.7}$$

Bu yöntem başlangıçta genomik verilerin analizi için geliştirilmiş olsa da daha sonra diğer omik veriler için uyarlaması yapılmıştır (Chen ve ark., 2015).

2.4. OTOMATİK MAKİNE ÖĞRENMESİ (AUTOML)

Otomatik makine öğrenimi, makine öğrenimi kısımlarını otomatikleştirmek için tasarlanmıştır. Kullanılır durumda olan otomatik makine öğrenimi araçları, makine öğrenmesi kullanan araştırmacıların işlerini kolaylaştırır ve gelişmiş analitik topluluk tarafından da kullanışlı bulunmaktadır. Araştırmacılar makine öğrenimi uygularken her aşamada birçok araç ve teknik bulunduğundan bütün bu bilgiler arasında kolayca kaybolabilir. Bu nedenle tüm süreci küçük parçalara ayırarak yönetmek gerekir. Otomatik makine öğrenimi süreci değişken ön işleme, değişken seçimi, model seçimi ve hiperparametre optimizasyonu gibi birçok bölümden oluşur. Bu bölümlerin her biri başarılı ve yansız sonuçlar elde etmek için özenle değerlendirilmelidir (Sibanjan Das ve Cakmak, 2018).

AutoML, makine öğrenim sürecinde tekrarlayan görevleri azaltarak makine öğrenmesi kullanıcıları, biyoistatistikçiler ve biyolog gibi bu yöntemleri kullanmaya ihtiyaç duyup alanda uzman olmayanlar için büyük verimlilik artışı vaat etmektedir. Şu anda bu görevleri otomatikleştirmeye çalışan çok sayıda farklı araç ve platform (hem açık kaynaklı hem de ticari olarak temin edilebilen) bulunmaktadır. Bu yöntemler ile ilgili olarak (i) autoML tarafından gerçekleştirilebilen mevcut makine öğrenmesi işlevleri nelerdir, (ii) geniş bir yelpazedeki gerçek veri setleri ile çalışıldığında bu yöntemlerin

nasıl performans gösterdiği, (iii) optimizasyon hızı ve sonuçların doğruluğu arasındaki denge, (iv) sonuçların tekrar üretilebilirliği, geçerlilik ve güvenilirlikleri ile ilgili olarak araştırmalar devam etmektedir (Truong ve ark., 2019)

AutoML kısa geçmişini inceleyecek olursak 1993-2015 yılları arasında Weka (1993), RapidMiner (2001), Scikit-learn (2007), caret, H2O (2011) ve Spark MLlib (2013) gibi birçok makine öğrenmesi kütüphanesi ve aracı geliştirilmiştir. Daha sonra son 5 yılda derin sinir ağları popülerlik kazanmaya başlamıştır. Tensorflow (2015), Keras (2015) ve MXNet (2015) gibi araçlar derin öğrenme modellerinin geniş çapta kabul görmesine katkı sağlamıştır. Bu süreçte makine öğrenmesi yöntemleri kullanarak iyi bir performans elde edebilmenin insan uzmanlığına bağlı olduğu görülmüştür. İlk olarak akademik alanda AutoML çalışmaları başlamış daha sonra ticari girişimlerde de bulunulmuştur.

İlk AutoML algoritması British Columbia ve Freiburg üniversitelerindeki araştırmacıların Weka kütüphanesinde bulunan makine öğrenimi algoritmalarını kullanarak geliştirmiş oldukları Auto-Weka'dır. (Thornton ve ark., 2013). Daha sonra yine Freiburg üniversitesinden araştırmacılar Auto-Sklearn'ü geliştirmişlerdir (Feurer ve ark., 2015). Auto-Sklearn'ün ardından Pennsylvania Üniversitesinden araştırmacılar açık kaynak kodlu Python kütüphanesi olan TPOT'u geliştirmişlerdir (Olson ve ark., 2016b). auto_ml (genel "AutoML" terimiyle karıştırılmamalıdır) açık kaynak kodlu Python paketidir ve 2016 yılında ortaya çıkmıştır (https://github.com/ClimbsRocks/auto_ml, 2019). TPOT, Auto-sklearn, Auto-ml Python makine öğrenimi paketi olan scikit-learn üzerine inşa edilmiştir. Auto-Keras sinir mimarisi aramaya yönelik olarak geliştirilmiş bir araçtır (Jin ve ark., 2019b). Bu yöntemlerden başkada AutoML araçları geliştirilmiş ve geliştirilmeye de devam etmektedir (Truong ve ark., 2019).

Geliştirilmekte olan, makine öğrenimi sürecinin otomatikleştirilmesini hedefleyen AutoML yöntemleri Tablo 2.1'de isimleri, hangi dilde geliştirildikleri ve kısa tanımları ile birlikte verilmiştir.

Tablo 2.1. Otomatik makine öğrenimi yaklaşımları ve özellikleri

İsim	Dil	Tanım
TPOT	Python	Ardışık düzende ağaç tabanlı optimizasyon aracı. TPOT genetik programlamanın bir versiyonunu kullanır.
Auto-Keras	Python	Derin öğrenme modellerinin mimarisini ve hiperparametrelerini otomatik olarak arar. Keras kütüphanesini kullanır.
SMAC	Python	SMAC (sıralı model tabanlı algoritma konfigürasyonu), algoritmaların yapılandırılması için parametrelerini optimize etmek için yetenekli bir araçtır. Makine öğrenme algoritmalarının hiperparametre optimizasyonu için de çok etkili ve hızlıdır. İki yapılandırmadan hangisinin daha iyi performans gösterdiğine karar vermek için agresif yarış mekanizması ile birlikte Bayes optimizasyonunu kullanır.
Auto-WEKA	Java	Weka yazılımında bulunan modeller için otomatik model seçimi ve hiperparametre ayarlaması yapar.
auto-sklearn	Python	Otomatik makine öğrenimi araç takımı. Scikit-learn paketinde bulunan modelleri kullanır.
auto_ml	Python	Analitik ve üretim için otomatik makine öğrenimi. Manuel olarak değişken türü bildirimlerini gerektirir.
H2O	Java	H2O, H2O AutoML olarak da adlandırılan açık kaynaklı ve makine öğrenim iş akışını otomatikleştirmek için kullanılan, kullanıcı tarafından belirlenen bir süre içinde birçok modelin otomatik eğitimini ve hiperparametre ayarını içeren otomatik bir makine öğrenme modülüdür.
devol	Python	Genetik programlama ile otomatik derin sinir ağı tasarımı yapar.
MLBox	Python	Dağıtılmış hesaplama desteği ile yüksek boyutlu uzayda doğru hiper-parametreyi belirlemek için optimizasyon yapar.
Recipe	C	Genetik programlama ile makine öğrenimi iş akışının optimizasyonu yapar.

Xcessiv	Python	Python'da hızlı, ölçeklenebilir ve otomatik hiper parametre ayarlama ve yığılmış topluluk modelleri için web tabanlı bir uygulamadır.
GAMA	Python	Asenkron değerlendirme tabanlı genetik programlama ile makine öğrenmesi sürecini optimize eder.
Amozon Lex	-	Konuşmayı metne ve doğal dil anlayışına (NLU) dönüştürmek için otomatik konuşma tanımanın (ASR) gelişmiş derin öğrenme işlevlerini sağlar ve kullanıcının oldukça gerçekçi konuşma etkileşimleriyle uygulamalar oluşturmasını ve ilgi çekici kullanıcı deneyimleri sağlar.
Auto-PyTorch	Python	Auto-PyTorch, en iyi ayarları aramak için çoklu uygunluk optimizasyonunu ve Bayesian optimizasyonunu (BOHB) kullanarak doğru mimariyi ve hiperparametre ayarlarını bulmayı otomatik hale getirir.
ROBO	Python	Sağlam Bayes Optimizasyonu (RoBO) sistemi farklı Bayes optimizasyon bileşenlerini kolayca ekleme ve değiştirmeyi sağlayan modüler bir sistemdir. Gauss prosesleri, RF veya Bayes sinir ağları ve farklı kazanım fonksiyonu içeren çeşitli regresyon modelleri içerir.
AutoFolio		AutoFolio, en iyi seçim yaklaşımını ve hiperparametreleri belirleyerek algoritma seçim sistemlerinin performansını optimize etmeye çalışır.
Flexfolio		Birkaç farklı portföy tabanlı algoritma seçimi yaklaşımı ve tekniklerini birleştiren modüler bir sistemdir. Algoritma seçim yaklaşımı ve tekniklerini karşılaştırmak ve birleştirmek için kullanışlı bir sistemdir.

Elde edilmiş olan bir veriden işe yarar anlamlı bilginin çıkarılması için çok fazla araç ve yönteme sahip olunduğunda avantajlı olan durum bir sorun haline gelir. Verilen probleme uygun doğru yaklaşımı belirlemek ve anlamak için oldukça fazla zaman ve uğraş gerekir. Yüksek performanslı sınıflandırma sonuçları elde etmek için makine öğrenimi basamaklarını özenli bir şekilde yerine getirmek gerekir. Her adım bir sonraki

adıma temel oluşturmakta olup sürecin düzgün işleyebilmesi ve genellenebilir sonuçlar elde edebilmek için son basamağa kadar dikkat edilmesi gerekir. Her adımda çok sayıda yöntem ve olası farklı kombinasyon sayısı düşünüldüğünde, makine öğrenimi sürecindeki tüm bu bileşenleri denemenin sistematik bir yoluna ihtiyaç duyulmaktadır (Sibanjan Das ve Cakmak, 2018). Bu ihtiyaca yönelik olarak otomatik makine öğrenimi araçları geliştirilmiştir. Problemin türüne, otomatik gerçekleştirilen makine öğrenimi basamağına ve otomatikleştirme yaklaşımına göre farklı otomatik makine öğrenimi yaklaşımları bulunmaktadır.

2.4.1. H2O

Makine öğrenim uzmanlarına olan talep son yıllarda bu alanda çalışmaya başlayanların da artmasına rağmen yeterli olmamıştır. Bu eksikliği gidermek için uzman olmayan kişiler tarafından da kullanılabilecek bir kullanıcı dostu makine öğrenme sisteminin geliştirilmesine yönelik çalışmalar başlamıştır. Makine öğrenmesini basitleştirmeye yönelik ilk adımlar, H2O gibi çeşitli makine öğrenme algoritmalarına yönelik basit ve birleşik arayüzler geliştirmeyi içeriyordu. Her ne kadar H2O bu alanda uzman olmayan kişilerinde makine öğrenmesi ile çalışabilmelerini kolaylaştırsa da genellenebilen, başarılı makine öğrenimi modelleri üretmek için bilgi ve deneyim sahibi olmayı gerektirmektedir. Özellikle Derin Sinir Ağları gibi daha kompleks yöntemler uzman olmayan kişiler tarafından uygulaması güç olmaya devam etmektedir. Makine öğrenimi yazılımının uzman olmayanlarında kullanabilir olmasını sağlamak için kullanımı kolay, çok sayıda modelin oluşturulmasını otomatikleştiren, bir arayüz tasarlanmıştır. H2O AutoML'i kodlama becerisi gerektiren çok sayıda yöntemin görevini otomatik olarak yerine getiren basit bir sarmalayıcı ile araştırmacıların zamanını veri ön işleme, değişken seçimi, iş akış süreci gibi diğer araştırma bölümlerine yöneltmesine yardımcı olur.

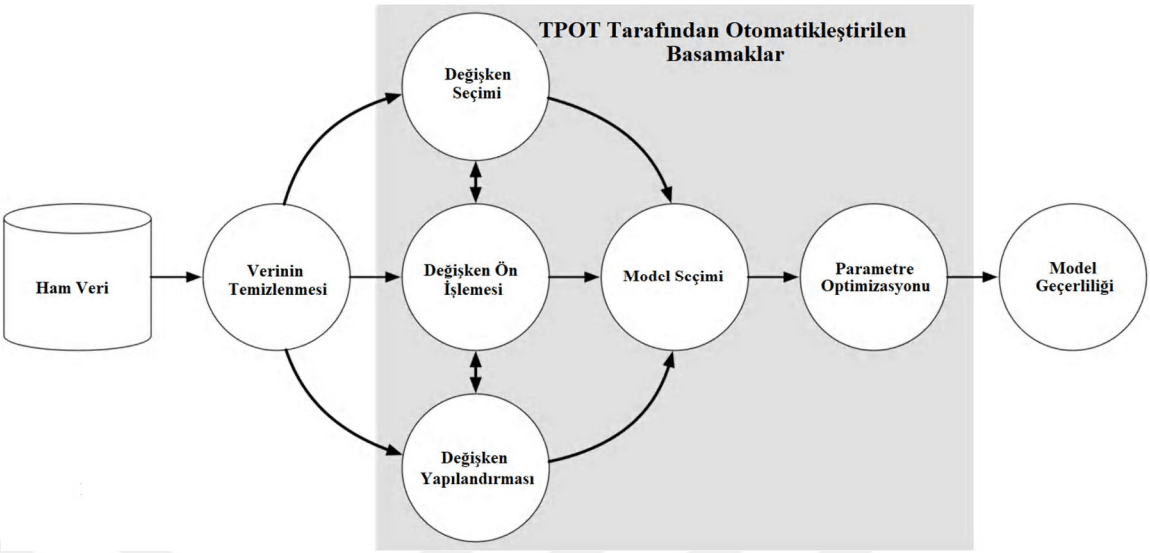
H2O'nun AutoML yöntemi kullanıcının belirtmiş olduğu süre zarfında birçok modelin otomatik olarak eğitimi ve hiperparametre ayarını içeren makine öğrenme kısmını otomatikleştirmek için kullanılabilir. H2O, çeşitli programlama dilleri ve arayüzler aracılığıyla kullanılabilen sunucu kümesinde yürütülmektedir. Genelleştirilmiş doğrusal modelleri (GLM), gradiyent artırma makinelerini (GBM), temel derin sinir ağları (DNN) modellerini ve yoğun rastgele ormanlar (DRF) gibi yöntemleri desteklemektedir. Parametre optimizasyonu için rastgele parametre araması (RGS) ve kartezyen parametre

arama (CGS) yöntemlerini destekler. Daha açık bir şekilde belirtecek olursak H2O mevcut sürümü sırayla aşağıdaki algoritmaları eğitir; önceden belirlenmiş üç farklı XGBoost GBM modelleri, sabit aralıklı parametre arayan GLM'ler, varsayılan RF (DRF), önceden belirlenmiş beş H2O GBM'leri, varsayılan derin sinir ağı (DNN), aşırı derecede rastgele ormanlar (XRT, Extremely Randomized Forest), XGBoost GBM'lerin rastgele parametre arama, H2O GBM'lerin rastgele parametre araması ve rastgele parametre arayan derin sinir ağları denenir. Mevcut veri için iyi çalışacak olan algoritmalar hakkında bir fikir sahibi olunduğunda bu algoritmalara yönelik denemeler yapmak yararlı olacaktır ancak bu bazen performans kaybına neden olabilir.

H2O Java programlama dili kullanılarak geliştirilmiştir. R, Tableau, Javascript, Python ve Flow (web UI) bağlamaları içerir (Balaji ve Allen, 2018; http://docs.h2o.ai/h2o/latest-stable/h2o-docs/automl.html, 01.12.2019)

2.4.2. TPOT

TPOT (Ağaç Tabanlı İş Akışı Optimizasyon Aracı) genetik programlama kullanmaktadır. Açık kaynak kodlu bir Python kütüphanesidir. TPOT scikit-learn algoritmalarını kullanarak çalışan bir scikit-learn sarımı (wrapper) olarak tanımlanır. Otomatik hiperparametre ayarlaması, scikit-learn algoritmaları modellemeler oluşturma, değişken seçimlerinin araştırılması, yinelenen model oluşturma ve model değerlendirmesi gibi makine öğrenmesi aşamalarını gerçekleştirebilmektedir. TPOT analiz sonucunda en iyi model performansı gösteren yöntemin scikit-learn kodlarını da vermektedir. Denenen tüm modellerin en iyi performansa sahip olanını temsil eden bu kod daha sonra üzerinde tekrar çalışılabilir, uygun yöntemi başlangıç noktası olarak kabul edilerek değiştirip incelenebilir. (Olson ve ark., 2016a).



Şekil 2.8 TPOT otomatik olarak gerçekleştirilen makine öğrenme basamakları

Şekilde makine öğrenme sürecinin TPOT tarafından otomatik olarak gerçekleştirilen aşamaları gösterilmektedir. Kullanılacak yöntemlerle ilgili fazlaca kod satırları ile uğraşmaya gerek kalmadan TPOT kullanılarak iyi performans gösteren sonuçlar alınabilmektedir. Makine öğrenmesi uygulamalarında sıklıkla denemeler yapılan örnek iris verisi için TPOT kod satırları şu şekildedir;

```
from tpot import TPOTClassifier
from sklearn.datasets import load_iris
from sklearn.model_selection import train_test_split
import numpy as np

iris = load_iris()
X_train, X_test, y_train, y_test = train_test_split(iris.data.astype(np.float64),
    iris.target.astype(np.float64), train_size=0.75, test_size=0.25, random_state=42)

tpot = TPOTClassifier(generations=5, population_size=50, verbosity=2, random_state=42)
tpot.fit(X_train, y_train)
print(tpot.score(X_test, y_test))
tpot.export('tpot_iris_pipeline.py')
```

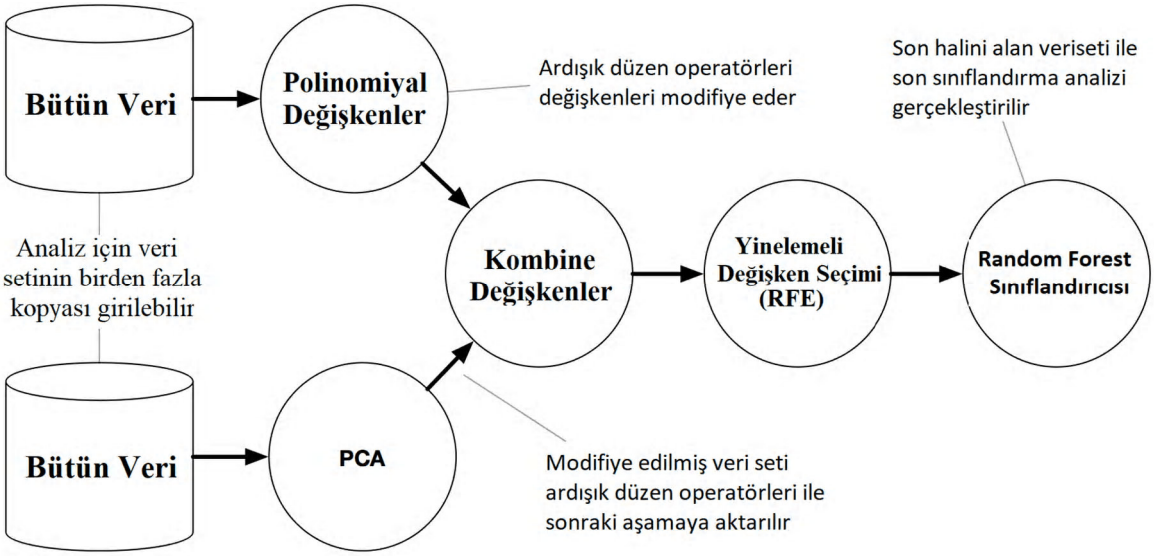
Son satırda yer almakta olan tpot.export kodu en iyi performansı gösteren kod satırlarını almamızı sağlar. Aşağıda yukarıda yazmış olduğumuz kod satırları yardımı ile eğitilmiş olan modellerden en iyi olarak elde etmiş olduğumuz modelin kod satırları görülmektedir.

```
import numpy as np
import pandas as pd
from sklearn.model_selection import train_test_split
from sklearn.neighbors import KNeighborsClassifier
from sklearn.pipeline import make_pipeline
from sklearn.preprocessing import Normalizer
from tpot.export_utils import set_param_recursive

# NOTE: Make sure that the outcome column is labeled 'target' in the data file
tpot_data = pd.read_csv('PATH/TO/DATA/FILE', sep='COLUMN_SEPARATOR', dtype=np.float64)
features = tpot_data.drop('target', axis=1)
training_features, testing_features, training_target, testing_target = \
            train_test_split(features, tpot_data['target'], random_state=42)

# Average CV score on the training set was: 0.9826086956521738
exported_pipeline = make_pipeline(
    Normalizer(norm="l2"),
    KNeighborsClassifier(n_neighbors=5, p=2, weights="distance")
)
# Fix random state for all the steps in exported pipeline
set_param_recursive(exported_pipeline.steps, 'random_state', 42)

exported_pipeline.fit(training_features, training_target)
results = exported_pipeline.predict(testing_features)
```

Elde etmiş olduğumuz kod satırlarını yeniden revize edebilir yeni özellikler ekleyip çıkarabiliriz. Diğer AutoML yöntemlerinde de olduğu gibi TPOT'un birkaç dakika içinde çalışıp hızlı bir şekilde sonuçlanması beklenmemelidir. Kısa süreli çalıştırılması sonucunda da iyi sonuçlar elde edilse bile en iyi sonucu alabilmek için daha uzun sürelerde çalıştırmak daha faydalı olur. Veri boyutu arttıkça TPOT çalışma süresi de arttırılmalıdır. TPOT'un avantajlı özelliklerinden biri de günlerce ya da saatlerce süren analizi istenilen noktada kesip o anki performansın yeterince iyi olup olmadığı kontrol edilebilir. Eğer yeterince iyi bir performans elde edilemediyse TPOT kaldığı yerden aramaya devam etmek üzere warm_start ismi verilen bir parametre kullanılarak tekrar başlatılır. Aynı veri seti için aynı kodu kullanarak çalıştırılan TPOT yeterince uzun süre çalıştırılmıyorsa her seferinde farklı yöntemleri en iyi olarak önerebilir. Buna veri için benzer sonuçlar veren birden fazla yöntemin olması ya da çalışma esnasında süre yeterli olmadığı için bir rastgele taranan çözüm uzayında aynı yönteme denk gelinememiş olması neden olabilir (Le ve ark., 2019)

Şekil 2.8'de örnek bir TPOT sistemi gösterilmektedir. Bu şekilde de görüldüğü gibi kullanılacak olan veri seti analizler için aynı anda birden fazla kez girilebilmektedir. Aynı veri farklı veri ön işleme yöntemleri ile değerlendirilip tekrar kombine edilerek kullanılabilir. Sonraki aşamada ise rutin makine öğrenmesi aşamalarından olan değişken seçimi uygulanıp veri setinin son halini almasıyla sınıflandırma modelinin belirlenmesi

aşamasına geçilmekte ve uygun model ve hiper parametre ile sınıflandırma performansı elde edilmektedir.

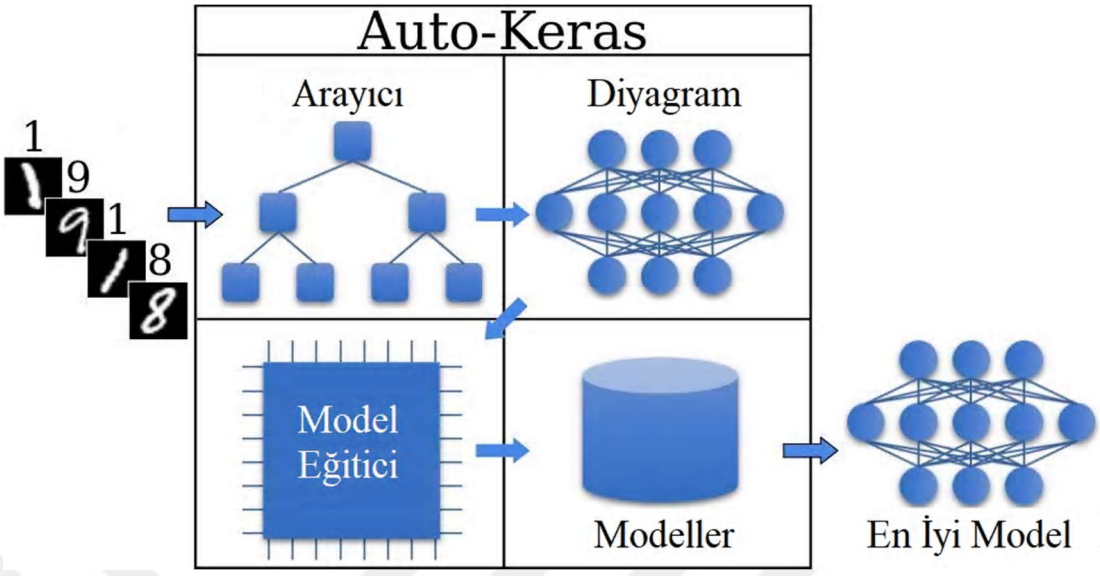


Şekil 2.9 Örnek bir TPOT model oluşturma süreci

TPOT halen aktif olarak geliştirilmekte olup çeşitli güncellemeler ve düzeltmeler devam etmektedir.

2.4.3. Auto-Keras

Auto-Keras AutoML için açık kaynaklı bir yazılım kütüphanesidir. Auto-Keras derin öğrenme modellerinin mimarisini ve hiperparametrelerini ayarlamak için otomatik fonksiyonlar sağlar. Texas A&M üniversitesinde, veri laboratuvarı topluluğu katkılarıyla geliştirilmiştir. AutoML'in nihai amacı doğrultusunda sınırlı veri bilimi ve makine öğrenmesi deneyimi olan kişilerin kolayca erişilebileceği ve uygulamalar yapabileceği derin öğrenme araçları sağlamaktır. Auto-Keras Pythonda yazılan bir kütüphanedir ve TensorFlow, CNTK veya Theano'nun üzerinden çalıştırılabilen yüksek düzeyli sinir ağları uygulaması gerçekleştirir (Jin ve ark., 2019a). R kullanıcıları içinde uygulama kullanıcı arayüzü (API) ile kullanılabilen Auto-Keras kütüphanesi geliştirilmiştir. Görüntü ve yazı verilerin sınıflandırma ve regresyon analizlerini yapabilmektedir. Auto-Keras görüntü sınıflandırıcısının kullanacağı nöral ağı bilmeyi gerektirmemek gibi önemli avantajlara sahiptir. Auto-Keras, farklı katman sayısına sahip birden fazla sinir ağını deneyerek en iyi mimariyi bulur. Aşağıdaki şekil 2.9'da Auto-Keras'ın model belirleme aşamaları gösterilmektedir.



Şekil 2.10. Auto-Keras model belirleme süreci

2.4.4. Auto_ml

Guyon ve ark. (2015) makine öğrenimi algoritmalarının başarısından ve bu başarıda etkili olan uygun özellikler, iş akışı, makine öğrenme paradigmaları, algoritmaları ve parametrelerini belirleyen araştırmacının öneminden bahsetmiştir. AutoML makine öğrenim süreci boyunca araştırmacının belirlediği bütün bu aşamaların otomatik gerçekleştirilmesini ifade eder. Otomatik makine öğreniminin model seçimi, parametre optimizasyonu ve model arama aşamalarının tüm yönleri insan gücü kullanmadan otomatik olarak gerçekleştirmesi ve en iyi performans sağlayan durumu göstermesi beklenmektedir. Auto_ml algoritması uygulamalar için Scikit-Learn, XGBoost, TensorFlow, Keras ve LightGBM gibi yüksek düzeyde optimize edilmiş kütüphaneleri kullanır. Auto_ml şirketlerin hızlı bir şekilde değer elde etmelerini sağlamak için üretim sistemlerinde kullanılmak üzere veriden elde edilen değeri müşterilere aktarmak amacıyla tasarlanmıştır. Auto_ml makine öğrenme sürecinin bir çok noktasını otomatik hale getirir. İlk olarak değişken mühendisliği kısmında kategorik değişkenleri kodlama sayısal değişkenleri ölçeklendirme, tarih işleme ve doğal dil işleme gibi işlevlerle bu kısmı otomatik hale getirir. Tarih ön işleme haftanın günlerini hafta içi ve hafta sonu şeklinde ikili ayırmayı ve gün ay yıl olarak bileşenlere ayırmayı içermektedir. Bu değişken mühendisliği yöntemleri dışında 100.000 üzerinde değişkenin bulunduğu durumlarda temel bileşenler analizi (PCA) yaparak boyut indirgemede yapabilir. Bu

kütüphane doğru ön işleme yapabilmek için her bir değişkenin tipini girdi olarak ister. Değişken mühendisliği dışında model oluşturma, ayarlama, seçme ve modellerin birlikte kullanımını otomatik hale getirir. Sklearn-deap Python kütüphanesini kullanarak evrimsel bir sistem arama (evolutionary grid search) algoritması kullanarak model optimizasyonu gerçekleştirir. Üstün özelliklerine rağmen zayıf genişletilebilirlik ve çok sınıflı sınıflandırma problemleriyle kötü performans sergileme gibi dezavantajları da vardır. Ayrıca bu yöntem için zaman sınırlaması yapılamamaktadır (Balaji ve Allen, 2018; Guyon ve ark., 2015).

2.4.5. Auto-sklearn

Auto-sklearn, araştırmacıları algoritma seçimi ve parametre optimizasyonun aşaması için çaba harcamaktan kurtarır ve otomatik olarak gerçekleştirir. Bu yöntem Bayes optimizasyon, meta-öğrenme ve topluluk oluşturma (ensemble construction) avantajlarından yararlanmaktadır. Auto-sklearn, aşağıdaki adımları yineleyerek ilerleyen Bayes optimizasyonu vasıtasıyla parametre optimizasyonunu gerçekleştirir (Feurer, 2015):

- Parametre optimizasyonu ve performansları arasındaki ilişkiyi yakalamak için olasılık modeli oluşturmak.
- En iyi performansı veren parametreyi belirlemek için oluşturulan modeli kullanarak parametre optimizasyonu yapmak.
- Makine öğrenme algoritmasını bu parametre ayarlarıyla çalıştırmak.

Auto-sklearn otomatik makine öğrenme süreci oluşturmak için scikit-learn kütüphanesini kullanır. Auto-sklearn kategorik değişkenin yeniden kodlanması, sayısal değişken standardizasyonu, temel bileşenler analizi (PCA) ile boyut indirgeme gibi değişken mühendisliği yöntemlerini içerir. Sınıflandırma ve regresyon problemleri için kullanılabilmektedir. Auto-sklearn bir iş akışı oluşturur ve bunu bayes araması kullanarak optimize eder. Hiperparametre optimizasyon süreci önceden eğitilmiş, OpenML'den 140 veri seti kullanılarak elde edilmiş varsayılan hiperparametre değerleri ile başlatılır. Bir veri seti için 38 istatistik hesaplanır ve hiperparametreleri eğitim verisine en yakın istatistiklere sahip optimize edilmiş parametrelere göre başlatır.

2.4.6. Auto-WEKA

Makine öğrenmesi araştırmacıları WEKA gibi açık kaynaklı paketler aracılığıyla çok çeşitli karmaşık öğrenme algoritmaları ve değişken seçim yöntemleri sunarak kullanıcılara yardımcı olmuştur. WEKA kullanıcıdan uygulama yapabilmesi için iki çeşit seçim yapmasını ister: (i)öğrenme algoritması seçmek ve (ii) parametre ayarlamak. Birçok farklı makine öğrenimi algoritmasının ve bunların çeşitli parametrelerinin ayarlanması çok sayıda alternatif oluşturmaktadır. WEKA 2 topluluk sınıflandırma metodu, 10 meta-öğrenme metodu, 27 temel sınıflandırıcı ve her sınıflandırıcı için parametre ayarlarını kapsar. Thornton ve ark. (2013) WEKA paketinde uygulanan tüm sınıflandırma algoritmaları ve değişken seçiciler/değerlendiriciler için bu alternatifleri otomatik olarak hesaplayan Auto-WEKA aracını oluşturdu. Auto-sklearn ile aynı prensip kullanılmaktadır. Çeşitli veriler üzerinde başarısı gösterilmiş olan bu yöntem ile uzman olmayan kullanıcıların, verilerine uygun makine öğrenme algoritmaları ve hiperparametre ayarlarını daha etkin bir şekilde belirlemelerine ve dolayısıyla daha iyi performans elde etmelerine yardımcı olmak amacıyla geliştirilmiştir (Guyon ve ark., 2019).

Bir veri seti ile otomatik ve eş zamanlı olarak bir öğrenme algoritması seçip ve deneysel olarak performansını optimize etmek için hiperparametrelerini ayarlama, kısaca kombine algoritma seçimi ve hiperparametre optimizasyonu (CASH) problemi auto-WEKA ile çözümlenmeye çalışılmaktadır. Bu sorunun pratikte önemine rağmen daha önce literatürde göz önüne alınmamıştır. Bunun nedeni öğrenme algoritmalarının ve hiperparametrelerinin birleşik uzayının arama yapmak için çok zorlayıcı olması, verilerin yüksek boyutluluğu, hem kategorik hem de nicel değişkenler içeren verilerin olması gibi durumlar olabilir (Thornton ve ark., 2013). Bu problemden yola çıkarak ilk otomatik makine öğrenimi yaklaşımı olan auto-WEKA geliştirilmiştir.

Otomatik makine öğrenimi yaklaşımları performans açısından çeşitli veriler ile denenmiştir. Henüz yeni geliştirilmekte olan bu yöntemlerin omik veriler ile denemesi araştırılmamıştır. Yöntemleri uygulamadan önce omik verileri anlamak, veri yapısını bilmek gerekir. Omik verilerde hem standart makine öğrenimi yaklaşımlarının hem de otomatik makine öğrenimi yaklaşımlarının performansı araştırılacaktır.

3. GEREÇ VE YÖNTEM

Bu çalışmada başarısı kanıtlanmış, birçok avantajlı özelliğe sahip sınıflandırma algoritmalarından, literatürde omik verilerin sınıflandırılmasında da sıklıkla tercih edilen DVM, RF algoritması, NSC yöntemi ve otomatik makine öğrenimi (AutoML) yöntemleri kullanıldı (Yi ve ark., 2016; Madsen ve ark., 2010). Genel bilgiler bölümünde de bahsedilmiş olduğu gibi birçok sınıflandırma yöntemi bulunmaktadır. Bu yöntemler arasından belirlemiş olduğumuz RF, SVM ve NSC yöntemleri diğer yöntemlere göre varsayımlarının nispeten daha az olma, omik verilere özgü problemlerle baş edebiliyor olma, omik verilere bu yöntemlerin uygulanmasıyla başarılı sonuçların alınmış olması, geniş bir kullanım alanına sahip olma gibi özellikleri nedeniyle tercih edilmiştir (Liu ve ark., 2005; Mahadevan ve ark., 2008; Schaalje ve Fields, 2012; Witten ve Tibshirani, 2011). AutoML yöntemlerinden H2O ve TPOT ile uygulamalar gerçekleştirildi. H2O ile yapılan analizler analiz süresi için 1, 2 ve 3 saatlik zaman dilimleri ayarlanarak denendi. Geliştirilmiş ve gelişmekte olan birçok otomatik makine öğrenmesi arasından H2O ve TPOT yöntemlerinin belirlenmesi aşamasında bu yöntemlerin kullanım kolaylığı, analizlerin maliyeti, literatürde kullanılacak yönteme ilişkin yeterli kaynağın bulunması, yeterli sayıda uygulamanın yapılmış olması, yöntemlerin farklı alandan veriler için kullanılabilir durumda olması gibi kriterler göz önüne alınarak seçim yapıldı. Uygulamalarda RNA dizileme verileri, mikrodizi verileri ve metabolomik verileri kullanıldı. Birçok omik bilimi bulunmasına rağmen bu omik bilimlerinden elde edilen verilerin yapısı kullanılan teknoloji benzer ise genel itibariyle gürültü, elde edilen verinin hacmi gibi durumlar açısından benzerlik gösterir. Metagenomik, epigenomik gibi omik bilimleri ile elde edilen verilerin RNA dizileme verilerine benzer olması nedeniyle bu omik bilimlerine ait veriler ile çalışılmadı (MacLean ve ark., 2009; Meng ve ark., 2013). Ayrıca metabolomik ve proteomik alanlarından elde edilen verilerde benzer teknolojiler kullanılarak elde edilebilir (Kamel

ve ark., 2012; Lange ve ark., 2008; Stoop ve ark., 2010). Analizlerin maliyetlerinin yüksek olması nedeniyle her omik bilimi ile uygulama yapmak yerine farklı yapıda olan omik bilimlerinden verilerin kullanılması tercih edildi. Kullanılan verilerin tamamı gerçek veri setleri olup çalışmalarda açık olarak paylaşılan veri setleridir. Toplamda 6 RNA dizileme verisi, 16 mikrodizi verisi ve 7 metabolomik verisi olmak üzere 29 farklı omik verisi kullanıldı.

Tablo 3.1. Çalışmada kullanılan omik verilerin özellikleri

Veri ismi	Sınıf sayısı	Örnek sayısı	Değişken sayısı	Veri türü
Rahim Ağzı Kanseri (Witten ve ark., 2010)	2	58	714	RNA dizileme verisi
Alzheimer (Leidinger ve ark., 2013)	2	70	2801	RNA dizileme verisi
Fare Embriyo (Deng ve ark., 2014)	10	268	22431	RNA dizileme verisi
Zebra Balığı (Singh ve ark., 2018)	3	212	20651	RNA dizileme verisi
Lemfoblastoid (Montgomery ve ark., 2010)	2	129	52580	RNA dizileme verisi
Fare Kök Hücre (Kolodziejczyk ve ark., 2015)	3	704	38616	RNA dizileme verisi
Alizadeh-V1 (Alizadeh ve ark., 2000)	2	42	1095	Mikrodizi verisi
Alizadeh-V2 (Alizadehve ark., 2000)	3	62	2093	Mikrodizi verisi
Armstrong-V1 (Armstrong ve ark., 2001)	2	72	1081	Mikrodizi verisi
Armstrong-V2 (Armstrongve ark., 2001)	3	72	2194	Mikrodizi verisi
Bhattacharjee (Bhattacharjee ve ark., 2001)	5	203	1543	Mikrodizi verisi
Bittner (Bittner ve ark., 2000)	2	38	2201	Mikrodizi verisi
Bredel (Bredel ve ark., 2005)	3	50	1739	Mikrodizi verisi
Chen (Chen ve ark., 2002)	2	180	85	Mikrodizi verisi
Chowdary (Chowdary ve ark., 2006)	2	104	182	Mikrodizi verisi
Dyrskjot (Dyrskjøt ve ark., 2002)	3	40	1203	Mikrodizi verisi

Garber (Garber ve ark., 2001)	4	66	4553	Mikrodizi verisi
Gordon (Gordon ve ark., 2002)	2	181	1626	Mikrodizi verisi
Khan (Khan ve ark., 2001)	4	83	1069	Mikrodizi verisi
Laiho (Laiho ve ark., 2007)	2	37	2202	Mikrodizi verisi
Lapointe-V1 (Lapointe ve ark., 2004)	3	69	1625	Mikrodizi verisi
West (West ve ark., 2001)	2	49	1198	Mikrodizi verisi
ST000369 (Fahrmann ve ark., 2015)	2	172	511	Metabolomik verisi
ST000389 (Fahrmann ve ark., 2016)	2	94	129	Metabolomik verisi
ST000388 (Fahrmannve ark., 2016)	2	94	989	Metabolomik verisi
ST000390 (Wikoff ve ark., 2015)	2	78	183	Metabolomik verisi
ST000391 (Wikoffve ark., 2015)	2	78	8869	Metabolomik verisi
ST000392 (Miyamoto ve ark., 2015)	2	164	158	Metabolomik verisi
ST000356 (Xie ve ark., 2015)	4	134	101	Metabolomik verisi

Tablo 3.1'de kullanılan verilerin isimleri ile birlikte kaynakları, grup değişkeninin sınıf sayısı, çalışmada bulunan örnek sayıları, içerdikleri değişken sayıları ve veri türü gösterilmiştir. Kullanılan verilerin özellikleri aşağıda kısaca bahsedilmiştir.

Rahim Ağzı Kanseri: Bu veri 29 rahim ağzı kanseri olan ve 29 sağlıklı gözlemden elde edilmiştir. Toplamda 58 küçük RNA kütüphanesinden oluşan bu verinin kanser alt tiplerine göre dağılım; 21 squamous hücre karsinoması, 6 adeno karsinoma ve 2 de alt tipi sınıflandırılmamış gözlem şeklindedir. Solexa/Illumina dizileme teknolojisi kullanılarak elde edilmiştir. Dizileme sonucunda toplamda 714 olgun miRNA tanımlanmıştır. Bu çalışma hem yeni miRNA'ların keşfedilmesi hem de normal ve kanserli örnekler arasındaki anlamlı farklılık bulunan genleri belirleme amacıyla gerçekleştirilmiştir (Witten ve ark., 2010).

Alzheimer: Bu verinin elde edilmiş olduğu çalışmada Alzheimer hastalığına erken evrede tanı koyabilmek amaçlanmıştır. Bu nedenle de kan örnekleri kullanılarak

miRNA'lar biyobelirteç olarak belirlenmeye çalışılmıştır. Çalışmada yer alan gözlemlerin dağılımı; 48 Alzheimer hastası ve 22 sağlıklı bireyler şeklindedir. miRNA'lar IlluminaHiSeq 2000 dizileme teknolojisi kullanılarak dizilenmiştir. Bu çalışma sonucunda 2801 satır ve 70 sütundan oluşan bir veri matrisi elde edilmiştir (Leidinger ve ark., 2013)

Fare Embriyo: Tek-hücreli RNA dizileri (scRNA-dizileri) ile memeli hücrelerinde dinamik, rastgele monoallel gen ifadesi açıklanması amaçlanmıştır. Bu çalışmada farklı dokular ve farklı gen ifade profillerine sahip ardışık embriyonik gelişim aşamalarındaki hücrelerin farklı transkriptlere sahip olacağı görülmüştür. Erken embriyonik dönemde fare embriyolarının farklı aşamalarından hücrelerin toplandığı bir veri seti kullanılmıştır. Allele özgü gen ifadesini araştırmak için farklı gelişim aşamalarından 268 hücre izole edilmiştir. Transkriptom profillerinin oluşturulması için her bağımsız hücre için Smart-dizileme ve Smart-dizileme2 kullanılmıştır. Araştırmacılar öncelikle kümeleme analizi yapıp 10 farklı kümenin oluştuğunu belirlemişlerdir. Bu çalışmadan elde edilen veri seti ile embriyonik aşamaya göre 10 gruba ayrılan veri ile sınıflandırma analizi gerçekleştirildi (Deng ve ark., 2014).

Zebra Balığı: Bu çalışmada döllenme sonrası sırasıyla 1 ay, 3 ay, 4 ay, 6 ay, 10 ay, 12 ay ve 14 ay olmak üzere 7 farklı yaş dönemindeki zebra balıklarından alınan örnekler çalışılmıştır. Hücrelerden scRNA-dizilemesi yapılarak beta hücreleri transkripsiyonel dinamiklerinin yaş değişkenine göre değişiminin belirlenmesi amaçlanmıştır. Bu 7 farklı yaş dönemi 3 grupta toplanmıştır: yavru (1 ay), genç (3,4 ve 6 ay) ve yetişkin (10,12 ve 14 ay). Hücreleri bu 3 grupta sınıflandırabilmek için çok terimli lojistik regresyon sınıflandırıcısı kullanılmıştır. Dizileme için Smart-Seq2 protokolü kullanılarak 637 beta-hücresi dizilenmiştir. Çalışma sonucunda 22431 satır ve 212 sütundan oluşan veri elde edilmiştir. Satırlarda genler sütunlarda gözlemler yer almaktadır (Singh ve ark., 2018).

Lemfoblastoid: Bu veri seti Montgomery ve ark.(2010) ve Pickrell ve ark. (2010) çalışmalarından alınarak birleştirilmiştir. Her iki çalışmada da genetik etkilerin ilk sorgulaması amacıyla yeni nesil dizileme teknolojilerini kullanarak transkriptom araştırması yapılmıştır. Montgomery ve ark. (2010) çift-sonlu Illımina kullanarak 37 baz-çifti (bp) 60 örnek örnekten lemfoblastoid hücre transkriptomunun mRNA'sı dizilenmiştir. Her örneğin transkriptomu Illumina GAII kullanılarak analiz edilmiştir.

Montgomery ve ark. (2010) çalışmada Avrupa popülasyonundan 60 birey alınırken, Pickrell ve ark. (2010) 69 adet Nijer birey ile lemfoblastoid hücresinden RNA dizilenmiş ve transkriptom düzeyindeki değişkenliği belirlemek amaçlanmıştır (Montgomery ve ark., 2010).

Fare Kök Hücre: Bu çalışmada kullanılan fare embriyonik kök hücreler ile her hücre tipine dönüşebilen durumlar arasındaki gen ifade heterojenliği özelliklerinin incelenmesi amaçlanmıştır. Fare embriyonik kök hücreleri; (1) serum+LIF'in 3 tekrarı, (2) 2i+LIF'in 4 tekrarı, (3) a2i+LIF'in 2 tekrarı olmak üzere 3 farklı koşul altında üretilmiştir. Illumina kütüphane hazırlığı için NexteraXTKit ve cDNA elde etmek için SMARTer Kit kullanarak ve Fluidigm C1 sistemi ile bu 3 koşul için toplamda 704 tek-hücre transkriptomu oluşturulmuştur. Okumaların %80'den fazlası musmuskulus genomu ile %60'ından fazlası ekzonla hizalanan bu çalışmada 704 sütun ve 38616 satırdan oluşan bir veri matrisi elde edilmiştir (Kolodziejczyk ve ark., 2015).

Alizadeh-V1 ve Alizadeh-V2: Bu çalışmada araştırmacılar en yaygın üç erişkin lenfoid malignitesinin gen ifade paternlerini karakterize etmek için mikrodizi kullanmıştır. Bu üç durum; yaygın büyük B hücreli lenfoma (DLBCL), foliküler lenfoma (FL) ve kronik lenfositik lösemi (KLL) şeklindedir. Ayrıca DLBCL1 ve DLBCL2 ile ifade edilen moleküler olarak farklı iki DLBCL formu tanımlanmıştır. Bu çalışmada 21 DLBCL1, 21 DLBCL2 olmak üzere toplamda 42 DLBCL, 9 FL ve 11 KLL olan kişiler kullanılmıştır. Alizadeh-V1 verisinde DLBCL alt tiplerinin belirlenebilmesi için DLBCL1 ve DLBCL2 olan gözlemler kullanılarak 1095 satır ve 42 sütundan oluşan veri seti kullanılmıştır. Alizadeh-V2 verisinde ise 42 DLBCL, 9 FL ve 11 KLL olan gözlemler kullanılarak 2093 satırdan, 62 sütundan oluşan veri seti kullanılmıştır (Alizadeh ve ark., 2000)**.**

Armstrong-V1 ve Armstrong-V2: Araştırmacılar yapmış oldukları bu çalışmada karışık kökenli lösemi genini içeren kromozomal translokasyonunu taşıyan AML'nin kötü prognoza sahip olduğunu belirtmişlerdir. Bireysel HOX geni ve seçili belirteçlerin ifadesine göre ALL, MLL, AML lösemi türleri oldukça farklı gen ifade profillerine sahip olduğu kümeleme algoritmaları kullanılarak gösterilmiştir. Affymetrix dizi tipi kullanılarak yapılan çalışmada 24 ALL, 48 MLL olan gözlem kullanılmıştır. Armstrong-V1 verisinde 2 grup şeklinde ALL ve MLL olan gözlemler kullanılırken, Armstrong-V2 verisinde 48 MLL olan gözlemler 20 MLL ve 28 AML olarak ayrılıp 24

ALL olan gözlemlerde dahil edilerek 3 gruba ayrılan verilerle analizler gerçekleştirilmiştir (Armstrong ve ark., 2001).

Bhattacharjee: Oligonükleotid mikrodizilerini kullanarak gerçekleştirilen bu çalışmada araştırmacılar 186 akciğer tümörü olan örnek ve 17 normal akciğer doku örneği (NL) olmak üzere 203 örnek ile 12.600 transkript dizisine karşılık gelen mRNA ifadesi seviyelerini analiz ettiler. Akciğer tümörleri arasında adenokarsinom (AD), küçük hücreli akciğer kanseri (SCLC), pulmoner karsinoidler (COID) ve yassı hücreli akciğer karsinomları (SQ) bulunmaktadır. Toplam 5 gruba ayrılan bu verinin örnek dağılımı 139 AD, 17 NL, 6 SCLC, 21 SQ ve 20 COID şeklindedir ve Affymetrix dizi tipi kullanılmıştır (Bhattacharjee ve ark., 2001).

Bittner: Bu çalışmada bir dizi örnekte gen ifade düzeyinin matematiksel analizi ile tanımlanan bir melanom alt kümesinin keşfedildiği bildirmektedir. Bu alt kümeleri belirlemek amacıyla 31 melanom ve 7 kontrol olmak üzere 38 örnek için ifade profili toplamışlardır. Bu örnekler kullanılarak hiyerarşik kümeleme analizi ile 31 melanom örneğinin 19'u çok iyi bir şekilde kümelenmiştir. Bu sıkı bir şekilde kümelenen örneklerin sınıfını ML1 ile belirtilmiştir. 7 kontrol dahil diğer melanom olan örnekler ML2 olarak adlandırılmıştır. Sonuç olarak 19 ML1, 19 ML2 olarak iki sınıf oluşturulmuş veri kullanılmıştır (Bittner ve ark., 2000).

Bredel: Yazarlar, 50 insan gliomunda gliomagenez ile ilişkili anahtar fonksiyonların ve yolakların analizine saf ağ bilgisini uyguladılar. Normal beyin ve glioma alt tiplerinin analizi için araştırmacılar tümörleri 31 saf glioblastoma (GBM) ve Oligodendroglial morfoloji (OG) için zenginleştirilmiş 14 tümörü gruplandırdı. Saf oligodendroglial ve zenginleştirilmiş oligoastroiktik tümörün birlikte gruplandırılması bu tümörlerin gözetimsiz, ortalama bağlantılı hiyerarşik kümeleme analizinde belirgin bir kümelenmesine dayanmaktadır. Analizler sonucunda 351 klonun gliomagenez ile önemli ölçüde ilişkili olduğu ortaya çıkmıştır. Normal beyine karşı tüm 50 glioma 31 GBM, 14 OG ve beş derece 1-3 astrositomlar (A) şeklinde 3 gruplu olarak sınıflandırma analizleri yapılmıştır (Bredel ve ark., 2005).

Chen: Hepatoselüler karsinom (HCC) dünya çapında önde gelen ölüm nedenleri arasında yer almaktadır. Bu çalışmada araştırmacılar HCC gen ifadesi örüntülerini karakterize etmek için cDNA mikrodizilerini kullanmışlardır. HCC olan örneklerin ifade düzeyi paternleri, tümörsüz karaciğer dokularında (LIVER) görülenlere kıyasla

tutarlı olarak farklılıklar bulunmuştur. Araştırmacılar tarafından farklı koşullar altında analiz edilen örnekler veri setinde sadece HCC ve LIVER ile ilgili durumlara göre ayrılmışlardır. 104 HCC 75 normal karaciğer doku örneği bulunmaktadır (Chen ve ark., 2002).

Chowdary: Bu çalışmada yazarlar, lenf nodu negatif meme tümörlerinden (B) ve Dukes'B kolon tümörlerinden (C) ani dondurulmuş ve RNAlater koruyucu süspansiyonlu doku çiftlerini karşılaştırmışlardır. Amaç B meme kanseri ile kolon kanserini ayırmaktır. Meme kanseri olan 62 kolon kanseri olan 42 örnek ile çalışılmıştır (Chowdary ve ark., 2006).

Dyrskjot: Mesane kanseri, sık sık yinelenen yaygın bir malign hastalıktır. Tanıda hastalığın evresi ve karsinomun varlığı bireyin hastalık seyrini belirlemede önemlidir. Bu bağlamda yazarlar, mesane karsinomunun klinik olarak anlamlı alt sınıflarının mikrodizi ifadesi analizi kullanılarak tanımlanmıştır. Örneklerin gruplara göre dağılımı 20 Ta, 11 T1 ve 9 T2 + şeklindedir. Çalışmada Affymetrix dizi tipi kullanılmıştır (Dyrskjøt ve ark., 2002).

Garber: 2001 yılında yapılan bu çalışmada 66 insan akciğer tümörü için global gen ifadesi profilleri cDNA mikrodizileri kullanılarak incelenmiştir. Tümörlerin gen ifadesi modellerine dayalı olarak alt bölümlere ayrılması, skuamöz (SCC), büyük hücreli (LCLC), küçük hücreli (SCLC) ve adenokarsinom (AC) olarak doğru bir şekilde sağlanmıştır. Veri seti oluşturulurken “normal” ve “fetal” akciğer dokuları olarak etiketlenmiş örnekleri çalışmaya dahil edilmemiştir (Garber ve ark., 2001).

Gordon: Akciğerin malign plevral mezotelyoma (MPM) ve adenokarsinomu (AD) arasındaki patolojik ayrım, süregelen yöntemler kullanılarak zayıf olabilir. Bu nedenle bu çalışmada az sayıda genin ifade seviyelerine dayanan basit bir tekniğin MPM ve akciğer kanserinin erken ve doğru tanısında yararlı olabileceğini önerilmiştir. Bu yöntem, gen ifadesi oranları ve rasyonel olarak seçilen kesim değerleri kullanılarak genetik olarak farklı dokular arasında doğru bir şekilde ayrım yapmak için tasarlanmıştır. Çalışmada 31 MPM, 150 AD örneği bulunmaktadır (Gordon ve ark., 2002).

Khan: Çalışmanın amacı, yapay sinir ağları (YSA) kullanarak gen ifadesi işaretlerine dayanarak kanserleri spesifik olarak tanı kategorilerine göre sınıflandırmak için bir yöntem geliştirmektir. Yazarlar yapay sinir ağları kullanarak; küçük, yuvarlak mavi hücreli tümörleri (SRBCT) model olarak kullanarak eğitmişlerdir. Bunlar nöroblastomlar (NB), Burkitt lenfoma (BL), rabdomiyosarkom (RMS) ve Ewing tümörler ailesi (EWS)'dir. Bu kanserler dört ayrı tanı kategorisine aittir ve klinik uygulamada sıklıkla tanı ikilemleri yaşanır. Orijinal veri setlerinden bahsedilen kanser türlerinden herhangi birine ait olmayan hücre hatları ve örnekler atılarak oluşturulan veride toplam 83 örnek bulunmaktadır (Khan ve ark., 2001).

Laiho: Tırtıklı kolorektal karsinomlar (CRC'ler) geleneksel CRC'lerden morfolojik olarak farklıdır ve farklı bir CRC oluşum yolunun izlemesi gerektiğini bu çalışmada öne sürülmüştür. Araştırmacılar, CRC gen ifade düzeyi profillerinin, tırtıklı ve konvansiyonel CRC olanların denetimsiz hiyerarşik kümeleme ile iki ayrı dala ayrıldığını göstermişlerdir. Çalışmada 8 tırtıklı CRC, 29 konvansiyonel CRC olan gözlem bulunmaktadır. Affymetrix dizi tipi kullanılan çalışmada toplam 2202 satır, 37 sütun bulunmaktadır (Laiho ve ark., 2007).

Lapointe-V1: Prostat kanseri nispeten yavaş olarak agresif metastatik hastalığa kadar geniş bir klinik davranış sergiler. Bu klinik heterojenitenin altında yatan potansiyel moleküler varyasyonu araştırmak için, yaklaşık 26.000 gen içeren cDNA mikrodizilerini kullanarak 62 primer prostat tümörünün yanı sıra 41 normal prostat örneğinde ve dokuz lenf nodu metastazında gen ifadesini belirledik. Denetimsiz hiyerarşik kümelenme, tümörleri normal örneklerden kolayca ayırdı ve ayrıca, farklı gen ifadesi kalıplarına dayanarak prostat tümörlerinin üç alt sınıfını (PT1, PT2 ve PT3) belirlemiştir. 11 PT1, 39 PT2 ve 19 PT3 örneği kullanılarak sınıflandırma analizi

yapılmıştır. Böylece yeni bir gözlemin hastalığın hangi evresinde olduğu tahmin edilir (Lapointe ve ark., 2004).

West: Yazarlar, bir dizi primer meme kanseri örneğinin DNA mikrodizi analizinden elde edilen gen ifade verilerine dayalı öngörme kabiliyeti sağlayan Bayes regresyon modelleri geliştirmiştir. Bu paternlerin göğüs tümörlerini östrojen reseptör durumuna ve ayrıca kategorize lenf nodu durumuna göre ayırt etme kapasitesine sahip olduğunu görmüşlerdir. Veriler östrojen-reseptör-pozitif (ER +) ve östrojen-reseptör-negatif (ER-) tümörlerden oluşmaktadır. Bu gruplarda sırayla 25 ve 24 örnek bulunmakta olup Affmetrix dizi tipi kullanılmıştır (West ve ark., 2001).

ST000369: Son yıllarda Ulusal Akciğer Kanseri Tarama Çalışması (NLST), düşük doz BT (LDCT) taramasının akciğer kanserine bağlı mortaliteyi %20 oranında azaltabildiğini göstermiştir. Fakat LDCT taraması, malign tümörlerin düşük prevelans oranları (% 2'den az) ve yüksek benign akciğer insidansı nedeniyle, özellikle yüksek riskli popülasyonlarda yüksek yanlış pozitif oranlara sahip olduğu bilinmektedir. Bu nedenle tanı kapasitelerini geliştirmek ve yanlış pozitif oranlarını azaltmak için LDCT taraması ile birlikte kullanılabilecek tamamlayıcı biyobelirteçler faydalı olacağı için bu çalışma tasarlanmıştır. Bu çalışmada iki grup olarak bağımsız vaka kontrol örneklerinin serum ve plazmaları ile çalışılmıştır. Metabolom analizi için gaz kromatografisi uçuş zamanlı kütle spektrometrisi (GC/TOF/MS) kullanılmıştır. Belli bir hedef olmaksızın bütün metabolitlerin tanımlanmaya çalışıldığı çalışmada hedefsiz metabolomik yaklaşım kullanılmıştır. Toplamda 43 adenokarsinomlu hastanın serum ve plazması ve 43 sağlıklı kontrolün serum ve plazması olmak üzere 172 örnek ile çalışılmıştır (Fahrmann ve ark., 2015).

ST000389: Son zamanlarda yapılan bilgisayarlı tomografi (BT) tarama çalışmaları akciğer kanserinin erken saptanmasında etkili ancak yüksek yanlış pozitif oranlara sahiptir. BT taramasını başka belirteçlerle de destekleyerek güçlendirmek ve yanlış pozitif oranları azaltmak için tasarlanmış ek olarak kan biyobelirteç testleri oldukça istenen bir durumdur. Bu çalışmada, benign ve malign olan bireylerden alınan serumdaki metabolitler gaz kromatografisi uçuş zamanlı kütle spektrometrisi (GC/TOF/MS) kullanılarak çalışılmıştır. Benign olan 29, malign olan 65 gözlem ile gerçekleştirilen bu çalışmada 989 metabolit bulunmaktadır (Fahrmann ve ark., 2016).

ST000388: ST000389 verisi ile bu veri aynı çalışmadan ve aynı örneklerden elde edilmiştir. ST000389 numaralı veri GC/TOF/MS ile elde edilirken ST000388 numaralı veri hidrofilik etkileşim sıvı kromatografisi-kuadratik kutuplu uçuş zamanlı kütle spektrometresi (HILIC/qTOF/MS) kullanılarak elde edilmiştir. Yine 94 örnek bulunan bu çalışmada HILIC/qTOF/MS teknolojisi ile 989 metabolit tespit edilmiştir (Fahrmann ve ark., 2016).

ST000390: Akciğer kanseri onlarca yıldır dünya genelinde kanser ölümlerinin önde gelen nedenleri arasındadır. Bu nedenle akciğer kanserinin erken tespiti için bu çalışmada belirleyici yeni metabolitlerin tespit edilmesi amaçlanmıştır. Sigarayla indüklenen akciğer kanserinin moleküler biyolojisi konusundaki bilginin son birkaç yılda çarpıcı bir şekilde arttığı göz önüne alındığında bu yaklaşım makul bir çözüm olarak düşünülmüştür. Hücresel metabolizmanın geniş kapsamı değerli tanısal biyobelirteçler sağlamaya ve potansiyel olarak tümör oluşumunun moleküler altyapısını tanımlamaya yardımcı olabileceği iyi bilinmektedir. Kütle spektrometrisindeki son gelişmeler ile birlikte çeşitli biyolojik örneklerdeki lipitlerin, karbonhidratların, aminoasitlerin ve nükleotitlerin kapsamlı metabolomik analizlerini gerçekleştirmek mümkün hale gelmiştir. Böylece yapılan bu çalışmada 39 malign ve 39 malign olmayan akciğer dokusu ile lipit, karbonhidrat, aminoasit, organik asit ve nükleotit metabolitlerini ölçmek için gaz kromatografisi uçuş zamanlı kütle spektrometrisi (GC/TOF/MS) kullanılmıştır. Kansere bağlı hücresel ve doku düzeyinde biyokimyasal değişikliklerin tanımlanması, dolaşımdaki biyobelirteçlerin tanımlanması ve adenokarsinom tümör oluşumunda yer alan biyokimyasal değişikliklerin belirlenmesi amacıyla bu çalışma gerçekleştirilmiştir (Wikoff ve ark., 2015).

ST000391: ST000390 numaralı çalışma ile aynı olup aynı gözlem ve örneklerden elde edilmiş bir veridir. Fakat bu verinin elde edilmesi için GC/TOF/MS değil HILIC/qTOF/MS teknolojisi kullanılmıştır (Wikoff ve ark., 2015).

ST000392: Akciğer kanserinin tespit edilmesi amacıyla gerçekleştirilmiş olan bu çalışmada araştırmacılar; tümör hücrelerindeki metabolik değişikliklerin tümör gelişimine konakçı yanıtının sistemik göstergeleri ile birleştiğinde tedavinin teşhisi ve izlenmesi için klinik fayda sağlayacağını düşünmüşlerdir. GC/TOF/MS kullanılarak elde edilen sonuçlar küçük hücreli olmayan akciğer kanseri (NSCLC) adenokarsinomu ve diğer akciğer kanseri vakalarından farklılığını göstermek için kullanılmıştır.

Çalışmada alınan kan örneklerinin metabolomik analizi sonucunda toplamı 437 metabolit elde edilmiş ve bunların 148'inin hangi bileşiğe karşılık geldiği tespit edilebilmiştir. 289 metabolit bilinemeyen bileşiklerdir ve analizlere dahil edilmemiştir. Metabolitlerin birçoğunun her iki grupta önemli ölçüde farklı olduğu bulunmuştur (Miyamoto ve ark., 2015).

ST000356: Bu çalışmada meme kanseri ve normal kontrol örneklerini profillemek için GC/MS ve LC/MS teknolojileri kullanılmıştır. Sınıflandırma analizleri gerçekleştirilirken meme kanseri aşamalarına göre 6 bilinmeyen, 50 2.aşama meme kanseri, 47 3. Aşama meme kanseri ve 31 sağlıklı kontrol örnekleri kullanılmıştır. Tespit edilen 101 metabolit ile analizler gerçekleştirilmiştir (Xie ve ark., 2015).

3.1. Omik verilerinin sınıflandırma için hazır hale getirilmesi

Omik veriler çeşitli teknolojiler kullanılarak elde edilir ve sınıflandırma analizine uygun bir veri elde edebilmek için çeşitli ön işleme basamaklarından geçer. Her omik biliminde ve ilgili omik bilimlerinde de kullanılan teknolojiye bağlı olarak bu basamaklar farklılık gösterebilmektedir. Aşağıda RNA dizileme, mikrodizi ve metabolomik verilerin sınıflandırmaya hazır hale getirilmesi süreci ifade edilmektedir.

3.1.1. RNA dizileme verilerinin sınıflandırma yöntemleri için hazırlanması

RNA dizileme de yeni nesil dizileme teknolojileri kullanılarak milyonlarca ham dizi okunur. Ham RNA dizileme verilerinin elde edilmesinden model oluşturulmasına kadar olan aşamalar aşağıda maddeler halinde gösterilmiştir;

1. Ham RNA dizilerinin elde edilmesi
2. Kalite değerlendirmesi (kesme, filtreleme vb.)
3. Referans genom ile eşleştirme
4. Eşleşme sayılarının elde edilmesiyle sayısal veri matrisinin oluşturulması
5. Verilerin normalleştirilmesi ve dönüşümü (deseq normalleştirmesi, vst dönüşümü)
6. Değişken seçimi
7. Sınıflandırma modellerinin oluşturulması
8. Oluşturulan modelin geçerliliğinin değerlendirilmesi

Çeşitli araçlarla ham veri elde edildikten sonra ilk olarak dizilenmiş verilerin kalitesi değerlendirilir. Düşük kalite dizilerin silinmesi ve düşük kaliteli okumaların elimine

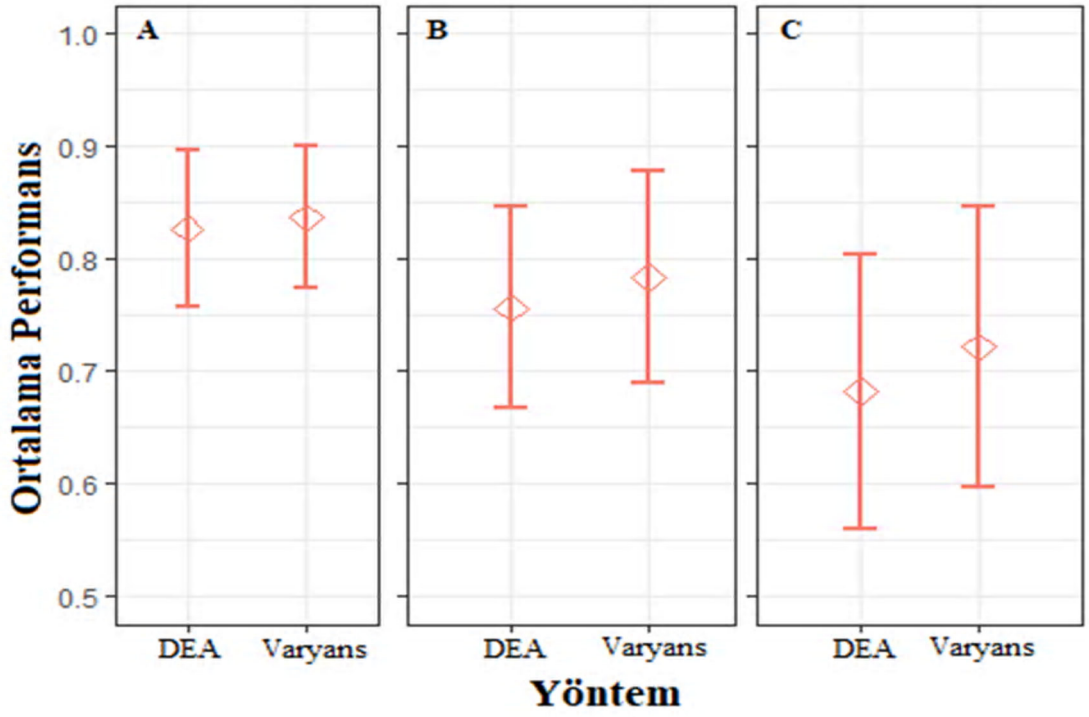
edilmesi gibi işlemler bu basamakta gerçekleştirilir. Filtreleme aşamasından sonra kalan veri referans genoma ya da transkriptoma göre hizalanır. Referans ile eşleşen okumaların sayısı sonraki adım için elde etmiş olduğumuz veridir. Bu veri ile doğrudan sınıflandırma analizi yapılmaz. Öncelikle veri normalleştirilmesi ve dönüşüm uygulanır. Sınıflandırma için en önemli olabilecek değişkenlerde seçildikten sonra sınıflandırma modeli oluşturulur (Zararsız ve ark., 2017).

Bu çalışmada çeşitli çalışmalardan elde etmiş olduğumuz RNA dizileme verilerine sınıflandırma yapmadan önce çeşitli işlemler uygulandı. İlk olarak elde etmiş olduğumuz nicel kesikli tipte değişkenlerden oluşan veri eğitim ve test verisi olarak ayrıldı. Ayırma oranı %70 eğitim %30 test verisi olacak şekilde yapıldı. Bu ayırmayı tamamen rastgele olarak yaparsak grup başına düşen gözlem sayısı eğitim ve test verilerinde dengesiz olabilir. Özellikle grup dağılımının dengesizlik gösterdiği durumlarda bir gruptan çok diğer gruptan daha az gözlem seçimi ile eğitim ve test verisi oluşturabiliriz. Bu durum sonuçları değerlendirirken problemlere neden olabilir. Bu nedenle veriler eğitim ve test seti olarak ayrılırken her gruptan grup başına düşen gözlem sayısı ile orantılı olacak şekilde ayırma yapıldı. Daha sonra eğitim verisini kullanarak ön filtreleme ve normalleştirme işlemleri gerçekleştirildi. Filtreleme için öncelikle sıfıra yakın varyans yöntemi kullanıldı. R programlama dilinde geliştirilmiş olan caret kütüphanesinin *nearZeroVar()* fonksiyonu kullanılarak bu yöntem uygulandı. Bu işlemle tek bir değere sahip varyansı sıfır olan ya da örnek sayısına göre çok az benzersiz değerlere sahip ve en yaygın değerin ikinci en yaygın değerin frekansına oranının büyük olduğu durumlar tespit edildi. Sonuç olarak varyansı sıfır ya da sıfıra yakın değerler alan değişkenler filtrelendi. Bu işlem ile eğitim verisinden çıkarılan değişkenler test verisinden de çıkarıldı.

Filtrelemeden sonra kalan eğitim verilerine normalleştirme ve dönüşüm işlemleri yapıldı. Normalleştirme işleminin amacı değişkenler arasında örnekten örneğe değişen farklılıkları ortadan kaldırarak, sınıflandırma analizi yaparken değişkenlerin değeri yüksek olanların düşük olanları baskılaması sonucunda oluşabilecek yanlı sonuçların önüne geçmektir. Böylece normalleştirme örnekler arasındaki karşılaştırmaları daha sağlıklı yapabilmemizi sağlar. Normalleştirme yöntemlerinden deseq yöntemi kullanılarak veriler normalleştirildi. Aynı normalleştirme tekniğine bağlı kalarak test verileri de normalleştirildi. Normalleştirmenin ardından elde edilen yeni veri değerleri

kullanılarak dönüşüm yapıldı. Dönüşümün amacı verideki aşırı yaygınlıkları önlemek ve ortalama-varyans ilişkisini daha iyi kestirebilmektir. Dönüşüm tekniği olarak vst (varyans-dengeleyici dönüşüm) yöntemi kullanıldı (Anders, 2010).

Elde edilen veri ile doğrudan analiz yapmak yerine değişkenler için filtreleme yapıldı. Bunun nedeni; omik teknolojiler kullanılarak elde edilen verilerde değişken sayıları genellikle gözlem sayısından oldukça fazla olmasına rağmen değişkenlerin tamamının sınıflandırma için önem arz etmemesidir. Bu aşamada kullanılan birçok yöntem bulunmaktadır. Maksimum varyans filtrelemesi ve anlamlılık analizi (DEA) sıklıkla tercih edilen yöntemler arasındadır. Bu nedenle vst dönüşümü sonrasında bu yöntemlerin performansını değerlendirmek için denemeler yapıldı.

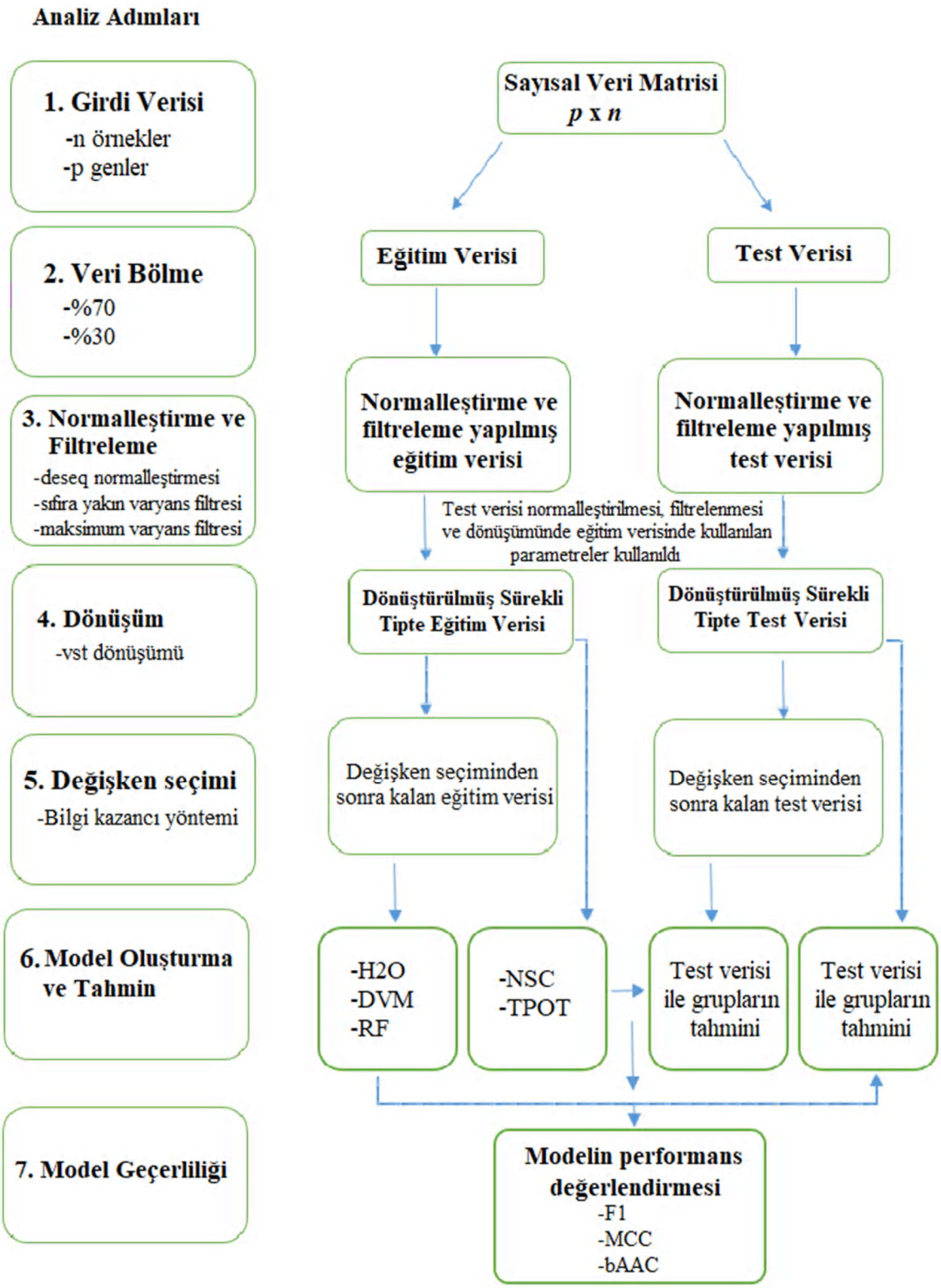


Şekil 3.1. Alzheimer verisinde filtreleme yöntemlerinin karşılaştırılması

Şekil 3.1'de bu denemelerin sonucunda elde edilen grafik gösterilmektedir. Alzheimer verisi kullanılarak oluşturulan bu grafikte A, B, C harfleri sırayla bACC, F1 ve MCC metriklerine karşılık gelmektedir. Verilerden 10 kez farklı örnekler seçilerek her iki yönteme göre filtreleme yapıldıktan sonra sınıflandırma performansları elde edildi. Grafik 10 farklı örnek grubuna ait sonuçların ortalama ± standart hatası kullanılarak oluşturuldu. Benzer bir analizde Lemfoblastoid verisi kullanılarak gerçekleştirilmiş olup Alzheimer veri ile benzer sonuçlar elde edildi. Her iki veri seti içinde maksimum

varyans filtrelemesi ile daha yüksek performans elde edildiği görüldü. Bu nedenle sonraki analizlerde maksimum varyansa göre filtreleme işlemleri gerçekleştirildi.

Bu yöntemin uygulaması esnasında her genin ayrı ayrı ortalama ve standart sapmaları belirlenip standart sapmanın ortalamaya bölünmesiyle değişim katsayısı hesaplandı. Genler değişim katsayısına göre büyükten küçüğe doğru sıralanıp gen sayısı 2000'den fazla olan verilerin ilk 2000 geni alındı. Gen sayısı 2000'den düşük olan verilere filtreleme uygulanmadı. Bu şekilde elde edilen veriler ile NSC ve TPOT yöntemleri için sınıflandırma aşamasına geçildi. Bilgi kazancı metodu ile verilerden sınıflandırmaya en çok katkı sağlayacak olan genlerin seçimi yapıldı, bu seçilmiş genlerden oluşan veri ile RF, DVM ve H2O yöntemleri ile sınıflandırma aşamasına geçildi. Eğitim verileri için uygulanan her işlem eğitim verisinden bağımsız olarak elde edildiği varsayılan test verisine de uygulandı. Uygulanan aşamaların şeması şekil 3.2'de gösterilmektedir.



Şekil 3.2. RNA dizileme verilerinin analiz aşamaları (Goksuluk ve ark., 2019)

Şekil 3.2'de de belirtilmiş olduğu gibi normalleştirme, filtreleme ve dönüşüm eğitim verisinde kullanılan parametrelere bağlı olarak gerçekleştirdi. Değişken seçimi eğitim verisi ile yapıldıktan sonra seçilen genler test verisinden de seçilerek hem test hem eğitim verisinden değişken seçimi yapıldı. Eğitim verisinde bulunan değişkenlerin tamamı test verisinde de bulmaktadır.

3.1.2. Mikrodizi verilerinin sınıflandırma yöntemleri için hazırlanması

Kullanılan 16 mikrodizi verisi iki farklı mikrodizi teknolojisi kullanılarak elde edilmiştir. Armstrong-V1, Armstrong-V2, Bhattacharjee, Chowdary, Dyrskjot, Gordon, Laiho ve West verileri tek kanallı Affymetrix çip teknolojisi kullanılarak, Bittner, Bredel, Chen, Garber, Khan, Lapointe-V1, Alizadeh-V1, Alizadeh-V2 verileri ise çift kanallı cDNA teknolojisi kullanılarak elde edilmiştir (de Souto ve ark., 2008). Yapılmış olan çalışmalardan elde edilen sayısal veriler çeşitli aşamalardan, ön işlemlerden geçirildi. İlk olarak elde etmiş olduğumuz veri %70 eğitim ve %30 test verisi olarak ayrıldı. Eğitim ve test verilerine ayırma işlemi dengesiz grup dağılımına sahip verilerde her grup için ayrı ayrı %70'e %30 olarak ayrıldı. Sınıflandırma analizi yapmadan önce bu verilerden nicel kesikli değişkene sahip olanlar için her değere 1 eklenip 10 tabanında logaritmik dönüşüm yapıldı, sürekli yapıda olanlara müdahele de bulunulmadı. Daha sonra yukarıda RNA dizileme verilerinde bahsetmiş olduğumuz şekilde filtreleme uygulandı. Yani gen sayısı 2000'den fazla olan verilerin ilk 2000 geni değişim katsayısına göre büyükten küçüğe doğru sıralanıp seçildi. Maksimum 2000 değişkeni bulunan mikrodizi verilerine içsel olarak değişken seçimini yapan NSC, TPOT kullanarak sınıflandırma analizi uygulandı. Değişken seçimini içsel olarak yapmayan yöntemlerden RF, DVM ve H2O uygulanması için önce bilgi kazancı yöntemi ile değişken seçimi yapıldı daha sonra bu yöntemlerin uygulanmasıyla sınıflandırma analizlerine geçildi.

3.1.3. Metabolomik verilerin sınıflandırma yöntemleri için hazırlanması

Metabolomik veriler https://www.metabolomicsworkbench.org/ sitesinden elde edildi ve veriler isimlendirilirken bu sitede bulunan çalışma kimlik numaraları kullanıldı. Metabolomik alanında elde kullanmış olduğumuz veriler gaz kromatografi – kuadratik kutuplu uçuş zamanlı kütle spektrometresi (GC/QTOF/MS), gaz kromatografi – uçuş zamanlı kütle spektrometresi (GC/TOF/MS), hidrofilik etkileşim sıvı kromatografisi-kuadratik kutuplu uçuş zamanlı kütle spektrometresi (HILIC/QTOF/MS) gibi sıklıkla

kullanılmakta olan teknolojilerden elde edilmiştir. Metabolit düzeylerini içeren veriler diğer omik verilerinde de yapmış olduğumuz gibi öncelikle gruplara düşen gözlem sayılarına göre dengeli olarak eğitim ve test verisine ayrıldı.

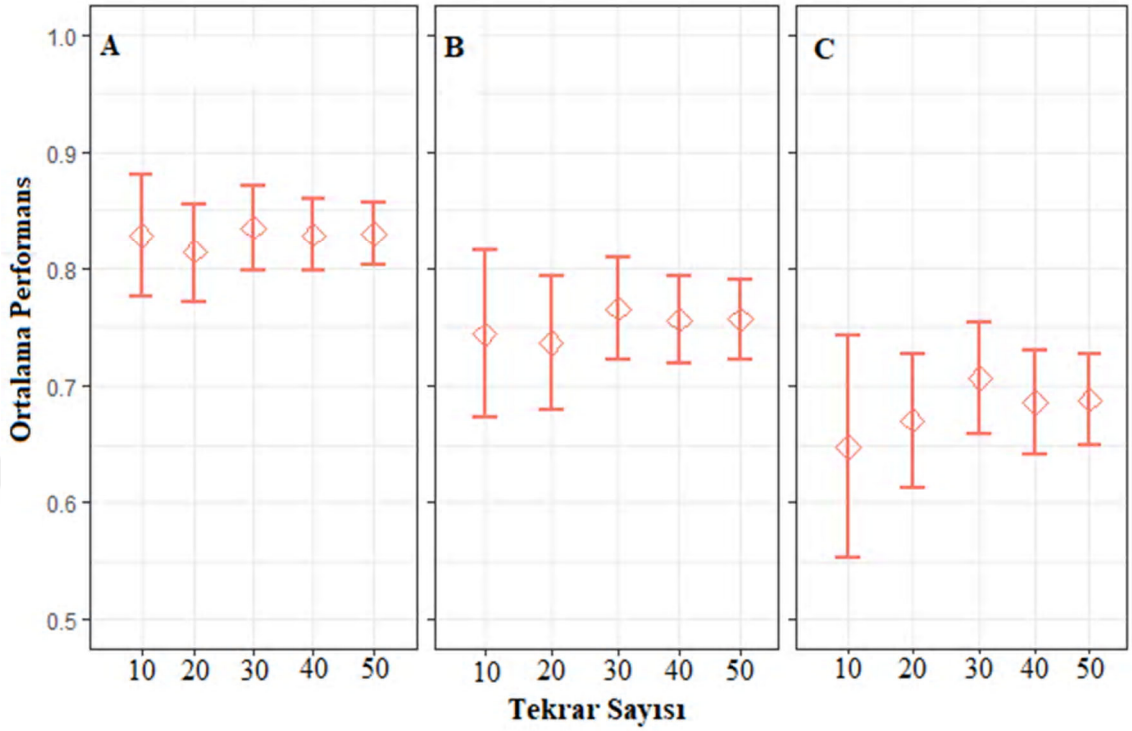
Daha sonra metabolit yoğunluğu her bir örneğin metabolit miktarının toplamına bölünerek normalleştirildi. Bazı metabolitler canlının sistemi gereği çok bol bulunurken bazıları nadir bulunur. Çok bulunan metabolitin düzeyi 1000'ler ile ölçülebilirken az bulunan metabolitlerin düzeyi 10'lar ile ölçülebilir. Böyle bir durumda bir metabolitin düzeyinde meydana gelen 5 birimlik fark klinik olarak anlamlı olabilirken başka bir metabolitin düzeyinde meydana gelen 800 birimlik fark anlamlı olmayabilir. Bu gibi durumlarda çok bulunan metabolit az bulunan metabolitin analiz sonucunu baskılayarak bu değişimi önemsiz olarak gösterebilir. Bu şekilde yanlı sonuçlar elde etmenin önüne geçmek için metabolit düzeylerinin normalleştirilmesi yapılarak analizler gerçekleştirildi. Böylelikle metabolit düzeyleri arasındaki farklılıkların analiz sonuçlarını etkilemesi önlenmiş oldu. Daha sonra veri setinde bulunan her bir değer 1 milyon ile çarpılarak 10 tabanında logaritmik dönüşüm uygulandı. Bu şekilde elde edilen yeni verilerden metabolit sayısı 2000'den fazla olanlar için RNA dizileme verilerinde bahsetmiş olduğumuz şekilde filtreleme uygulandı. Yani metabolitlerin değişim katsayısına göre büyükten küçüğe doğru sıralandı. Daha sonra metabolit sayısı 2000'den fazla olan verilerin değişim katsayısı en büyük olan ilk 2000 metaboliti alındı. Eğitim verisinde gerçekleştirilen bu aşamaların tamamı test verisinde de gerçekleştirildi.

3.2. Omik Verilerin Sınıflandırılması

Omik teknolojiler kullanılarak elde edilen veriler çeşitli ön işleme aşamalarından geçirildi. Ham veriler elde edilen çalışmalarda çalışmacılar tarafından çeşitli ön işlemlerden geçmiştir (farklı cihazlardan elde edilen verilerin sayısal veri matrisi aşamasına getirilmesine kadar olan aşamalar). Sayısal veri matrisi olarak elde etmiş olduğumuz veriler için verilerin eğitim ve test seti olacak şekilde bölünmesi, filtreleme, normalleştirme, dönüşüm gibi işlemler uygulanarak sınıflandırma için hazır hale getirildi.

Bu çalışmada kullanılan her bir veriden eğitim ve test verileri 20 kez rastgele seçildi. Yani her veri için yapılan tüm ön işleme ve sınıflandırma işlemleri 20 kez yapıldı. Verilerin tekrarlı olarak analiz edilmesinin nedeni elde edilen sonuçların kesinliğini sağlamaktır. Tekrar sayısı arttıkça elde edilen standart hatalar kullanılarak oluşturulan

güven aralığı azalır ve daha kesin sonuçlar elde edilir.



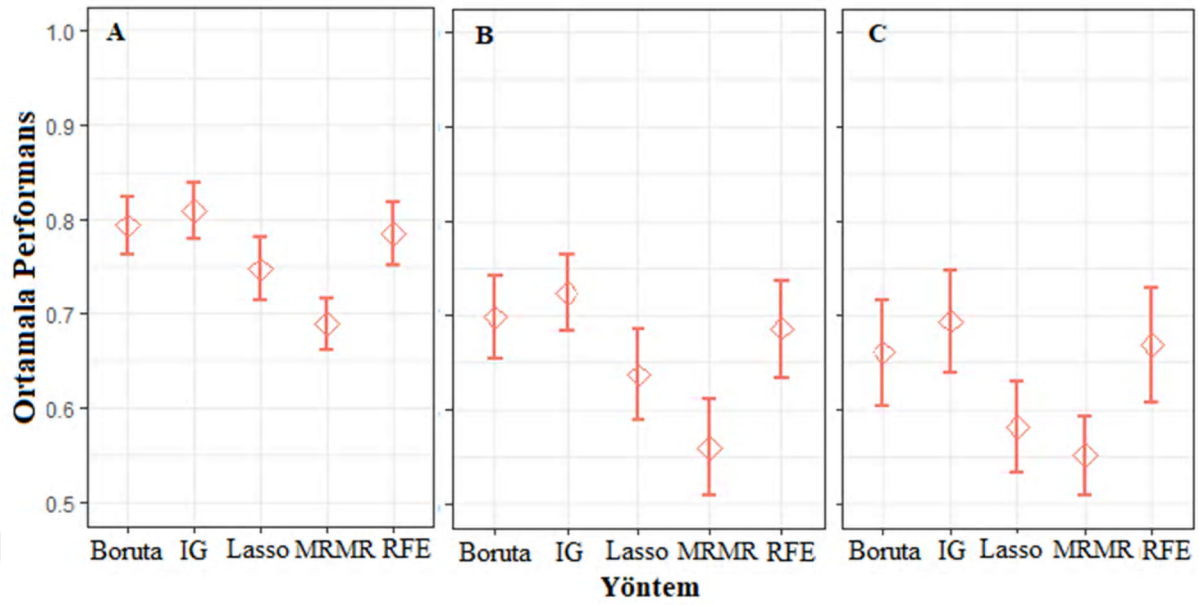
Şekil 3. 3. Tekrar sayılarına göre standart hata değişimi

Tekrar sayısı arttıkça elde edilen standart hata azalsa da verilerin analiz süresi ve iş yükü artar. Bu nedenle optimum tekrar sayısını belirlemek gerekir. Optimum tekrar sayısını belirlemek için Alzheimer verisi kullanılarak 10, 20, 30, 40 ve 50 tekrar ile verilerin performansı ortalama ± 2*standart hata olarak değerlendirildi. Şekil 3.3'de değerlendirme sonuçları grafik ile gösterildi. Bu grafik DVM yöntemi kullanılarak çizdirildi. Grafikte sırayla A,B,C bACC, F1 ve MCC metriklerine göre elde edilen sonuçlara karşılık gelmektedir. Her bir metrik için 10 tekrar için ± 2*standart hata aralığı oldukça yüksek iken 20 ve üzeri tekrar için standart hata aralığı giderek azalmaktadır. Tekrar sayısı 10'dan 20'ye çıktığında standart hata aralığı fazlaca düşmesine rağmen 20'den 30, 40 ve 50'ye çıkarıldığında bu denli bir düşme olmadı. Analizlerin maliyeti de göz önüne alınarak optimum tekrar sayısı 20 olarak belirlendi. Bütün verilerin analizi 20 kez farklı örnek grupları çekilerek tekrar tekrar analiz edildi.

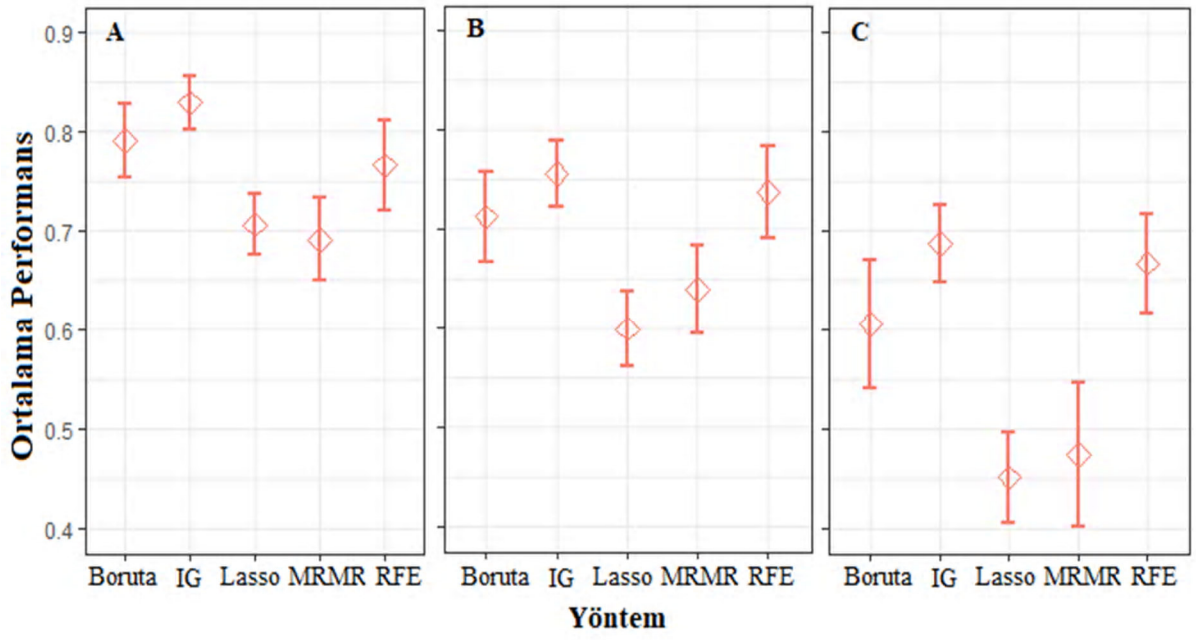
Bu çalışmada yer alan bütün sınıflandırma analizleri TÜBİTAK ULAKBİM, Yüksek Başarım ve Grid Hesaplama Merkezinde (TRUBA kaynaklarında) gerçekleştirildi. TRUBA kuyruk sistemlerinde Centos Enterprise 7.3 Linux işletim sistemi kuruludur.

Kullanıcı arayüzü sunucularına SSH ile bağlanarak hesaplama kümelerine iş gönderildi. TRUBA sisteminde verimli kaynak kullanımı sağlamak için hesaplama sunucuları SLURM kaynak yöneticisi ile hesaplama kuyruklarında organize edilip kullanılır. Analizleri TRUBA hesaplama kuyruklarında çalıştırmak için çekirdek sayısı, sunucu sayısı, işin gönderileceği kuyruk, kullanılacak modül gibi özelliklerin belirtildiği SLURM betik dosyaları kullanıldı. RF, NSC, DVM ve H2O analizleri R 3.4.3 sürümünde TPOT analizleri Python 3.6.5 sürümü kullanılarak gerçekleştirildi.

RF, DVM ve H2O analizlerine başlamadan önce değişken seçimi yapıldı. NSC yöntemi çalışma prensibi gereği kendi önemli bulduğu değişkenler ile analiz yaptığı için NSC yönteminde kullanılan verilerde değişken seçimi yapılmadı. TPOT yöntemi ise değişken seçimini kendi otomatik yaptığı için bu yöntemde de değişken seçimi yapılmadan analizler gerçekleştirildi. Değişken seçimi için 3 farklı yaklaşım ve bu yaklaşımlar içinde farklı yöntemler bulunmaktadır. Literatürde omik verilerde kullanılmakta olan birçok değişken seçimi yöntemi bulunmaktadır. Bu yöntemlerin performansı veriden veriye farklılık gösterir. İyi sonuçlar verdiği bilinen belli başlı yöntemler arasında boruta, bilgi kazancı (IG), lasso, minimum kalabalık maksimum ilgililik (Minimum Redundancy Maximal Relevancy, MRMR) ve yinelemeli değişken seçimi (RFE) yer almaktadır. Bu yöntemlerden hangisi ile daha iyi sonuçlar alınabileceğini değerlendirmek için bazı verilerde denemeler yapıldı. Yapılan denemelerde 50 tekrar yapıldı için elde edilen sonuçlardan performans ortalaması ± 2*standart hata kullanılarak grafikler elde edildi. Şekil 3.4 ve 3.5'de RF ve DVM yöntemleri kullanılarak bACC, F1 ve MCC metriklerine göre elde edilen sonuçlar görülmektedir. Grafiklerde A, B, C sırasıyla bACC, F1 ve MCC metriklerine karşılık gelmektedir. Grafik ile gösterilen sonuçlarda da görüldüğü gibi IG metodunda daha yüksek performansa sahip sonuçlar elde edildi. Bu nedenle RF, DVM ve H2O yöntemleri ile analizler yapılmadan önce IG yöntemi ile değişken seçimi yapıldı, NSC, TPOT yöntemlerinin analizi ise verilerin tamamı kullanılarak gerçekleştirildi.



Şekil 3.4. RF yöntemine göre değişken seçme yöntemlerinin performansı



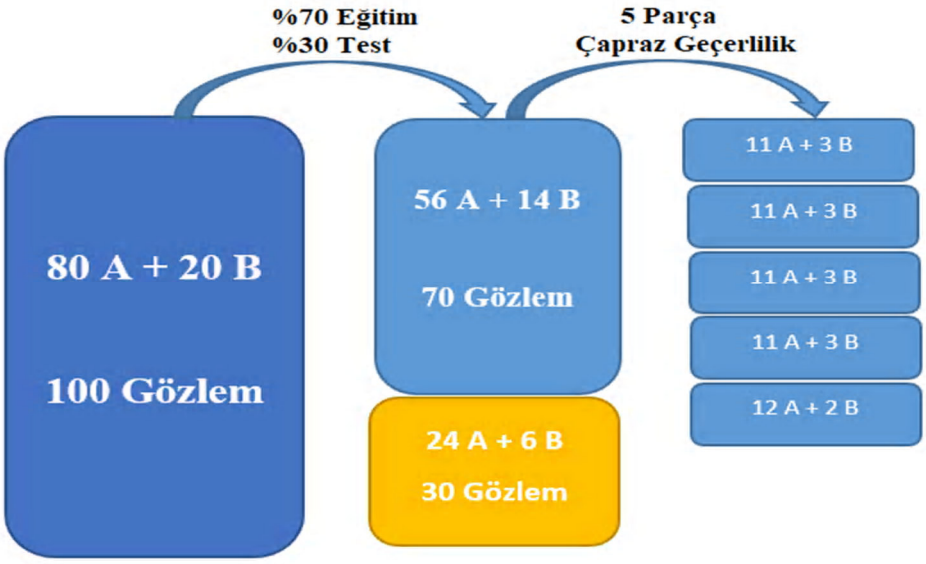
Şekil 3.5. DVM yöntemine göre değişken seçme yöntemlerinin performansı

RF, DVM ve NSC yöntem analizleri R CRAN ağında bulunan caret kütüphanesinin 6.0.84 versiyonu kullanılarak gerçekleştirildi.

```
1 rpt <- 10
2 fold <- 5
3 train.set <- unlist(lapply(1:rpt, function(x){
4   createFolds(y = cond.tr, k = fold, returnTrain = TRUE)
5   }), recursive = FALSE)
6
7 Model <- train(x = tr.data, y = cond.tr, method = "pam", tuneLength = 10,
8                 trControl = trainControl(method = "repeatedcv", number = fold,
9                 repeats = rpt, index = train.set))
10 pred <- predict(Model, ts.data)
```

Şekil 3.6. NSC yöntem kodları

NSC yöntemi uygulanırken yukarıda yer alan kod satırlarından faydalanıldı. Kullanılan kütüphanelerin çağırılması ve verilerin yüklenmesinden sonra değeri 10 olan rpt, 5 olan fold değişkenleri oluşturuldu. fold değişkeni kaç parça çapraz geçerlilik yapılacağını, rpt değişkeni ise bu çapraz geçerlilik yönteminin kaç kez yapıldığını belirtmektedir. Daha sonra 3.-5. satır aralığındaki kodlar kullanılarak daha öncesinde grup başına düşen gözlem sayıları ile orantılı olarak ayrılan eğitim ve test verisinde olduğu gibi, veriyi 5 parçaya ayırırken de yine grup başına düşen gözlem sayıları ile orantılı olarak ayırma işlemi sağlandı. Böylece eğitim verisi ile model eğitilirken gruplardan gelen gözlemler rastgele fakat belirli bir orana göre elde edilmiş oldu. Örneğin iki gruplu bir verinin A grubunda 80, B grubunda 20 gözlem olduğunu var sayarsak bu veriyi Şekil 3.7'deki gibi parçalara ayırırız.



Şekil 3.7. Verilerin test, eğitim ve çapraz geçerlilik için bölünmesi işlemi

Verileri bu şekilde ayırdığımızda bölünen veri parçalarına giden gözlemlerin tek bir gruptan seçilmesi gibi istenmeyen yanlı sonuçlara neden olabilecek durumların önüne geçilmiş olur. Bu iki aşamalı veri bölme işleminde; ilk aşamada model performansı ikinci aşamada ise eğitim verisi ile oluşturulan modelin hiperparametreleri değerlendirildi.

Şekil 3.8'de yer alan 7.- 9. satır aralığı ise modelin oluşturulduğu aşamadır. Model oluşturmak için *train* fonksiyonu kullanılarak çeşitli model parametreleri bu fonksiyon içinde belirtilir. Bu fonksiyonun ilk parametresi tr.data'dır. tr.data sayısal veri matrisidir. cond.tr ise her bir gözlemin sınıfını bulunduran grup değişkenidir. Method parametresi ile hangi sınıflandırma yöntemini kullanacağımızı seçeriz. (https://topepo.github.io/caret/). Son satırda yer alan predict fonksiyonu ise oluşturulan model ile test verisini kullanarak nihai model performansını elde etmemizi sağladı. Daha sonra çeşitli fonksiyonlar ve hesaplamalar yardımıyla model performansının değerlendirileceği metrikler hesaplandı.

H2O ile analizler paket versiyonu 3.24.0.5 kullanılarak gerçekleştirilmiştir. Aşağıdaki şekil 3.8'de otomatik H2O modeli oluşturmak ve model performansı elde etmek için yazılmış kod satırları gösterilmektedir. *h2o.automl* fonksiyonu kullanılarak sayısal veri matrisi, grup değişkeni ve maksimum çalışma süresinden oluşan parametreler ayarlandı.

H2O ile oluşturulan modellerin farklı çalışma süreleri içinde karşılaştırılması için *max_runtime_secs* parametresi 3600, 7200 ve 10800 saniye olarak ayarlandı.

```
Automated_Model <- h2o.automl(y = "cond.tr", training_frame=tr_dataH2o,
                              max_runtime_secs=3600)
performans = h2o.performance(Automated_Model@leader, ts_dataH2o)
```

Şekil 3.8 H2O yöntem kodları

Modeller otomatik olarak denendikten sonra en iyi performansı gösteren model leader objesinde saklanır. Bir sonraki aşamada en iyi performansa sahip lider model ile test verisi kullanılarak lider modelin genellenebilir bir model olup olmadığı değerlendirildi.

Bir diğer otomatik makine öğrenmesi yöntemlerinden TPOT yöntemi şekil 3.9 da görülmekte olan kod satırları kullanılarak gerçekleştirildi. Analizler esnasında kullanılacak olan kütüphaneler ve işlem görecek olan veri yüklendikten sonra *TPOTClassifier()* fonksiyonu kullanılarak analizler için ilgili parametreler belirlendi. Daha sonra 3.satırda yer alan kod ile belirlenen parametreler doğrultusunda eğitim verisi ile model eğitim süreci başlatıldı. Daha sonra oluşturulan model test verisi ile denendi ve sınıflandırma performansı elde edildi.

```
tpot = TPOTClassifier(verbosity=2, generations=5, population_size=20, cv=5,
                      random_state=42)
    tpot.fit(InputTrain, condTr)
    print(tpot.score(InputTest, condTs), flush = True)
    tpot.score(InputTest, condTs)
    y_predictions = tpot.predict(InputTest)
    conf_mat = sklearn.metrics.confusion_matrix(y_true=condTs,
                                                y_pred=y_predictions)
```

Şekil 3.9. TPOT yöntemi kodları

3.3. Sınıflandırma Performansının Değerlendirilmesi

Makine öğrenmesinde kullanılan sınıflandırma modellerinin performansını değerlendirmek için hedef gruba ilişkin tahminlerin ve gerçek değerlerin karşılaştırılmasında hata matrisi sıklıkla kullanılmaktadır. Sınıflandırma performansını değerlendirmek için çok sayıda metrik bulunmaktadır ve bu metriklerin birçoğunun hesaplanmasında hata matrisinde yer alan 4 bileşenden faydalanılır (Zararsız, 2015). Bu bileşenler;

DP: Doğru pozitif gözlem sayısı, gerçekte pozitif olan durumu sınıflandırma sonucunda da pozitif olarak tahmin etmek

DN: Doğru negatif sayısı, gerçekte negatif olan durumu sınıflandırma sonucunda da negatif olarak tahmi etmek

YP: Yanlış pozitif sayısı, gerçekte negatif olan durumu sınıflandırma sonucunda pozitif olarak tahmin etmek

YN: Yanlış negatif sayısı, gerçekte pozitif olan durumu sınıflandırma sonucunda negatif olarak tahmin etme şeklindedir.

Bu bileşenleri hata matrisindeki yeri tablo 3.2'de gösterilmektedir.

Tablo 3.2. İki sınıflı veriler için hata matrisi

Sınıflandırma ile tahmin edilen durum	**Gerçek durum**		**Toplam**
	Pozitif	**Negatif**	
Pozitif	DP	YP	DP+ YP
Negatif	YN	DN	YN+ DN
Toplam	DP+ YN	YP+ DN	n

n: toplam gözlem sayısı

Uygulanan yöntemlerin performans değerlendirmesi için iki gruplu verilerde ağırlıklandırılmış dengeli doğruluk oranı (bACC), Matthew korelasyon katsayısı (MCC) ve ağırlıklandırılmış F1 metrikleri kullanıldı. Çok sınıflı veriler için bahsetmiş olduğumuz metriklere ek olarak her bir sınıfın F1 ve bACC değerleri de hesaplandı. Bu metrikler aşağıdaki gibi hesaplanmaktadır;

Matthew korelasyon katsayısı:

$$\text{MCC} = \frac{DP*DN - YP*YN}{\sqrt{(DP+YP)(DP+YN)(DN+YP)(DN+YN)}} \tag{3.1}$$

formülü ile hesaplanır.

F1 metriğinin formülü:

$$\text{F1} = 2 * \frac{Kesinlik * Duyarlılık}{Kesinlik + Duyarlılık} \tag{3.2}$$

şeklinde hesaplanır. Formülünden de anlaşıldığı gibi F1 kesinlik ve duyarlılığın harmonik ortalamasıdır. F1 formülünde yer alan kesinlik

$$Kesinlik = \frac{DP}{DP+YP} \tag{3.3}$$

formülü ile doğruluk ise

$$Duyarlılık = \frac{DP}{DP+YN} \tag{3.4}$$

formülü ile elde edilir.

Dengeli doğruluk oranı hesaplanırken aşağıda yer alan formül kullanılır:

$$\text{bACC} = \frac{Duyarlılık + Seçicilik}{2} \tag{3.5}$$

Bu formülde duyarlılığı hesaplamak için denklem (3.4) kullanılır. Seçicilik ise aşağıdaki gibi hesaplanır:

$$Seçicilik = \frac{DN}{DN+YP} \tag{3.6}$$

F1 ve bACC hesaplanırken gruplara göre ağırlıklandırma yapıldı. Özellikle dengesiz sınıf dağılımına sahip olan verilerde ağırlıklandırma yapılan ve yapılmayan metriklerin değerleri farklılık gösterir. Grup başına düşen gözlem sayılarında dengesizlik yoksa ağırlıklandırılmış ve normal metrikler arasında herhangi fark oluşmayacağı için metrikler ağırlıklandırma yapılarak hesaplandı.

Sınıflandırma yöntemlerini karşılaştırırken bu metriklere ek olarak sparslık derecesi (en az değişkenli olarak en iyi performansın ölçütü) ve hesaplama maliyeti de kullanılmıştır. Hesaplama maliyeti tek bir veri ile sınıflandırma analizlerinin sonlanma süresi olarak hesaplandı. Sparslık ölçütü ise sınıflandırma esnasında her modelin 20 tekrar için kullanmış olduğu değişken sayılarının ortalama ve standart hatası kullanılarak değerlendirildi. Elde edilen analiz sonuçları bulgular bölümünde ifade edilmiştir.

4. BULGULAR

Her veri için RF, DVM, NSC makine öğrenmesi sınıflandırma metotları ve H2O, TPOT otomatik makine öğrenmesi metodları uygulanmıştır. H2O için analizler 1, 2, 3 saat için denenmiş ve sınıflandırma performansları kaydedilmiştir. Sınıflandırma performansını değerlendirmek için MCC, bACC, F1 ve ağırlıklandırılmış bACC, F1 metrikleri kullanılarak sonuçlar raporlanmıştır. Çok sınıflı verilerde herbir grup için F1 ve bACC metrikleri ifade edilmiştir. Her veride 20 kez sınıflandırma yapılmış olduğu için elde edilen sonuçlar performansın ortalama ve standart hatası kullanılarak özetlenmiştir.

Alizadeh-V1 verisi için 20 kez gerçekleştirilen sınıflandırma analizleri sonucunda elde edilen F1 metriğine göre performanslar aşağıda yer alan tablo 4.1'de görüldüğü gibidir.

Tablo 4.1. Alizadeh-V1 verisi sınıflandırma sonucu F1 metrikleri

Çalışma sayısı	Otomatik Makine Öğrenmesi				Standart Makine Öğrenmesi		
	H2O (1 saat)	H2O (2 saat)	H2O (3 saat)	TPOT	NSC	RF	DVM
1	1.000	1.000	1.000	0.833	1.000	1.000	1.000
2	1.000	0.909	1.000	0.909	1.000	0.800	0.909
3	0.909	0.800	0.909	0.727	0.769	0.833	0.833
4	1.000	0.909	1.000	0.769	1.000	0.909	0.909
5	1.000	0.923	1.000	0.800	1.000	0.857	1.000
6	0.909	0.800	0.909	0.800	0.909	0.833	0.923
7	1.000	1.000	1.000	0.833	1.000	0.923	0.923
8	0.909	0.909	1.000	0.833	0.909	0.909	0.909
9	1.000	1.000	1.000	0.909	0.909	0.909	1.000
10	0.909	1.000	0.909	0.833	0.833	0.833	1.000
11	1.000	0.800	1.000	1.000	0.923	1.000	0.923
12	1.000	1.000	1.000	0.909	1.000	1.000	1.000
13	0.923	0.923	0.800	0.800	0.909	0.727	0.727
14	0.923	0.923	0.923	0.833	1.000	0.923	1.000
15	0.800	1.000	0.909	0.857	0.923	0.923	0.923
16	1.000	0.600	1.000	0.923	0.923	0.923	0.923
17	0.923	0.857	0.909	0.857	1.000	0.923	0.769
18	1.000	1.000	1.000	0.800	1.000	1.000	0.909
19	0.727	1.000	1.000	0.833	1.000	1.000	1.000
20	1.000	0.909	1.000	0.769	1.000	0.909	1.000

Tablo 4.1'den de anlaşılacağı gibi 20 kez yapılan sınıflandırma işleminde her seferinde farklı gözlemler ile eğitim-test verisi kullanıldığı için birbirinden farklı sonuçlar elde edilebilmektedir. Alizadeh-V1 verisi iki sınıflı bir veridir ve tablo 4.1 de sadece F1 metriği ifade edilmiştir. Her bir veriyi, çok sınıflı verilerde grupları, her bir grup içinde kullanılacak olan metrikleri düşünecek olursak bütün analiz sonuçlarını bu şekilde göstermek mantıklı olmayacaktır. Sonuçların kolay anlaşılabilmesi, yorumlanabilmesi ve karşılaştırılabilmesi açısından elde edilen sonuçlar ortalama (standart hata) olarak ifade edilmiştir.

Aşağıda yer alan tablo 4.2'de mikrodizi verilerinin, 4.3'de RNA dizileme verilerinin, 4.4'de metabolomik verilerinin sınıflandırma analizleri sonucunda her bir yönteme göre çeşitli metriklerle sınıflandırma performanslarını ifade eden sonuçlar yer almaktadır.

Tablo 4.2. Mikrodizi verilerinin sınıflandırma analiz sonuçları

VERİ	Otomatik Makine Öğrenmesi				Standart Makine Öğrenmesi		
	H2O (1 saat)	H2O (2 saat)	H2O (3 saat)	TPOT	NSC	RF	DVM
Alizadeh-V1							
MCC	0.907(0.029)	0.855(0.037)	**0.939(0.020)**	0.701(0.028)	0.904(0.030)	0.826(0.031)	0.864(0.034)
F1	0.947(0.017)	0.913(0.023)	**0.963(0.013)**	0.842(0.014)	0.950(0.015)	0.907(0.017)	0.929(0.018)
bACC	0.950(0.015)	0.921(0.020)	**0.967(0.011)**	0.842(0.015)	0.950(0.015)	0.908(0.016)	0.929(0.017)
Alizadeh-V2							
MCC	0.895(0.023)	0.919(0.023)	0.922(0.018)	0.948(0.013)	0.944(0.013)	0.922(0.020)	**1.000(0.000)**
***F1**	0.946(0.012)	0.959(0.011)	0.962(0.009)	0.974(0.007)	0.973(0.006)	0.958(0.011)	**1.000(0.000)**
CLL	0.948(0.028)	0.969(0.019)	0.986(0.010)	0.990(0.010)	**1.000(0.000)**	0.980(0.014)	**1.000(0.000)**
DLBCL	0.963(0.009)	0.972(0.007)	0.959(0.013)	0.983(0.005)	0.978(0.005)	0.976(0.006)	**1.000(0.000)**
FL	0.844(0.037)	0.865(0.035)	0.873(0.038)	0.897(0.028)	0.900(0.023)	0.816(0.060)	**1.000(0.000)**
***bACC**	0.955(0.012)	0.963(0.011)	0.961(0.010)	0.978(0.007)	0.983(0.004)	0.940(0.015)	**1.000(0.000)**
CLL	0.965(0.019)	0.986(0.010)	0.980(0.011)	0.992(0.008)	**1.000(0.000)**	0.983(0.011)	**1.000(0.000)**
DLBCL	0.953(0.012)	0.961(0.010)	0.960(0.010)	0.978(0.007)	0.979(0.005)	0.940(0.015)	**1.000(0.000)**
FL	0.952(0.018)	0.938(0.023)	0.940(0.024)	0.963(0.017)	0.983(0.004)	0.875(0.034)	**1.000(0.000)**
Armstrong-V1							
MCC	0.974(0.013)	0.995(0.005)	0.990(0.007)	**1.000(0.000)**	**1.000(0.000)**	0.958(0.018)	0.989(0.007)
bACC	0.982(0.009)	0.996(0.004)	0.995(0.004)	**1.000(0.000)**	**1.000(0.000)**	0.973(0.011)	0.993(0.005)
F1	0.980(0.010)	0.996(0.004)	0.993(0.005)	**1.000(0.000)**	**1.000(0.000)**	0.968(0.013)	0.992(0.005)
Armstrong-V2							

MCC	0.900(0.019)	0.896(0.025)	0.906(0.015)	**0.945(0.011)**	0.941(0.009)	0.935(0.014)	**0.945(0.009)**
***F1**	0.929(0.014)	0.927(0.018)	0.934(0.011)	**0.962(0.007)**	0.959(0.006)	0.954(0.010)	**0.962(0.007)**
ALL	0.968(0.013)	0.960(0.018)	0.969(0.010)	0.997(0.003)	**1.000(0.000)**	**1.000(0.000)**	**1.000(0.000)**
MLL	0.875(0.023)	0.881(0.031)	0.888(0.022)	**0.933(0.013)**	0.925(0.012)	0.917(0.017)	0.931(0.011)
AML	0.936(0.010)	0.932(0.013)	0.939(0.012)	**0.953(0.009)**	0.948(0.008)	0.945(0.011)	0.951(0.009)
***bACC**	0.947(0.010)	0.945(0.013)	0.951(0.008)	**0.971(0.005)**	0.969(0.005)	0.964(0.007)	**0.971(0.005)**
ALL	0.977(0.010)	0.968(0.014)	0.979(0.008)	0.998(0.002)	**1.000(0.000)**	**1.000(0.000)**	**1.000(0.000)**
MLL	0.908(0.018)	0.915(0.021)	0.916(0.016)	**0.951(0.010)**	0.942(0.010)	0.930(0.013)	0.946(0.009)
AML	0.951(0.008)	0.949(0.011)	0.953(0.010)	**0.963(0.008)**	0.961(0.006)	0.961(0.008)	**0.963(0.007)**
Bhattacharjee							
MCC	0.901(0.016)	0.918(0.011)	0.917(0.010)	0.910(0.008)	0.813(0.011)	0.878(0.013)	**0.922(0.066)**
***F1**	0.857(0.008)	0.864(0.005)	0.860(0.004)	0.863(0.003)	0.815(0.005)	0.810(0.031)	**0.865(0.006)**
AD	0.965(0.006)	**0.973(0.004)**	0.971(0.003)	0.970(0.003)	0.925(0.004)	0.910(0.048)	**0.973(0.004)**
COID	0.985(0.009)	0.987(0.006)	0.992(0.006)	0.991(0.006)	0.933(0.012)	**0.996(0.004)**	0.995(0.005)
NL	0.909(0.020)	0.899(0.022)	0.901(0.021)	**0.925(0.018)**	0.862(0.020)	0.887(0.025)	0.906(0.024)
SCLC	0.783(0.091)	0.867(0.070)	**0.908(0.056)**	0.859(0.071)	0.615(0.096)	0.500(0.115)	0.800(0.092)
SQ	0.849(0.041)	0.868(0.049)	0.898(0.018)	0.848(0.023)	0.777(0.026)	0.799(0.036)	**0.905(0.020)**
***bACC**	0.948(0.011)	**0.961(0.006)**	0.959(0.007)	0.944(0.007)	0.928(0.006)	0.913(0.009)	0.950(0.008)
AD	0.944(0.011)	**0.958(0.006)**	0.956(0.007)	0.942(0.007)	0.916(0.006)	0.913(0.009)	0.948(0.008)
COID	0.998(0.001)	0.999(0.001)	0.995(0.004)	0.992(0.006)	0.992(0.002)	**1.000(0.000)**	0.996(0.004)
NL	**0.964(0.010)**	0.954(0.013)	0.954(0.011)	0.961(0.011)	0.980(0.005)	0.927(0.017)	0.950(0.014)
SCLC	0.900(0.046)	0.924(0.041)	**0.974(0.025)**	0.924(0.041)	0.869(0.049)	0.750(0.057)	0.875(0.050)
SQ	0.919(0.026)	**0.953(0.013)**	0.944(0.014)	0.893(0.019)	0.909(0.018)	0.849(0.022)	0.928(0.017)

Bittner							
MCC	0.739(0.046)	**0.837(0.031)**	0.791(0.032)	0.773(0.036)	0.814(0.038)	0.732(0.036)	0.686(0.028)
F1	0.821(0.039)	0.900(0.023)	0.890(0.016)	0.884(0.019)	**0.906(0.019)**	0.863(0.017)	0.817(0.020)
bACC	0.855(0.026)	**0.910(0.018)**	0.890(0.016)	0.879(0.023)	0.900(0.021)	0.855(0.020)	0.830(0.015)
Bredel							
MCC	0.637(0.032)	0.675(0.039)	0.687(0.030)	0.701(0.042)	**0.751(0.032)**	0.740(0.030)	0.720(0.028)
***F1**	0.781(0.015)	0.788(0.023)	0.795(0.013)	0.796 (0.021)	**0.812(0.017)**	**0.812(0.014)**	0.816(0.013)
A	**0.050(0.034)**	**0.050(0.050)**	0.000(0.000)	**0.050(0.050)**	0.000(0.000)	0.000(0.000)	0.025(0.025)
GBM	0.907(0.015)	0.902(0.031)	0.921(0.015)	0.914(0.017)	0.928(0.017)	0.937(0.013)	**0.945(0.012)**
OG	0.763(0.027)	0.801(0.026)	0.798(0.020)	0.802(0.035)	**0.845(0.026)**	0.827(0.024)	0.815(0.026)
***bACC**	0.853(0.014)	0.862(0.018)	0.868(0.015)	0.854(0.023)	**0.886(0.015)**	0.873(0.016)	0.876(0.015)
A	0.506(0.033)	0.485(0.030)	0.481(0.006)	**0.521(0.025)**	0.498(0.002)	0.500(0.000)	0.504(0.023)
GBM	0.896(0.015)	0.899(0.021)	0.907(0.019)	0.878(0.025)	0.916(0.018)	0.905(0.019)	**0.918(0.018)**
OG	0.844(0.021)	0.874(0.019)	0.878(0.015)	0.881(0.024)	**0.915(0.018)**	0.893(0.018)	0.876(0.020)
Chen							
MCC	**0.923(0.013)**	0.922(0.012)	0.919(0.011)	0.858(0.014)	0.824(0.019)	0.881(0.014)	0.852(0.015)
F1	**0.968(0.006)**	0.967(0.005)	0.963(0.006)	0.939(0.006)	0.924(0.008)	0.951(0.006)	0.936(0.007)
bACC	**0.961(0.007)**	0.960(0.006)	0.961(0.005)	0.928(0.008)	0.911(0.009)	0.936(0.008)	0.925(0.007)
Chowdary							
MCC	**0.966(0.011)**	**0.966(0.011)**	**0.966(0.011)**	0.953(0.013)	0.909(0.013)	0.929(0.012)	0.960(0.011)
F1	**0.986(0.005)**	**0.986(0.004)**	**0.986(0.004)**	0.981(0.005)	0.964(0.005)	0.972(0.005)	0.983(0.005)
bACC	**0.985(0.005)**	0.983(0.005)	0.983(0.005)	0.975(0.007)	0.948(0.008)	0.960(0.007)	0.980(0.006)
Dyrskjot							

MCC	0.697(0.045)	0.679(0.047)	0.713(0.043)	0.727(0.034)	0.782(0.036)	**0.825(0.030)**	0.788(0.034)
***F1**	0.783(0.033)	0.767(0.035)	0.797(0.031)	0.832(0.021)	0.848(0.025)	**0.883(0.020)**	0.862(0.021)
T1	0.598(0.069)	0.571(0.072)	0.656(0.070)	0.765(0.041)	**0.847(0.038)**	0.778(0.044)	0.740(0.044)
T2+	0.810(0.055)	0.732(0.063)	0.832(0.054)	0.796(0.038)	0.860(0.038)	**0.940(0.028)**	0.937(0.026)
TA	0.873(0.021)	0.890(0.015)	0.858(0.020)	0.881(0.019)	0.844(0.025)	**0.917(0.013)**	0.898(0.018)
***bACC**	0.841(0.023)	0.837(0.024)	0.848(0.023)	0.861(0.019)	0.893(0.019)	**0.897(0.017)**	0.880(0.018)
T1	0.755(0.038)	0.735(0.035)	0.795(0.039)	0.826(0.032)	**0.914(0.025)**	0.850(0.030)	0.819(0.029)
T2+	0.907(0.032)	0.890(0.033)	0.919(0.034)	0.879(0.031)	0.949(0.019)	0.960(0.020)	**0.967(0.017)**
TA	0.863(0.021)	0.869(0.023)	0.850(0.022)	0.873(0.021)	0.864(0.022)	**0.900(0.016)**	0.883(0.020)
Garber							
MCC	**0.718(0.030)**	0.673(0.036)	0.654(0.025)	0.628(0.031)	0.705(0.026)	0.604(0.035)	0.662(0.029)
***F1**	**0.828(0.019)**	0.808(0.020)	0.801(0.013)	0.786(0.017)	0.819(0.013)	0.749(0.018)	0.800(0.016)
AC	0.882(0.011)	0.868(0.013)	0.863(0.009)	0.857(0.012)	**0.885(0.010)**	0.861(0.011)	0.868(0.010)
LCLC	**0.338(0.084)**	0.300(0.087)	0.192(0.062)	0.108(0.062)	0.067(0.046)	0.000(0.000)	0.183(0.065)
SCC	**0.829(0.031)**	0.824(0.025)	0.812(0.019)	0.793(0.028)	0.819(0.016)	0.777(0.031)	0.813(0.024)
SCLC	0.783(0.084)	0.675(0.104)	0.750(0.099)	0.733(0.099)	**0.981(0.018)**	0.350(0.109)	0.700(0.105)
***bACC**	**0.848(0.017)**	0.822(0.019)	0.819(0.015)	0.797(0.017)	0.812(0.013)	0.754(0.017)	0.802(0.017)
AC	**0.832(0.015)**	0.802(0.017)	0.810(0.015)	0.787(0.018)	0.806(0.014)	0.751(0.018)	0.790(0.018)
LCLC	**0.725(0.055)**	0.681(0.056)	0.643(0.052)	0.557(0.040)	0.540(0.033)	0.492(0.003)	0.629(0.050)
SCC	**0.879(0.020)**	0.872(0.018)	0.864(0.015)	0.853(0.020)	0.855(0.012)	0.831(0.021)	0.856(0.018)
SCLC	0.918(0.040)	0.842(0.054)	0.872(0.051)	0.874(0.050)	**0.949(0.034)**	0.675(0.055)	0.850(0.053)
Gordon							
MCC	0.994(0.004)	**0.997(0.003)**	**0.997(0.003)**	0.974(0.007)	0.972(0.007)	0.980(0.007)	0.980(0.007)

F1	**0.999(0.001)**	**0.999(0.001)**	**0.999(0.001)**	0.996(0.001)	0.995(0.001)	0.997(0.001)	0.997(0.001)
bACC	**0.999(0.001)**	**0.999(0.001)**	0.997(0.003)	0.982(0.006)	0.995(0.001)	0.983(0.006)	0.983(0.006)
Khan							
MCC	0.969(0.012)	0.966(0.009)	0.983(0.006)	0.961(0.014)	0.980(0.009)	**0.986(0.008)**	0.978(0.012)
***F1**	0.976(0.010)	0.974(0.007)	0.987(0.005)	0.968(0.012)	0.985(0.007)	**0.988(0.007)**	0.982(0.010)
BL	0.986(0.010)	**1.000(0.000)**	0.993(0.007)	**1.000(0.000)**	**1.000(0.000)**	**1.000(0.000)**	**1.000(0.000)**
EWS	0.961(0.017)	0.967(0.012)	0.984(0.006)	0.959(0.017)	0.979(0.012)	**0.997(0.003)**	0.977(0.016)
NB	0.984(0.009)	0.964(0.011)	0.980(0.009)	0.950(0.024)	0.984(0.009)	0.969(0.018)	**0.988(0.013)**
RMS	0.982(0.011)	0.977(0.010)	**0.993(0.005)**	0.977(0.010)	0.985(0.010)	0.987(0.007)	0.976(0.013)
***bACC**	0.983(0.007)	0.981(0.005)	0.991(0.003)	0.978(0.008)	0.989(0.005)	**0.992(0.005)**	0.988(0.007)
BL	0.998(0.002)	**1.000(0.000)**	0.999(0.001)	**1.000(0.000)**	**1.000(0.000)**	**1.000(0.000)**	**1.000(0.000)**
EWS	0.969(0.013)	0.975(0.010)	0.987(0.005)	0.968(0.013)	0.983(0.010)	**0.998(0.002)**	0.981(0.013)
NB	0.989(0.007)	0.976(0.009)	0.987(0.007)	0.969(0.016)	0.989(0.007)	0.975(0.014)	**0.990(0.010)**
RMS	0.991(0.006)	0.985(0.007)	**0.995(0.004)**	0.987(0.006)	0.992(0.005)	0.994(0.004)	0.988(0.007)
Laiho							
MCC	0.813(0.042)	**0.833(0.036)**	0.808(0.049)	0.727(0.064)	0.686(0.056)	0.517(0.092)	0.762(0.057)
F1	0.953(0.017)	**0.962(0.009)**	0.955(0.012)	0.948(0.013)	0.916(0.016)	0.938(0.010)	0.958(0.009)
bACC	0.881(0.028)	**0.906(0.025)**	0.900(0.027)	0.834(0.035)	0.856(0.030)	0.725(0.044)	0.853(0.033)
Lapointe-V1							
MCC	**0.732(0.022)**	0.688(0.037)	0.726(0.023)	0.569(0.041)	0.578(0.049)	0.688(0.028)	0.707(0.029)
***F1**	**0.840(0.012)**	0.813(0.020)	0.833(0.015)	0.741(0.023)	0.752(0.027)	0.794(0.018)	0.824(0.016)
PT1	**0.712(0.035)**	0.651(0.058)	0.664(0.057)	0.511(0.050)	0.670(0.059)	0.465(0.068)	0.686(0.042)
PT2	0.885(0.010)	0.864(0.016)	**0.889(0.010)**	0.818(0.018)	0.812(0.020)	0.877(0.011)	0.878(0.010)

PT3	**0.818(0.029)**	0.800(0.029)	0.815(0.022)	0.717(0.045)	0.677(0.030)	0.814(0.030)	0.788(0.033)
***bACC**	**0.844(0.014)**	0.827(0.019)	**0.844(0.014)**	0.775(0.022)	0.782(0.025)	0.807(0.016)	0.828(0.016)
PT1	**0.789(0.033)**	0.780(0.033)	**0.789(0.033)**	0.719(0.029)	0.775(0.028)	0.675(0.028)	0.785(0.026)
PT2	**0.849(0.015)**	0.824(0.019)	**0.849(0.015)**	0.774(0.023)	0.777(0.028)	0.814(0.018)	0.830(0.016)
PT3	0.867(0.016)	0.860(0.021)	0.867(0.016)	0.810(0.029)	0.797(0.022)	**0.869(0.021)**	0.850(0.022)
West							
MCC	**0.839(0.030)**	0.827(0.031)	0.832(0.024)	0.739(0.029)	0.738(0.029)	0.746(0.031)	0.72(0.029)
F1	**0.911(0.018)**	0.899(0.020)	0.907(0.014)	0.849(0.019)	0.850(0.018)	0.854(0.02)	0.849(0.017)
bACC	**0.914(0.016)**	0.907(0.016)	0.911(0.013)	0.861(0.015)	0.861(0.015)	0.864(0.016)	0.854(0.015)

Veriler 20 tekrardan elde edilen sonuçların ortalama (standart hata) değeri olarak ifade edilmiştir. Tabloda yer alan *F1 değerleri ağırlıklandırılmış F1 değerini, *bACC değerleri ağırlıklandırılmış bACC değerlerini ifade etmektedir.

Tablo 4.2'de mikrodizi verilerinin analiz sonuçları özetlenmiş ve en iyi performansa sahip sonuçlar kalın yazı tipi ile işaretlenmiştir. Alizadeh-V1 verisini en iyi sınıflandıran yöntem 3 saat süreyle çalıştırılan H2O yöntemidir. Alizadeh-V2 verisi için bazı gruplarda NSC yöntemi iyi performans göstersede genel olarak en iyi sınıflandıran yöntem DVM'dir. Armstrong-V1 verisinde TPOT ve NSC yöntemleri bütün verileri tamamen doğru olarak sınıflandırmayı başarmıştır. Armstrong-V2 verisinde gruplara göre en iyi performansı gösteren sonuçlar farklılık gösterse de ağırlıklandırma yapılarak elde edilen F1, bACC ve MCC metriklerine göre en yüksek performansa sahip yöntem DVM ve TPOT'dur. Grup sayısı 5 olan Bhattacharjee verisinde en yüksek performansa sahip yöntem gruplara göre değişiklik gösterse de genel olarak DVM yönteminde en iyi sınıflandırma performansına sahip denilebilir. Bittner verisinde MCC ve bACC metriklerine göre en iyi sınıflandırıcı 2 saat çalışan H2O yöntemi iken F1 ölçüsüne göre NSC yöntemi bu veri için en iyi sınıflandırıcıdır. Bredel verisinde NSC yöntemi diğer yöntemlere göre genel olarak oldukça başarılı sınıflandırma performansına sahiptir. Chen verisi için 1 saat çalışan H2O ile elde edilen performans her bir metrik ölçüsüne göre en iyi sınıflandırıcı olarak bulunmuştur. Chowdary verisi için 1, 2 ve 3 saat çalışan H2O yönteminin performans sonuçları birbirine oldukça yakın değerlerdir ve bu yöntemin en yüksek sınıflandırma yeteneğine sahip olduğu görülmektedir. Dyrskjot verisinde RF yönteminin performansı diğer yöntemlerden daha iyi bulunmuştur. Garber verisinde 1 saat çalıştırılan H2O yöntemi en iyi bulunmuştur. Oldukça iyi sınıflandırılabilen Gordon verisinde bütün yöntemler oldukça iyi sınıflandırma yapabilmiş olsa da 1, 2 ve 3 saat çalışan H2O yöntemi her bir metriğe göre en yüksek sınıflandırma becerisine sahip bulunmuştur. Khan verisinde ağırlıklandırılmış F1, bACC ve MCC metriklerine göre en iyi sınıflandırıcı RF yöntemi bulunurken bu verinin bazı gruplarının sınıflandırılmasında diğer yöntemler daha başarılı olabilmiştir. Grup sayısı 2 olan Laiho verisinde 2 saatlik çalışma sonucunda elde edilen H2O performansı en iyi bulunmuştur. Lapointe-V1 verisine göre ağırlıklandırılmış F1, bACC ve MCC metrikleri için en iyi performans 1 ve 3 saat çalışan H2O yöntemi sonucunda alınmıştır. West verisine göre de en iyi sınıflandırma performansı 1 saatlik çalışma sonucunda elde edilen H2O performansıdır.

Mikrodizi verilerinde genel itibariyle H2O yönteminin diğer yöntemlerden daha yüksek sınıflandırma performansı gösterdiği veri sayısı daha fazladır. H2O yönteminin çalışma süresine göre 1, 2 ve 3 saatlik denemeler arasında performans farklılığı zamana göre

artış gösteren bir trend izlememekte olup en iyi performansın elde edildiği çalışma saati farklılık göstermektedir. Genel olarak sonuçları değerlendirecek olursak 4 veride 1 saat çalıştırılan H2O, 2 veride 2 saat çalıştırılan H2O, 1 veride 3 saat çalıştırılan H2O ve 2 veride de 1, 2 ve 3 saat çalıştırılan H2O en iyi sınıflandırma performansına sahiptir. Çalıştırılma süresi dikkate alınmaksızın toplam 9 veride H2O yöntemi en iyi sınıflandırıcı olarak bulunmuştur. Diğer verilerin ikisinde RF, ikisinde DVM, birinde NSC yöntemleri ile en iyi sınıflandırma performansı elde edilmiştir. Ayrıca verilerin birinde TPOT ile NSC ve bir diğerinde TPOT ile DVM yöntemleri aynı olacak şekilde en iyi sınıflandırma performansına sahiptir.

Tablo 4.3. RNA-dizileme verilerinin sınıflandırma analiz sonuçları

VERİ	Otomatik Makine Öğrenmesi				Standart Makine Öğrenmesi		
	H2O (1 saat)	H2O (2 saat)	H2O (3 saat)	TPOT	NSC	RF	DVM
Alzheimer							
MCC	**0.844(0.024)**	0.805(0.027)	0.820(0.026)	0.579(0.072)	0.607(0.028)	0.623(0.031)	0.637(0.045)
F1	**0.952(0.008)**	0.942(0.007)	0.946(0.007)	0.648(0.069)	0.725(0.018)	0.718(0.023)	0.736(0.029)
bACC	**0.916(0.014)**	0.891(0.016)	0.902(0.017)	0.783(0.036)	0.824(0.016)	0.804(0.018)	0.811(0.024)
Rahim Ağzı Kanseri							
MCC	0.911(0.019)	0.905(0.020)	0.898(0.025)	**0.947(0.016)**	0.795(0.034)	0.851(0.028)	0.889(0.022)
F1	0.955(0.010)	0.950(0.011)	0.949(0.012)	**0.973(0.008)**	0.900(0.016)	0.925(0.013)	0.944(0.011)
bACC	0.953(0.010)	0.950(0.011)	0.944(0.014)	**0.972(0.008)**	0.894(0.018)	0.922(0.014)	0.941(0.012)
Lemfoblastoid							
MCC	0.982(0.008)	0.977(0.009)	**0.987(0.005)**	0.975(0.008)	0.943(0.011)	0.951(0.011)	0.985(0.005)
F1	0.992(0.004)	0.989(0.004)	**0.994(0.002)**	0.986(0.004)	0.969(0.006)	0.973(0.006)	0.991(0.003)
bACC	0.990(0.004)	0.988(0.005)	**0.993(0.003)**	0.987(0.004)	0.97(0.005)	0.974(0.006)	0.992(0.003)
Zebra Balığı							
MCC	0.988(0.005)	0.967(0.012)	0.976(0.007)	0.982(0.004)	0.966(0.007)	**0.991(0.003)**	0.981(0.006)
***F1**	0.992(0.003)	0.976(0.009)	0.984(0.005)	0.988(0.003)	0.977(0.005)	**0.994(0.002)**	0.987(0.004)
Juvenile	0.993(0.004)	0.985(0.005)	0.992(0.005)	0.995(0.003)	0.980(0.006)	**0.998(0.002)**	0.984(0.005)
Adolescent	**0.993(0.003)**	0.982(0.009)	0.982(0.005)	0.988(0.003)	0.984(0.004)	**0.993(0.002)**	0.989(0.004)
Adult	0.989(0.005)	0.956(0.021)	0.972(0.007)	0.976(0.006)	0.963(0.007)	**0.990(0.003)**	0.989(0.004)

***bACC**	0.994(0.003)	0.983(0.006)	0.987(0.004)	0.990(0.002)	0.983(0.004)	**0.996(0.002)**	0.990(0.003)
Juvenile	0.994(0.003)	0.975(0.004)	0.994(0.004)	0.997(0.002)	0.984(0.005)	**0.998(0.002)**	0.987(0.004)
Adolescent	0.994(0.003)	0.988(0.006)	0.988(0.004)	0.993(0.002)	0.989(0.003)	**0.996(0.001)**	0.990(0.004)
Adult	**0.993(0.003)**	0.969(0.015)	0.976(0.006)	0.978(0.005)	0.972(0.006)	0.992(0.003)	**0.993(0.003)**
Fare Kök Hücre (Kolodz)							
MCC	0.996(0.002)	0.997(0.002)	0.997(0.001)	**1.000(0.000)**	0.998(0.001)	0.999(0.001)	**1.000(0.000)**
***F1**	0.997(0.001)	0.998(0.001)	0.998(0.001)	**1.000(0.000)**	0.999(0.000)	0.999(0.000)	**1.000(0.000)**
serum+LIF	0.997(0.001)	0.998(0.001)	0.998(0.001)	**1.000(0.000)**	0.999(0.001)	0.999(0.000)	**1.000(0.000)**
2i+LIF	0.994(0.002)	0.995(0.003)	0.996(0.001)	**0.999(0.001)**	0.997(0.001)	0.998(0.001)	**0.999(0.001)**
a2i+LIF	0.999(0.001)	**1.000(0.000)**	**1.000(0.000)**	**1.000(0.000)**	**1.000(0.000)**	**1.000(0.000)**	**1.000(0.000)**
***bACC**	0.998(0.001)	0.998(0.001)	0.999(0.001)	**1.000(0.000)**	0.999(0.000)	0.999(0.000)	**1.000(0.000)**
serum+LIF	0.998(0.001)	0.998(0.001)	0.999(0.001)	**1.000(0.000)**	0.999(0.000)	0.999(0.000)	**1.000(0.000)**
2i+LIF	0.995(0.002)	0.996(0.002)	0.997(0.001)	**1.000(0.000)**	0.998(0.001)	0.998(0.001)	0.999(0.001)
a2i+LIF	0.999(0.001)	**1.000(0.000)**	**1.000(0.000)**	**1.000(0.000)**	**1.000(0.000)**	**1.000(0.000)**	**1.000(0.000)**
Fare Embriyo							
MCC	0.860(0.018)	0.834(0.023)	0.829(0.020)	**0.899(0.020)**	0.653(0.019)	0.894(0.012)	0.831(0.014)
***F1**	0.856(0.022)	0.826(0.026)	0.828(0.022)	0.894(0.024)	0.624(0.022)	**0.899(0.013)**	0.821(0.018)
16-cell	0.568(0.071)	0.517(0.069)	0.538(0.064)	0.675(0.079)	0.237(0.046)	**0.776(0.042)**	0.434(0.067)
zygote	0.700(0.105)	0.696(0.100)	0.750(0.100)	**1.000(0.000)**	0.100(0.069)	0.700(0.105)	**1.000(0.000)**
4-cell	0.993(0.007)	0.993(0.007)	0.994(0.006)	0.920(0.057)	0.939(0.027)	**1.000(0.000)**	**1.000(0.000)**
8-cell	0.749(0.023)	0.725(0.024)	0.732(0.023)	0.812(0.033)	0.631(0.011)	**0.834(0.022)**	0.685(0.015)

early 2-cell	0.896(0.034)	0.849(0.043)	0.844(0.049)	**0.940(0.021)**	0.780(0.022)	0.880(0.022)	**0.940(0.021)**
early blast	0.946(0.014)	0.927(0.023)	0.896(0.025)	**0.954(0.015)**	0.470(0.063)	0.945(0.016)	0.917(0.026)
late 2-cel	0.960(0.025)	0.948(0.021)	0.945(0.029)	**0.990(0.010)**	0.810(0.039)	0.920(0.036)	0.960(0.018)
late blast	0.983(0.007)	0.977(0.012)	0.955(0.018)	**1.000(0.000)**	0.879(0.020)	0.961(0.010)	**1.000(0.000)**
mid 2-cell	0.979(0.012)	0.924(0.025)	0.886(0.030)	**0.993(0.007)**	0.925(0.020)	0.954(0.019)	0.971(0.013)
mid blast	0.963(0.009)	0.946(0.028)	0.949(0.010)	**0.974(0.007)**	0.775(0.017)	0.955(0.009)	0.948(0.011)
***bACC**	0.924(0.010)	0.910(0.013)	0.908(0.011)	**0.944(0.012)**	0.808(0.010)	**0.944(0.007)**	0.906(0.008)
16-cell	0.730(0.034)	0.693(0.034)	0.710(0.032)	0.800(0.040)	0.575(0.016)	**0.836(0.026)**	0.660(0.026)
zygote	0.849(0.053)	0.873(0.049)	0.875(0.050)	**1.000(0.000)**	0.550(0.034)	0.850(0.053)	**1.000(0.000)**
4-cell	0.994(0.006)	0.994(0.006)	**1.000(0.000)**	0.956(0.030)	0.967(0.018)	**1.000(0.000)**	**1.000(0.000)**
8-cell	0.939(0.007)	0.932(0.008)	0.933(0.008)	0.955(0.009)	0.897(0.007)	**0.961(0.005)**	0.915(0.006)
early 2-cell	0.996(0.002)	0.991(0.005)	0.991(0.004)	**0.998(0.001)**	0.992(0.001)	0.996(0.001)	**0.998(0.001)**
early blast	0.962(0.012)	0.952(0.017)	0.924(0.019)	0.967(0.012)	0.679(0.031)	**0.968(0.012)**	0.942(0.017)
late 2-cel	0.990(0.008)	0.966(0.015)	0.958(0.021)	**0.992(0.008)**	0.905(0.025)	0.942(0.025)	0.967(0.015)
late blast	0.991(0.005)	0.987(0.009)	0.968(0.013)	**1.000(0.000)**	0.904(0.015)	0.964(0.009)	**1.000(0.000)**
mid 2-cell	0.999(0.001)	0.996(0.001)	0.994(0.002)	**1.000(0.000)**	0.996(0.001)	0.998(0.001)	0.999(0.001)
mid blast	0.975(0.005)	0.963(0.018)	0.969(0.006)	**0.987(0.003)**	0.897(0.008)	0.976(0.005)	0.975(0.005)

Veriler 20 tekrardan elde edilen sonuçların ortalama (standart hata) değeri olarak ifade edilmiştir. Tabloda yer alan *F1 değerleri ağırlıklandırılmış F1 değerini, *bACC değerleri ağırlıklandırılmış bACC değerlerini ifade etmektedir.

Tablo 4.3'de RNA dizileme verilerinin sınıflandırma performansları gösterilmekte olup en iyi sınıflandırma performansına sahip sonuçlar kalın yazı tipi ile işaretlenmiştir. Alzheimer verisinde en iyi sınıflandırma performansına sahip yöntem 1 saat çalıştırılan H2O yöntemi olup H2O ile yapılan analizlerin performansları NSC, RF, DVM ve TPOT yöntemlerinden oldukça yüksek bulunmuştur. Rahim Ağzı Kanseri verisinde TPOT yöntemi ile elde edilen sınıflandırma performansları diğer yöntemlere göre oldukça başarılı bulunmuştur. Lemfoblastoid verisi genel itibariyle her bir yönteme göre iyi sınıflandırılabilen bir veri olup en iyi sınıflandırma performansı 3 saat çalıştırılarak elde edilen H2O yönteminde elde edilmiştir. Zebra balığı verisine göre en iyi sınıflandırma performansı RF yöntemi ile elde edilmiştir. Fare Kök Hücre verisi uygulanan her bir yöntem içinde oldukça iyi bir şekilde sınıflandırılmış olsa da en yüksek performans DVM ve TPOT ile elde edilmiştir. Fare Embriyo verisi için en iyi sınıflandırma performansı MCC metriğine göre TPOT, ağırlıklandırılmış F1 metriğine göre RF ve ağırlıklandırılmış bACC metriğine göre hem RF hem de TPOT en iyi sınıflandırıcılar olarak bulunmuştur. Grup sayısı 10 olan bu verinin her bir grubunun sınıflandırma performansı ise genellikle RF, DVM ve TPOT yöntemlerinde en iyi olarak bulunmuştur.

RNA dizileme verilerinde genel itibariyle TPOT yönteminin diğer yöntemlerden daha yüksek sınıflandırma performansı gösterdiği veri sayısı daha fazladır. Toplamda 6 RNA dizileme verisinden ikisinde 1 saat ve 3 saat çalıştırılan H2O yöntemi en iyi sınıflandırma performansına sahiptir. Verilerin 1 tanesinde RF, 1 tanesinde TPOT ve 1 tanesinde de hem RF hem TPOT yöntemi en iyi en iyi sınıflandırma performansına sahip yöntemler olarak bulunmuştur. Bir veride de hem TPOT hem de DVM yöntemi en iyi sınıflandırıcı olarak bulunmuştur. Yani toplamda 3 veri için TPOT yöntemi diğer yöntemlerle birlikte de olsa en iyi sınıflandırmayı yapmıştır.

Tablo 4.4. Metabolomik verilerinin sınıflandırma analizi sonuçları

VERİ	Otomatik Makine Öğrenmesi				Standart Makine Öğrenmesi		
	H2O (1 saat)	H2O (2 saat)	H2O (3 saat)	TPOT	NSC	RF	DVM
ST000369							
MCC	0.624(0.026)	0.643(0.021)	**0.647(0.021)**	0.515(0.038)	0.524(0.023)	0.559(0.025)	0.524(0.023)
F1	0.780(0.024)	0.798(0.017)	**0.806(0.017)**	0.758(0.020)	0.758(0.013)	0.782(0.013)	0.760(0.013)
bACC	0.805(0.015)	0.815(0.012)	**0.819(0.012)**	0.755(0.019)	0.760(0.011)	0.778(0.013)	0.759(0.011)
ST000389							
MCC	0.361(0.052)	0.342(0.040)	**0.382(0.051)**	0.150(0.050)	0.099(0.036)	0.348(0.034)	0.171(0.050)
F1	0.407(0.051)	0.371(0.044)	**0.449(0.052)**	0.312(0.039)	0.204(0.030)	0.363(0.037)	0.417(0.038)
bACC	0.638(0.024)	0.619(0.019)	**0.651(0.026)**	0.559(0.021)	0.533(0.012)	0.611(0.016)	0.590(0.025)
ST000388							
MCC	**0.123(0.027)**	0.089(0.012)	0.078(0.000)	0.061(0.020)	0.059(0.018)	0.006(0.022)	-0.192(0.039)
F1	0.142(0.026)	0.118(0.018)	0.100(0.000)	0.103(0.005)	0.101(0.001)	0.100(0.004)	**0.156(0.028)**
bACC	**0.517(0.011)**	0.507(0.007)	0.500(0.000)	0.495(0.007)	0.493(0.007)	0.484(0.008)	0.406(0.020)
ST000390							
MCC	**0.695(0.027)**	0.675(0.026)	0.659(0.019)	0.566(0.023)	0.599(0.023)	0.618(0.022)	0.509(0.025)
F1	**0.815(0.029)**	0.800(0.019)	0.799(0.013)	0.787(0.011)	0.803(0.012)	0.809(0.013)	0.757(0.013)
bACC	**0.836(0.016)**	0.825(0.014)	0.818(0.010)	0.777(0.011)	0.795(0.011)	0.805(0.011)	0.752(0.013)
ST000391							
MCC	0.564(0.025)	**0.575(0.030)**	0.524(0.023)	0.464(0.048)	0.562(0.036)	0.559(0.042)	0.311(0.045)
F1	0.731(0.028)	0.744(0.029)	0.758(0.013)	0.730(0.023)	0.756(0.022)	**0.772(0.024)**	0.658(0.020)

bACC	0.766(0.014)	**0.773(0.017)**	0.760(0.011)	0.725(0.023)	**0.773(0.018)**	0.770(0.021)	0.650(0.021)
ST000392							
MCC	0.677(0.026)	**0.690(0.023)**	0.664(0.024)	0.629(0.026)	0.609(0.025)	0.595(0.024)	0.521(0.033)
F1	0.800(0.021)	**0.819(0.016)**	0.800(0.018)	0.810(0.015)	0.790(0.014)	0.790(0.014)	0.744(0.021)
bACC	0.827(0.015)	**0.836(0.012)**	0.822(0.013)	0.811(0.013)	0.801(0.012)	0.795(0.012)	0.756(0.016)
ST000356							
MCC	0.389(0.019)	0.367(0.014)	0.386(0.014)	0.430(0.017)	**0.443(0.018)**	0.391(0.020)	0.355(0.023)
***F1**	0.575(0.013)	0.547(0.011)	0.570(0.010)	**0.600(0.013)**	0.551(0.016)	0.575(0.012)	0.556(0.016)
breast cancer -	0.000(0.000)	0.000(0.000)	0.000(0.000)	0.000(0.000)	0.000(0.000)	0.000(0.000)	**0.017(0.017)**
breast cancer 2	0.483(0.021)	0.512(0.017)	0.493(0.022)	0.531(0.024)	**0.583(0.021)**	0.507(0.021)	0.456(0.025)
breast cancer 3	0.475(0.018)	0.363(0.030)	0.448(0.021)	0.512(0.032)	0.310(0.050)	0.458(0.022)	0.439(0.024)
control	0.989(0.005)	0.989(0.005)	0.989(0.005)	0.960(0.018)	0.969(0.008)	0.975(0.006)	**0.997(0.003)**
***bACC**	0.681(0.010)	0.668(0.007)	0.679(0.007)	**0.698(0.009)**	0.685(0.008)	0.679(0.010)	0.669(0.012)
breast cancer -	0.480(0.005)	0.467(0.007)	0.474(0.005)	0.499(0.001)	**0.500(0.000)**	0.499(0.001)	0.469(0.026)
breast cancer 2	0.593(0.014)	0.591(0.011)	0.596(0.011)	0.620(0.014)	**0.638(0.012)**	0.598(0.016)	0.572(0.019)
breast cancer 3	0.589(0.013)	0.556(0.012)	0.583(0.012)	**0.611(0.012)**	0.559(0.018)	0.580(0.015)	0.577(0.016)
control	0.995(0.003)	0.991(0.005)	0.991(0.005)	0.986(0.005)	0.979(0.006)	0.987(0.004)	**0.997(0.003)**

Tablo 4.4'de metabolomik verilerinin sınıflandırma performansları gösterilmekte olup en iyi sınıflandırma performansına sahip sonuçlar kalın yazı tipi ile işaretlenmiştir. ST000369 ve ST000389 verileri için en iyi sınıflandırma performansı 3 saat çalıştırılan H2O yönteminde elde edilmiştir. ST000388 verisi için en iyi sınıflandırma yöntemi MCC ve bACC metrikleri için H2O olup 1 saatlik çalışma sonucunda elde edilmiştir. F1 metriğine göre bu veride en iyi sınıflandırma yapan yöntem DVM yöntemidir. ST000390 verisi için en iyi sınıflandırıcı 1 saat çalışma ile H2O yöntemi olmuştur. ST000391 verisi için en iyi sınıflandırma performansı MCC ve bACC metriklerine göre 2 saat çalışan H2O yöntemi ile elde edilirken iken F1 metriğine göre RF yöntemi ile elde edilmiştir. ST000392 verisi için en iyi sınıflandırma performansı 2 saat çalıştırılan H2O yöntemi ile elde edilmiştir. ST000356 verisi için en iyi sınıflandırma performansı MCC metriğine göre NSC, ağırlıklandırılmış F1 ve bACC metriğine göre TPOT yöntemi en iyi sınıflandırma performansına sahiptir. Grup sayısı 4 olan bu verinin gruplarının sınıflandırma performansları ise NSC, DVM ve TPOT yöntemleri ile en iyi performansa sahip olarak bulunmuştur.

Metabolomik verilerin genel sınıflandırma performanslarını her bir sınıflandırma yöntemi açısından değerlendirecek olursak genel itibariyle H2O yöntemi diğer yöntemlere göre daha başarılı bulunmuştur. Toplam 7 verinin 6'sında H2O yöntemi ile elde edilen sonuçlar diğer yöntemlerden daha iyi bulunmuştur. Çalışma süresi 1, 2 ve 3 saat olan H2O yöntemi ile 2'şer veri en iyi performansa sahip olacak şekilde sınıflandırılmıştır. Metabolomik verilerinden bir tanesi ise TPOT ve NSC yöntemi ile en iyi sınıflandırılabilmiştir.

Kullanılan genomik, transkriptomik ve metabolomik verilerin tamamı açısından daha genel olarak sonuçları inceleyecek olursak; toplam 29 veriden 23'ünde otomatik makine öğrenmesi yaklaşımlarıyla en iyi şekilde sınıflandırma performansı elde edilmiştir. Bu 23 verinin 5 tanesinde otomatik olmayan sınıflandırma yöntemleriyle de en iyi sınıflandırma performansı elde edilebilmiştir. Yani sadece AutoML yöntemlerinin en iyi sınıflandırma performansı gösterdiği veri sayısı 18'dir. RF, DVM ve NSC yöntemleri ile en iyi sınıflandırma performanslarının elde edildiği veri sayısı 11'dir. Bu 11 verinin 5 tanesinde otomatik TPOT yöntemi de en iyi sınıflandırma performansına sahiptir.

Tablo 4.5. Sınıflandırma yöntemlerinin kullanmış oldukları değişken sayıları

VERİ	H2, RF, DVM	NSC	TPOT
Rahim Ağzı Kanseri	**118.45(2.18)**	213.85(37.87)	714.00(0.00)
Alzheimer	**137.90(4.25)**	201.90(49.63)	1620.55(174.26)
Fare Embriyo	1870.75(10.14)	2000.00(0.00)	**1343.15(195.72)**
Zebra Balığı	**1172.45(6.79)**	1803.45(50.84)	2000.00(0.00)
Lemfoblastoid	**915.50(7.73)**	526.95(79.18)	1804.35(134.66)
Fare Kök Hücre	906.20(3.61)	**495.60(52.61)**	1814.65(127.77)
Alizadeh-V1	**92.05(2.47)**	112.45(13.75)	990.85(71.71)
Alizadeh-V2	**1016.45(7.85)**	1950.40(1.93)	1722.75(151.71)
Armstrong-V1	271.85(21.31)	**221.20(31.97)**	1081.00(0.00)
Armstrong-V2	**476.00(2.18)**	630.15(73.11)	2000.00(0.00)
Bhattacharjee	946.65(23.80)	**677.75(129.22)**	1349.35(107.73)
Bittner	**92.50(4.33)**	169.35(23.50)	2000.00(0.00)
Bredel	171.50(4.67)	**40.30(6.27)**	1406.25(152.75)
Chen	**64.00(0.00)**	76.15(1.20)	82.50(2.50)
Chowdary	135.95(0.41)	**110.45(9.91)**	174.10(7.90)
Dyrskjot	**237.15(5.98)**	703.65(93.72)	748.70(127.73)
Garber	**167.55(5.49)**	179.50(43.15)	1544.10(182.73)
Gordon	**580.60(3.68)**	1343.25(4.53)	1626.00(0.00)
Khan	352.75(5.38)	**256.00(49.07)**	1069.00(0.00)
Laiho	**123.25(7.45)**	1742.20(38.72)	2000.00(0.00)
Lapointe-V1	**192.25(33.38)**	338.25(28.64)	1548.15(76.85)
West	**83.95(2.26)**	571.75(102.94)	1198.00(0.00)
ST000369	109.55(20.02)	**52.35(5.71)**	511.00(0.00)
ST000389	64.00(0.00)	**32.30(7.86)**	129.00(0.00)
ST000388	494.00(0.00)	**5.20(3.59)**	989.00(0.00)
ST000390	**33.95(3.39)**	80.45(12.18)	166.70(11.24)
ST000391	**132.40(7.92)**	191.65(33.04)	1901.80(98.20)
ST000392	**22.00(0.70)**	81.55(5.55)	158.00(0.00)
ST000356	**46.75(0.58)**	64.20(8.85)	97.35(3.65)

Tabloda yer alan değerler 20 verinin ortalama(standart hata) değeri olarak ifade edilmiştir.

Tablo 4.5'de kullanılan sınıflandırma yöntemlerinin sınıflandırma esnasında kullanmış oldukları değişken sayıları gösterilmekte olup en az değişkene sahip olan kalın yazı tipi ile işaretlenmiştir. Çalışma süresi fark etmeksizin H2O, RF ve DVM yöntemlerinin kullanmış oldukları değişken sayıları aynıdır. Çünkü bu yöntemler içsel olarak değişken seçimi yapmamakta olup sınıflandırma öncesinde değişken seçimi yöntemlerinden IG yöntemi kullanılarak elde edilen değişkenler ile sınıflandırma yapılmıştır.

Omik verileri ile sınıflandırma yaparken en az değişken kullanılarak en yüksek sınıflandırma performansı elde edilmesi hedeflenir. Birçok çalışmada sınıfların ayrılmasında en etkili olan gen, protein ya da metabolitin belirlenmesi istenir. Bu nedenle yöntemleri karşılaştırırken doğru bir değerlendirme yapabilmek için her verilerin sınıflandırma performansı ile birlikte kullanılan değişken sayıları da değerlendirilmelidir.

Kullanılan 29 omik verisinden 19'unda en az değişkeni kullanarak sınıflandırma yapan yöntemler H2O, DVM ve RF'dir. Yani IG yönteminin kullanımıyla değişken seçimi yapılarak bu yöntemlerin 19 veride en az değişkenle analiz yapması sağlanmıştır. Toplam 9 veride NSC yöntemi en az değişkeni kullanarak sınıflandırma yapmıştır. Yalnızca 1 veride TPOT yöntemi ile daha az değişken kullanılarak sınıflandırma yapılmıştır. TPOT yöntemi otomatik olarak kendisi değişken seçimini gerçekleştirmektedir. Bazı verilerde 20 tekrarın hiçbirinde de değişken seçimi uygulamazken bazı verilerinde sadece birkaç tekrarında değişken seçimi yapmıştır. Bu nedenle değişken seçimi olmaksızın her tekrarda bütün değişkenlerin kullanılarak analize tabi tutulduğu TPOT yönteminde kullanılan değişken sayıları diğer yöntemlere göre genel itibariyle oldukça fazla bulunmuştur.

Tablo 4.6. Omik verilerin sınıflandırma yöntemleri için hesaplama maliyetleri

VERİ	NSC	RF	DVM	TPOT
Rahim Ağzı Kanseri	1.52	22.40	1.17	197.91
Alzheimer	4.70	25.86	1.19	653.73
Fare Embriyo	24.63	2234.05	27.80	338.71
Zebra Balığı	8.86	1027.50	7.76	41.68
Lemfoblastoid	6.72	287.68	4.50	64.40
Fare Kök Hücre	13.88	2519.48	13.87	71.10
Alizadeh-V1	3.76	14.23	1.17	18.40
Alizadeh-V2	6.02	151.30	4.43	60.72
Armstrong-V1	4.79	36.21	1.49	25.33
Armstrong-V2	4.23	107.22	2.54	43.93
Bhattacharjee	7.45	839.71	7.27	90.17
Bittner	7.41	12.01	1.02	23.96
Bredel	6.64	34.02	1.45	52.28
Chen	1.84	50.01	1.03	3.06
Chowdary	2.20	37.87	1.26	0.49
Dyrskjot	5.21	31.14	1.61	25.04
Garber	9.73	41.26	1.68	27.87
Gordon	7.43	183.20	3.19	31.85
Khan	4.23	97.52	2.17	9.60
Laiho	8.27	16.11	1.22	24.86
Lapointe-V1	6.73	39.34	1.26	12.62
West	5.14	12.84	1.02	14.92
ST000369	3.36	25.53	0.92	19.22
ST000389	1.83	27.70	0.91	22.45
ST000388	6.84	158.17	2.44	21.01
ST000390	2.95	12.47	0.81	2.07
ST000391	8.64	25.26	0.94	50.87
ST000392	2.36	19.61	0.79	4.35
ST000356	1.58	36.62	1.07	7.04

Hesaplama maliyetleri saniye cinsinden ifade edilmiştir.

Yukarıda tablo 4.2, 4.3, 4.4'de omik verilerin sınıflandırma performansları ve tablo 4.5'de model oluşturulurken kullanılan değişken sayıları, sparslık dereceleri değerlendirilmiştir. Fakat yöntemlerin kullanılabilirliği açısından hesaplama maliyetlerinin de değerlendirilmesi gerekmektedir. Araştırmacıların kullanacakları yönteme karar vermeden önce bu üç durumu optimize etmeleri gerekebilir. Örneğin bir birine oldukça yakın sınıflandırma performansına sahip olan iki yöntemden biri 3 gün diğeri ise 3 saatte sınıflandırma analizi gerçekleştiriyorsa kişiler performansı kısmen düşük olup daha hızlı çalışan yöntemi kullanmak isteyebilirler. Bu nedenle bizde bu çalışmada kullanılan yöntemlerin hesaplama maliyetlerini de değerlendirdik. Tablo 4.6'da her bir veri için her bir yöntemin hesaplama maliyetleri özetlenmiştir.

H2O yöntemi için analizler 1, 2 ve 3 saat olarak, yani sırasıyla 3600, 7200 ve 10800 saniye çalışacak şekilde ayarlanmıştır. Bu nedenle analizlerin sonlanması dakika bazında bazen belirtilen süreyi aşmış, bazen de daha kısa sürede bitmiş olsa da yaklaşık olarak belirtilen sürelerde sona ermiştir. H2O hesaplama maliyetini her veri için aynı olacak şekilde biz kendimiz belirlediğimiz için tablo 4.6'da bu yöntemin hesaplama maliyeti gösterilmemiştir. Diğer yöntemlerin analiz sürelerine baktığımızda Fare Embriyo verisi hariç tüm verilerde DVM yöntemi maliyeti en düşük olarak bulunmuştur. Fare Embriyo verisinde ise en düşük maliyetli yöntem NCS'dir. DVM'den sonra en düşük maliyetli yöntem, ST000390 ve Chowdary verileri hariç tüm veriler için NSC'dir. ST000390 ve Chowdary verileri içinse en düşük maliyetli yöntem TPOT'dur. RF ve TPOT yöntemleri arasında ise genel bir sıralama yapmak söz konusu değildir. Bazı verilerde RF yöntemi TPOT yönteminden çok daha hızlı sonuçlar verirken bazı verilerde tam tersi durumlar söz konusu olabilmektedir.

TPOT yöntemi için her veride her bir tekrar için aynı parametreler kullanılmış olsa da birbirine çok yakın verilerde dahi farklı çalışma süreleri kaydedilmiş olup tablo 4.6'da yaklaşık maliyetlerine yer verilmiştir. Ayrıca TPOT yönteminde kullanılan parametrelere bağlı olarak bu süre değişiklik gösterebilir. Her ne kadar diğer yöntemlerin performansı da kullanılan parametrelere bağlı olsa, TPOT yönteminde bu süre farklılığı çok daha belirgin olarak ortaya çıkacaktır. Geliştiriciler bu yöntemi kullanırken bizim elde etmiş olduğumuz sürelerden çok daha uzun sürelerde çalıştırılmasını tavsiye etmektedirler. Elde edilen hesaplama maliyetleri Intel(R) Core

i7-7700 HQ CPU 2.80 GHz 16 GB 8 çekirdekli 64 bit özelliğinde bilgisayar kullanılarak hesaplanmıştır.

Kullanılan 29 omik veri için AutoML yöntemleri ile 12 çekirdek kullanılmış olup toplam hesaplama maliyeti yaklaşık olarak 3625 saattir. RF, DVM ve NSC yöntemlerinin maliyetini de düşünecek olursak bu çalışmanın sınıflandırma işi için toplam hesaplama maliyeti yaklaşık olarak 3915 saattir.

5. TARTIŞMA VE SONUÇ

Bu çalışmada omik verilerin sınıflandırılmasında sıklıkla başvurulan güçlü sınıflandırma yöntemlerinden RF, NSC ve DVM yöntemleri ile henüz omik veriler için denenmemiş olan otomatik makine öğrenimi yaklaşımları karşılaştırılmıştır. Otomatik makine öğrenimi yaklaşımlarından H2O ve TPOT kullanılmıştır. Literatürde bu yöntemlerin performansı birçok veri için denenmiş olsa da omik veriler için bu yöntemlerin performansını değerlendiren çalışmalar bulunmamaktadır. Bu çalışma sonucunda; RNA dizileme verilerinde, farklı teknolojilerle elde edilen mikrodizi verilerinde ve yine farklı teknoloji ve yöntemlerle elde edilen metabolomik verilerde hem NSC, RF ve DVM hem de otomatik makine öğrenmesi yöntemlerinden H2O ve TPOT yöntemlerinin uygulanmasıyla bu yöntemlerin performansları, birbirlerine karşı üstünlükleri ortaya koyulmuştur.

Hem standart NSC, RF ve DVM yöntemleri hem de otomatik H2O ve TPOT yöntemleri ile elde edilmiş sonuçlara bakıldığında her zaman kullanılabilecek, her veride en iyi performansı gösteren bir yaklaşım bulunamamıştır. Fakat H2O ve TPOT olarak ayırmaksızın AutoML yöntemleri ile birçok veride en iyi sınıflandırma performansı elde edilebilmiştir. Otomatik makine öğrenmesi yaklaşımları ile yapılan uygulamalarda arama uzayındaki denenebilecek durumları H2O için 1, 2 ve 3 saatlik sürelerle, TPOT için 5 yineleme ve 20 popülasyon büyüklüğü ile sınırlandırdık. Bu sınırlamaların artırılmasıyla denenecek olan yöntemler denenmiş olan yöntemlerden daha yüksek sınıflandırma performansı gösteriyor ise sınıflandırma performansı artış gösterebilir. Fakat analiz sonuçlarına göre H2O yönteminde süre artışı ile birlikte performans artışı gibi bir durumdan söz edilememektedir. Saat sınırlamasının artmasıyla elde edilecek olan sınıflandırma performansının artması beklenirken; bazı verilerde zamana göre performans artışı gözlenmemiş, bazı verilerde ise zamanla daha düşük performanslı sonuçların elde edildiği de olmuştur. Bu sınırlamaların arttırılmasıyla sınıflandırma

performansı artış gösterebilirken hesaplama maliyeti artışı, verilerin ezberlenmesiyle yeni gözlemlerde düşük performansların elde edilmesi gibi istenmeyen durumlarda da artış görülebilir. Bu noktada araştırmacıların bu sınırlamalara kendilerinin karar vermeleri gerekmektedir ve performans-maliyet açısından optimizasyon yapılmalıdır. Düşük maliyetli bir sınırlama kullanılmasıyla da yeterince iyi sınıflandırma performansı elde edilebiliyorsa yüksek maliyetli sınırlama parametrelerinin kullanılmasına gerek duyulmayabilir. H2O yönteminde olduğu gibi bazen daha uzun süreli analizlerle daha düşük performanslı sonuçlar da elde edilebilir.

Otomatik makine öğrenmesi yöntemleri ile bazı verilerde daha yüksek maliyetli sınırlamalar kullanılarak denemeler yapıldı. Elde edilen sonuçlar ve yapmış olduğumuz analizler incelendiğinde H2O yönteminde sürenin artmasıyla performansta belirgin bir artış durumu gözlenmezken TPOT yönteminde daha fazla yineleme ve popülasyon büyüklüğünün kullanılmasıyla literatür ile uyumlu olarak performansta artışların olduğu görüldü (Gijsbers ve ark., 2019). Her ne kadar AutoML yaklaşımları veri analizi aşamalarını baştan sona en iyi şekilde optimize etmeyi hedeflese de bu yöntemler için bulunan birçok parametreyi araştırmacının kendi deneyim ve bilgi birikimi ile belirlemesi gerekmektedir. Ayrıca araştırmacılar giderek sayısı artmakta olan otomatik makine öğrenmesi yaklaşımlarından hangisini tercih etmesi gerektiğini, bu yöntemlerin kapsamı ve başarısını, hangi verilerde nasıl sonuçlar verdiğini bilmek ve ona göre bir yol haritası belirlemek durumdadırlar.

Kullanılan 29 omik veriden 23'ünü otomatik makine öğrenmesi ile en iyi performansı elde edecek şekilde sınıflandırabildiğimiz için gelecek çalışmalarda araştırmacılar bu yöntemleri omik verilerin analizlerinde tercih edebilirler. Her ne kadar TPOT ve H2O otomatik olarak birçok yöntemi farklı parametreler ile deniyor olsa da bazı verilerde RF, DVM ve NSC uygulamalarıyla daha iyi sonuçlar elde edilmiştir. Kullanılan 29 veriden 11'inde RF, DVM ve NSC yöntemleri ile en iyi performansı elde edildi. Maliyeti AutoML yöntemlerine göre düşük olan bu yöntemler her zaman en iyi sınıflandırma performansına sahip olmasa bile bazı verilerde en iyi sınıflandırma performansına yakın değerler almıştır. Otomatik olmayan sınıflandırma yöntemlerinden RF, DVM ve NSC yöntemleri ile en iyi sınıflandırma performansının elde edildiği 11 verinin 5'inde TPOT yöntemi de en iyi sınıflandırma performansını gösteren yöntemler arasında yer almıştır.

Truong ve ark. (2019) birçok otomatik makine öğrenimi yaklaşımını karşılaştırmıştır. Bu yöntemler arasında Auto-Keras, Auto-sklearn, Darwin, Ludwig, H2O ve TPOT yer almaktadır. Denemiş oldukları veriler için doğruluk oranı ve F1 metriklerini kullanmışlardır. Analizler sonucunda çok sınıflı verilerde H2O yönteminin diğer yöntemlerden daha düşük performanslar gösterdiği görülmüştür. Sonuç olarak Truong ve ark. (2019) birçok sınıflandırma ve regresyon problemlerine ilişkin veriler kullanarak yapmış oldukları çalışma ile çeşitli kriterlere göre H2O, Auto-Keras ve auto-sklearn yöntemlerini Ludwig, Darwin, TPOT ve auto_ml yöntemlerinden daha başarılı bulmuşlardır. Bizim yapmış olduğumuz analizler sonucunda ise H2O ile 17 veride en iyi sınıflandırma performansı elde edilmiştir ve bu verilerden sadece 2 tanesi çok sınıflıdır. Başka bir deyişle kullanılan verilerden 13 tanesi çok sınıflıdır ve bu verilerin sadece 2 tanesinde H2O ile en iyi sınıflandırma performansı elde edilebilmiştir. Truong ve ark. (2019) çalışma sonucuna benzer olarak bizim yapmış olduğumuz çalışmada da H2O yönteminin çok sınıflı verilerde sınıflandırma performansı diğer yöntemlere göre zayıf bulunmuştur. Bunun nedeninin, iki sınıflı verilerde H2O yönteminin yığılmış topluluk (stacked ensembles) sınıflandırma yöntemlerini desteklerken henüz çok sınıflı veriler için bu yöntemlerin kullanılamamakta olmasından ve sadece tekli modellerin denenebiliyor olmasından kaynaklanabileceği düşünülmektedir.

Balaji ve Allen (2018) yılında yapmış oldukları çalışmada sınıflandırma ve regresyon problemleri için otomatik makine öğrenmesi yöntemlerinin performansını değerlendirmişlerdir. Karşılaştırma yaptıkları yöntemler arasında H2O, TPOT, auto_ml ve auto-sklearn yöntemleri yer almaktadır. 57 sınıflandırma 30 regresyon probleminin çözüldüğü bu çalışmanın genel sonucu değerlendirildiğinde sınıflandırma için en iyi performans auto-sklearn yöntemi ile elde edilirken regresyon problemlerinde en iyi performansı gösteren yöntem TPOT olmuştur (Balaji ve Allen, 2018).

Gijsbers ve ark. (2019) otomatik makine öğrenmesi yöntemlerinin doğru bir şekilde karşılaştırılabilmesi için açık kaynak kodlu, genişletilebilir, devam etmekte olan bir sistem oluşturmuşlardır. Bu çalışmada oluşturdukları sistemle 39 veri kullanarak 4 farklı AutoML ve RF yöntemlerini karşılaştırmışlardır. Hesaplama maliyeti 1 ve 4 saat için sınırlandırılarak gerçekleştirilmiş ve zamana göre yöntemlerin performanslarında kayda değer artış görülmezken TPOT yöntemi için bazı analizlerin performans skorunda artışların olduğu görülmüştür. Süre artışıyla Auto-WEKA yönteminin çok sınıflı

verilerde ezberleme gibi durumlardan muzdarip olduğunu kaydeden çalışmacılar bizim çalışmamızda da olduğu gibi diğerlerinden her zaman daha iyi bir sonuç veren yöntemin olmadığını kaydetmişlerdir. Genel itibariyle sonuçları değerlendirdiklerinde iki veri için en iyi sınıflandırma performansı veren yöntem RF olarak bulunurken bazı verilerde AutoML yöntemleri belirgin şekilde daha iyi sonuçlar vermiştir. Diğer verilerde ise AutoML yöntemlerinin RF'den çok az daha iyi sonuçlar verdiğini belirtmişlerdir (Gijsbers ve ark., 2019). Yapmış olduğumuz analizler sonucunda, literatürde farklı veriler için elde edilen sonuçlar ile uyumlu sonuçlar elde ettik. RF yöntemi ile ilgi olarak, mikrodizi verilerinin sınıflandırılması ve gen seçimi için standart bir araç olarak kullanılması önerilmiştir (Díaz-Uriarte ve Alvarez de Andrés, 2006). Kullanmış olduğumuz 16 mikrodizi verisinin 5 tanesinde RF yöntemi en iyi sınıflandırma yöntemleri arasında yer almıştır. RF yöntemi gözlem sayının düşük, değişken sayısının fazla olduğu, dengesiz grup dağılımının olduğu ve çok sınıflı veriler ile başa çıkabilme gibi avantajlara sahip olduğu için RF yönteminin en iyi olduğu bir veri yapısından söz edilememektedir. Çok farklı özellikte veri yapıları ile RF yöntemi en iyi sınıflandırma performansına sahip olabilmektedir. 6 RNA dizileme verisinden 2'sinde RF yöntemi en iyi sınıflandırıcı yöntemler arasında yer almaktadır. Bu verilerin her ikisinde çok sınıflı ve değişken sayısı fazla olan verilerdir. Metabolomik verilerinde RF yöntemi ile en iyi sınıflandırma performansları elde edilememiştir. Metabolomik verileri genel itibariyle AutoML yöntemleri ile en iyi şekilde sınıflandırılmıştır ve en kötü sınıflandırma performansının DVM yönteminde olduğu görülmüştür.

Otomatik makine öğrenimi yaklaşımları ile yapılan her analizde rastgele model araması ve hiperparametre ayarlaması yapıldığı için aynı veri ile aynı parametreler kullanılarak aynı sürede analiz tekrar edildiğinde dahi aynı sonuçlar alınamamaktadır. Yani otomatik makine öğrenme yöntemlerinde tekrarlanabilirlik sorunu bulunmaktadır. Özellikle de TPOT yönteminde aynı veri kullanılarak yapılan analiz sonuçlarında bile farklı sınıflandırma performansları elde edilebilir. RF, DVM ve NSC yöntemlerinde böyle bir durum söz konusu değilken AutoML yöntemlerinin tekrarlanabilirliği düşüktür. TPOT yönteminde bu problemin üstesinden gelinebilmesi için sınıflandırma yaparken arama uzayının tamamını arayacak şekilde sınırlama düzeyi düşük parametrelerle çalıştırılması önerilmektedir. Fakat bu kez de en iyi performansı veren başka bir yöntem daha bulunuyorsa yine farklı modeller en iyi olarak belirlenmiş olacaktır. Yani hesaplama maliyeti artırılsa da tekrarlanabilirlik garantisi bulunmayacaktır.

H2O yöntemi ile analizler esnasında stabil bir şekilde çalışmalar gerçekleştirilememiş ve zaman zaman bazı hatalar oluşmuştur. Yöntemin tekrar başlatılmasıyla bu hatalar çözülmüştür. Balaji ve Allen (2018) tarafından yapılan çalışmada da araştırmacılar H2O ile yapılan analizler esnasında 23 kez sistemin başarısız olduğunu kaydetmiştir. Bu gibi hatalar yeni sürümlerle düzeltilmektedir. TPOT yöntemi ile uygulamalar esnasında ise H2O yönteminde olduğu gibi bir problem ile karşılaşılmamıştır. Bizim çalışmamızda TPOT yöntemi için karşılaşılan problem max_time_mins parametresinin kullanıcılar tarafından yanlış anlaşılabilir durumda olması olmuştur. Bu parametre ile belirtilen süreye ulaşıldığında yöntemin çalışması sonlanmamakta olup gelecek sürümlerde bu durumun düzeltileceği belirtilmektedir. Bu gibi nedenlerden dolayı oldukça yeni olan otomatik makine öğrenmesi yaklaşımlarının stabil hale gelip en iyi versiyonlarına ulaşması için zamana ihtiyaç olduğu düşünülmektedir. Yukarıda bahsetmiş olduğumuz gibi her veri yapısına uygun yöntemi bulamama, otomatik yöntemleri kullanırken de optimize edilmesi gereken durumlar, bu yöntemlerin bazı hatalarının bulunması gibi durumlardan dolayı üstün özelliklerine rağmen AutoML yaklaşımlarıyla analizler halen konuya uzak araştırmacılar için sanıldığı kadar kolay değildir. Her ne kadar bu yöntemleri kullanabilmek için birkaç satır kod kullanılıyor olsa da temel düzeyde kodlama becerisi gerektirmektedir.

AutoML yaklaşımları ile analizler yaparken tek bir yöntem ile yapılan analizin hesaplama maliyeti gibi düşük sonuçlar beklenmesi mantıklı değildir. Araştırmacıların analizlerini gerçekleştirme aşamasında hesaplama maliyetleri açısından problemleri yoksa otomatik makine öğrenmesi yaklaşımlarını denemeleri bu çalışmanın sonuçlarına bakılarak oldukça tavsiye edilen bir durumdur. Çünkü birçok veride AutoML yaklaşımları ile daha iyi performansların elde edilebildiği durumlar görülmüştür. Üstelik bu performanslar bizim denemiş olduğumuz yöntemlerden farklı olan lojistik regresyon, gradiyent artırma makineleri gibi otomatik makine öğrenmesi yaklaşımlarının kendisinin belirlemiş olduğu modeller ile elde edilmiştir. Otomatik makine öğrenmesi yaklaşımlarının en iyi sınıflandırma performansını bulabilmesi için ciddi bir hesaplama maliyeti göze alınması gerektiği unutulmamalıdır. Bu çalışmada AutoML yaklaşımları için toplam hesaplama maliyeti yaklaşık 3625 saatdir. Otomatik olmayan yöntemlerin maliyeti ise yaklaşık 290 saat olup AutoML yaklaşımlarına göre oldukça düşüktür. Araştırmacının parametrelerini kendisinin belirlemiş olduğu bir yöntemi kullanması ile belirlenen yöntem ve parametreler gibi yüzlercesini deneyen,

bazen ön işleme basamaklarını da kapsayan AutoML yaklaşımlarının tek bir yöntemde olduğu gibi çok kısa zamanlarda sonlanması beklemek mantıklı olmayacaktır. Öte yandan bu yöntemin geliştiricileri bu problemleri çözebilmek için bu yöntemlerin kodlarını paralel olarak çalışacak şekilde geliştirmektedirler. TPOT ve H2O yöntemleri de paralel çalışmaya uygun yöntemler olup aldıkları parametreler ile paralel çalışma durumları ayarlanabilmektedir. Araştırmacının bildiği ve uygulayabildiği sınırlı makine öğrenmesi yöntemine ek olarak farklı yöntemlerin performanslarını da görebilmesi için, özellikle zaman kısıtlamasının olmadığı durumlarda AutoML yaklaşımları ile elde edilen sonuçları, en iyi performansı veren modeli görmek araştırmacılara yeni bir bakış açısı kazandırabilir. Üstelik AutoML yöntemleri ile kişilerin kendi denedikleri sonuçlardan daha yüksek performanslı sınıflandırma sonuçları elde edilebilir. AutoML yaklaşımları şuanda veri analiz sürecini baştan sona yönetebilecek, bu alanda uzman biyoistatistikçi birine ihtiyaç olmadan veri analizi yapacak düzeyde gelişmiş olmasa da, makine öğrenmesi kullanacak olan kişilerin deneme yanılma yoluyla uzun uğraşlar ile gerçekleştireceği aşamaları kolaylaştırmaktadır.

Kullanılan değişken sayıları bakımından yöntemleri değerlendirdiğimizde NSC yöntemi en az değişkeni kullanmaktadır. En çok değişken kullanan yöntem TPOT olmuştur. Analizlerin gerçekleştirilmesi aşamasında değişken seçimi, değişken yapılandırması TPOT yönteminde otomatik olarak gerçekleştirildiği için bu yöntem uygulanmadan önce değişken seçimi yapılmamıştır. Bazı verilerde temel bileşenler analizi gibi boyut indirgeyen, bazı verilerde ise polinomiyal değişkenleri de veriye ekleyen TPOT yönteminde zaman zaman değişken seçimi yöntemleri de kullanılmıştır. Bazı verilerde hiçbir tekrarda değişken seçimi uygulanmamış, bazı verilerde ise sadece birkaç tekrarda değişken seçimi uygulanmış olduğu için genel itibariyle bu yöntemin kullanmış olduğu değişken sayısı diğer yöntemlerden daha fazladır. TPOT yöntemi ile yapılan uygulamalarda birden fazla aşama otomatik gerçekleştirilmektedir. Değişken seçimi H2O, RF, DVM yöntemlerinde olduğu gibi TPOT uygulanmadan önce yapılarak daha temiz bir veri ile gerçekleştirilseydi TPOT sonuçlarında değişiklikler olabilirdi. Çünkü genel itibariyle TPOT değişkenlerin yapılandırılması ve seçiminde pasif kalmıştır. Bu yöntemleri kullanacak olan araştırmacılar çok yüksek boyutlu gen ifade verilerine sahipse NSC yöntemini kullanmayı tercih edebilirler. Çünkü hem değişken seçimi gibi karar verilmesi ve uygulanmasında ayrı bir iş yükü gerektiren bir aşama ile uğraşmak zorunda kalınmaz hem de iyi bir sınıflandırma performansı elde edilebilir. Ayrıca NSC

yönteminin hesaplama maliyeti yüksek boyutlu ve çok sınıflı verilerde dahi oldukça düşüktür. Boyutu gen ifade verilerine göre daha düşük olan, kullanmış olduğumuz metabolomik verileri için AutoML yaklaşımları genel itibariyle diğer yöntemlerden daha yüksek performanslar gösterdiği için bu tür verilerde araştırmacılar verilerini özellikle H2O gibi bir yaklaşımla analiz edebilirler.

AutoML yaklaşımları oldukça yeni yaklaşımlar olup her geçen gün yeni özellikler kazanmaktadır. Aynı veride dahi farklı sonuçların elde edilebildiği bu yöntemlerin daha geniş veri çeşitliliği kullanarak denenmesi ile daha fazla uygulama yapılması bu yöntemlerin etkinliğini daha iyi bir şekilde değerlendirebilmeyi sağlayacaktır. Ayrıca simülasyon çalışmaları yapılarak farklı veri boyutlarında, daha farklı heterojenliğe sahip, grup sayısı ve gruplara düşen gözlem sayıları farklı olan, olabildiğince farklı omik veri senaryolarının oluşturulmasıyla daha detaylı değerlendirmeler yapılmalıdır. Gelecek çalışmalarda H2O ve TPOT yöntemleri farklı çalışma süreleri ve farklı parametreler ile daha farklı veriler kullanılarak denenebilir. Ayrıca bu çalışmada kullanılan yöntemlerden farklı AutoML yaklaşımları da omik verilerde denenebilir.

Omik veriler, gerek elde edilen teknolojiye gerekse elde edildiği örneğe göre farklılık gösteren yapılardadır. Bu nedenle omik verileri kendilerine özgü ön işleme aşamalarına ihtiyaç duymaktadır. Verilerin sayısallaştırılması, dönüşümü ve normalleştirilmesi gibi aşamalar omik verilerle doğru bilgi edinebilmek için önemlidir. Omik verilere özgü veri ön işleme aşamalarını otomatikleştiren bir sistem bulunmamaktadır. Farklı omik veri yapıları için biyoinformatik alanına özgü aşamaları otomatikleştiren yeni sistemler geliştirilebilir. AutoML-Omik gibi yeni bir sistem tasarlanarak sadece omik verilere özgü olan aşamaları da otomatik hale getirebilecek, farklı omik veriler için uygulanabilirliğe sahip, çeşitli kontrol parametreleri kullanarak her omik bilimi için ona özgü makine öğrenmesi sistemini inşa edebilecek yeni araçlar geliştirilmesi bu alandaki ihtiyaçlar düşünüldüğünde oldukça faydalı olacaktır. Yeni bir sistem geliştirmekten ziyade mevcut AutoML yöntemlerinin omik verilerde performanslarını iyileştirecek, omik verilere özgü aşamaları bu sistemlerle entegre olarak kullanılabileceği yapılar geliştirilebilir.

Çalışmada kullanılan omik veri setlerine göre, otomatik makine öğrenme yöntemleri standart makine öğrenme yöntemleri kadar başarılı F1, bACC ve MCC değerlerine sahip olduğu tespit edilmiştir. Genel olarak değerlendirildiğinde AutoML yöntemleri

standart makine öğrenme yöntemlerinden daha iyi sınıflandırma performanslarına sahip olduğu söylenebilir. Omik veri yapılarına uygun çok çeşitli senaryolar ile veri benzetim yöntemleri kullanılarak AutoML ve standart makine öğrenmesi yöntemlerinin birbirlerine karşı üstünlükleri daha belirgin olarak ifade edilebilir.

6. KAYNAKLAR

Afanador, N. L., Smolinska, A., Tran, T. N. Blanchet, L. Unsupervised Random Forest: A Tutorial with Case Studies. J Chemometr, 2016; 30**:** 232-241.

Alizadeh, A. A., Eisen, M. B., Davis, R. E., Ma, C., Lossos, I. S., Rosenwald, A., Boldrick, J. C., Sabet, H., Tran, T. Yu, X. J. N. Distinct Types of Diffuse Large B-Cell Lymphoma Identified by Gene Expression Profiling. Nature, 2000; 403**:** 503.

Anders, S. Analysing RNA-Seq Data with the DESeq Package. Mol. Biol., 2010; 43**:** 1-17.

Anders, S. Huber, W. Differential Expression Analysis for Sequence Count Data. Genome Biol., 2010; 11**:** R106.

Armstrong, S. A., Staunton, J. E., Silverman, L. B., Pieters, R., den Boer, M. L., Minden, M. D., Sallan, S. E., Lander, E. S., Golub, T. R. Korsmeyer, S. J. MLL Translocations Specify a Distinct Gene Expression Profile that Distinguishes a Unique Leukemia. Genome Biol, 2001; 30**:** 41.

Bal, S. H. BUDAK, F. Mikroarray Teknolojisi. Uludağ Üniversitesi Tıp Fakültesi Derg, 2012; 38**:** 227-233.

Balaji, A. Allen, A. Benchmarking Automatic Machine Learning Frameworks. CoRR, 2018; abs/1808.06492.

Başaran, E., Aras, S. Cansaran-Duman, D. Genomik,Proteomik, Metabolomik Kavramlarına Genel Bakış ve Uygulama Alanları. Turk Hij Den Biyol Derg, 2010; 67**:** 85-96.

Bhattacharjee, A., Richards, W. G., Staunton, J., Li, C., Monti, S., Vasa, P., Ladd, C., Beheshti, J., Bueno, R. Gillette, M. Classification of Human Lung Carcinomas by mRNA Expression Profiling Reveals Distinct Adenocarcinoma Subclasses. P. Natl. Acad. Sci. USA 2001; 98**:** 13790-13795.

Bittner, M., Meltzer, P., Chen, Y., Jiang, Y., Seftor, E., Hendrix, M., Radmacher, M., Simon, R., Yakhini, Z. Ben-Dor, A. Molecular Classification of Cutaneous Malignant Melanoma by Gene Expression Profiling. Nature, 2000; 406**:** 536.

Braga-Neto, U. Dougherty, E. Bolstered Error Estimation. Pattern Recogn., 2004; 37**:** 1267-1281.

Brazma, A. Vilo, J. Gene Expression Data Analysis. FEBS Lett., 2000; 480**:** 17-24.

Bredel, M., Bredel, C., Juric, D., Harsh, G. R., Vogel, H., Recht, L. D. Sikic, B. Functional Network Analysis Reveals Extended Gliomagenesis Pathway Maps and Three Novel MYC-Interacting Genes in Human Gliomas. Cancer Res., 2005; 65**:** 8679-8689.

Breiman, L. Random Forests. Mach. Learn., 2001; 45**:** 5-32.

Brown, M. P., Grundy, W. N., Lin, D., Cristianini, N., Sugnet, C. W., Furey, T. S., Ares, M., Jr. Haussler, D. Knowledge-Based Analysis of Microarray Gene Expression Data by Using Support Vector Machines. P. Natl. Acad. Sci. USA 2000; 97**:** 262-267.

Buriani, A., Garcia-Bermejo, M. L., Bosisio, E., Xu, Q., Li, H., Dong, X., Simmonds, M. S., Carrara, M., Tejedor, N., Lucio-Cazana, J. Hylands, P. J. Omic Techniques in Systems Biology Approaches to Traditional Chinese Medicine Research: Present and Future. J Ethnopharmacol, 2012; 140**:** 535-44.

Carlberg, C. Molnár, F. Human Epigenomics, Springer, 2018

Celis, J. E., Kruhøffer, M., Gromova, I., Frederiksen, C., Østergaard, M., Thykjaer, T., Gromov, P., Yu, J., Pálsdóttir, H., Magnusson, N. Ørntoft, T. F. Gene Expression Profiling: Monitoring Transcription and Translation Products Using DNA Microarrays and Proteomics. FEBS Lett., 2000; 480**:** 2-16.

Chen, C., Zhang, Z.-M., Ouyang, M.-L., Liu, X., Yi, L., Liang, Y.-Z. Zhang, C.-P. Shrunken Centroids Regularized Discriminant Analysis as a Promising Strategy for Metabolomics Data Exploration. J Chemometr, 2015; 29**:** 154-164.

Chen, X., Cheung, S. T., So, S., Fan, S. T., Barry, C., Higgins, J., Lai, K.-M., Ji, J., Dudoit, S. Ng, I. Gene Expression Patterns in Human Liver Cancers. MBoC, 2002; 13**:** 1929-1939.

Chowdary, D., Lathrop, J., Skelton, J., Curtin, K., Briggs, T., Zhang, Y., Yu, J., Wang, Y. Mazumder, A. Prognostic Gene Exepression Signatures Can Be Measured in Tissues Collected in RNAlater Preservative. J. Mol. Diagn., 2006; 8**:** 31-39.

Cover, T. Hart, P. Nearest Neighbor Pattern Classification. IEEE Trans. Inf. Theory, 1967; 13**:** 21-27.

Çalış, A., Kayapınar, S. Çetinyokuş, T. Veri Madenciliğinde Karar Ağacı Algoritmaları ile Bilgisayar ve İnternet Güvenliği Üzerine Bir Uygulama. Endüstri Mühendisliði Dergisi, 2014.

Dalkılıç, H. Dalkılıç, F. Karar Ağaçları Destekli Vadeli Mevduat Analizi, XVII Akademik Bilişim Konferansı–AB, 2015.

de Souto, M. C. P., Costa, I. G., de Araujo, D. S. A., Ludermir, T. B. Schliep, A. Clustering Cancer Gene Expression Data: A Comparative Study. BMC Bioinformatics, 2008; 9**:** 497-497.

Debnath, M., Prasad, G. B. Bisen, P. S. Molecular Diagnostics: Promises and Possibilities, Springer Science & Business Media, 2010

Deng, Q., Ramsköld, D., Reinius, B. Sandberg, R. Single-Cell RNA-Seq Reveals Dynamic, Random Monoallelic Gene Expression in Mammalian Cells. Science, 2014; 343**:** 193-196.

Díaz-Uriarte, R. Alvarez de Andrés, S. Gene Selection and Classification of Microarray Data Using Random Forest. BMC Bioinformatics, 2006; 7**:** 3.

Dirks, R. A., Stunnenberg, H. G. Marks, H. Genome-Wide Epigenomic Profiling for Biomarker Discovery. Clin. Epigenetics, 2016; 8**:** 122.

Dong, Z. Chen, Y. Transcriptomics: Advances and Approaches. Sci. China Life Sci., 2013; 56**:** 960-967.

Durmuşçelebi, A. RNA-Dizileme Verilerinin Kümelenmesinde Yeni İstatistiksel Yaklaşımlar. Yüksek lisans tezi, Erciyes Üniversitesi Sağlık Bilimleri Enstitüsü, Kayseri, 2019.

Dyrskjøt, L., Thykjaer, T., Kruhøffer, M., Jensen, J. L., Marcussen, N., Hamilton-Dutoit, S., Wolf, H. Ørntoft, T. F. Identifying Distinct Classes of Bladder Carcinoma Using Microarrays. Nat. Genet., 2002; 33**:** 90.

Fahrmann, J. F., Grapov, D., DeFelice, B. C., Taylor, S., Kim, K., Kelly, K., Wikoff, W. R., Pass, H., Rom, W. N. Fiehn, O. Serum Phosphatidylethanolamine Levels Distinguish Benign from Malignant Solitary Pulmonary Nodules and Represent a Potential Diagnostic Biomarker for Lung Cancer. Cancer Biomark., 2016; 16**:** 609-617.

Fahrmann, J. F., Kim, K., DeFelice, B. C., Taylor, S. L., Gandara, D. R., Yoneda, K. Y., Cooke, D. T., Fiehn, O., Kelly, K. Miyamoto, S. Investigation of Metabolomic Blood Biomarkers for Detection of Adenocarcinoma Lung Cancer. Cancer Epidemiol. Biomark. Prev., 2015; 24**:** 1716-1723.

Feurer, M., Klein, A., Eggensperger, K., Springenberg, J., Blum, M. Hutter, F. Efficient and Robust Automated Machine Learning, Advances in Neural Information Processing Systems, 2015. 2962-2970.

Fisher, R. A. The Use of Multiple Measurements in Taxonomic Problems. Ann. Eugen., 1936; 7**:** 179-188.

Fondi, M. Liò, P. Multi –omics and Metabolic Modelling Pipelines: Challenges and Tools for Systems Microbiology. Microbiol. Res., 2015; 171**:** 52-64.

Forecast, C. The Chipping Forecast. Special Supplement. Nat. Genet., 1999; 21.

Garber, M. E., Troyanskaya, O. G., Schluens, K., Petersen, S., Thaesler, Z., Pacyna-Gengelbach, M., Van De Rijn, M., Rosen, G. D., Perou, C. M. Whyte, R. Diversity of Gene Expression in Adenocarcinoma of the Lung. P. Natl. Acad. Sci. USA 2001; 98**:** 13784-13789.

Ge, Y., Wang, D.-Z., Chiu, J.-F., Cristobal, S., Sheehan, D., Silvestre, F., Peng, X., Li, H., Gong, Z., Lam, S. H., Wentao, H., Iwahashi, H., Liu, J., Mei, N., Shi, L., Bruno, M., Foth, H. Teichman, K. Environmental OMICS: Current Status and Future Directions. J. Integr. OMICS, 2013; 3.

Géron, A. Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems, O'Reilly Media, Incorporated, 2019.

Gijsbers, P., LeDell, E., Thomas, J., Poirier, S., Bischl, B. Vanschoren, J., An Open Source AutoML Benchmark. 6th ICML Workshop on Automated Machine Learning. 2019, Long Beach, United States.

Glaab, E. Analysing Functional Genomics Data Using Novel Ensemble, Consensus and Data Fusion Techniques. PhD thesis, University of Nottingham, 2011.

Glass, G. V. Primary, Secondary and Meta-Analysis of Research. LEARN, 1976; 5**:** 3-8.

Goksuluk, D., Zararsiz, G., Korkmaz, S., Eldem, V., Zararsiz, G. E., Ozcetin, E., Ozturk, A. Karaagaoglu, A. E. MLSeq: Machine Learning Interface for RNA-Sequencing Data. Comput. Methods Programs Biomed., 2019; 175**:** 223-231.

Gordon, G. J., Jensen, R. V., Hsiao, L.-L., Gullans, S. R., Blumenstock, J. E., Ramaswamy, S., Richards, W. G., Sugarbaker, D. J. Bueno, R. Translation of Microarray Data into Clinically Relevant Cancer Diagnostic Tests Using Gene Expression Ratios in Lung Cancer and Mesothelioma. Cancer Res., 2002; 62**:** 4963-4967.

Greally, J. M. Jacobs, M. N. In Vitro and in Vivo Testing Methods of Epigenomic Endpoints for Evaluating Endocrine Disruptors. ALTEX, 2013; 30**:** 445-471.

Guyon, I., Bennett, K., Cawley, G., Escalante, H. J., Escalera, S., Tin Kam, H., Macià, N., Ray, B., Saeed, M., Statnikov, A. Viegas, E. Design of the 2015 ChaLearn AutoML Challenge, 2015 International Joint Conference on Neural Networks (IJCNN), 12-17 July 2015 2015. 1-8.

Guyon, I., Sun-Hosoya, L., Boullé, M., Escalante, H. J., Escalera, S., Liu, Z., Jajetic, D., Ray, B., Saeed, M., Sebag, M., Statnikov, A., Tu, W.-W. Viegas, E. Analysis of the AutoML Challenge Series 2015–2018. In: HUTTER, F., KOTTHOFF, L. & VANSCHOREN, J. (eds.) Automated Machine Learning: Methods, Systems, Challenges. Cham: Springer International Publishing, 2019

Handelsman, J. Metagenomics: Application of Genomics to Uncultured Microorganisms. Microbiol. Mol. Biol., 2004; 68**:** 669-685.

Handelsman, J., Rondon, M. R., Brady, S. F., Clardy, J. Goodman, R. M. Molecular Biological Access to the Chemistry of Unknown Soil Microbes: A New Frontier for Natural Products. Chem. Biol., 1998; 5**:** R245-R249.

Hastie, T., Tibshirani, R. Friedman, J. The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Springer Science & Business Media, 2009

Heinemann, J., Mazurie, A., Tokmina-Lukaszewska, M., Beilman, G. Bothner, B. Application of support vector machines to metabolomics experiments with limited replicates. Metabolomics, 2014; 10.

Heller, M. J. DNA microarray technology: devices, systems, and applications. Annu. Rev. Biomed. Eng., 2002; 4**:** 129-153.

http://docs.h2o.ai/h2o/latest-stable/h2o-docs/automl.html. Erişim tarihi: 01.12.2019.

https://github.com/ClimbsRocks/auto_ml. Auto-ml: Automated machine learning for production and analytics. Erişim tarihi: 04.12.2019.

International Human Genome Sequencing, C. Finishing the euchromatic sequence of the human genome. Nature, 2004; 431**:** 931-945.

James, G., Witten, D., Hastie, T. Tibshirani, R. An Introduction to Statistical Learning: with Applications in R, Springer Publishing Company, Incorporated, 2014

Jerez, C. A. The use of genomics, proteomics and other OMICS technologies for the global understanding of biomining microorganisms. Hydrometallurgy, 2008; 94**:** 162-169.

Jin, H., Song, Q. Hu, X. 2019a. Auto-Keras: An Efficient Neural Architecture Search System. Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. Anchorage, AK, USA: ACM.

Jin, H., Song, Q. Hu, X. Auto-keras: An efficient neural architecture search system, Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, 2019b. ACM, 1946-1956.

Kamel, D., Brady, B., Tabchy, A., B Mills, G. Hennessy, B. Proteomic Classification of Breast Cancer. Curr. Drug Targets, 2012; 13**:** 1495-1509.

Karabulut, E. Lojistik Regresyon. Uygulamalı Çok Değişkenli İstatistiksel Yöntemler Ankara: Detay Yayıncılık, 2017

Karahalil, B. Overview of Systems Biology and Omics Technologies. Curr. Med. Chem., 2016; 23**:** 4221-4230.

Khan, J., Wei, J. S., Ringner, M., Saal, L. H., Ladanyi, M., Westermann, F., Berthold, F., Schwab, M., Antonescu, C. R. Peterson, C. Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks. Nat. Med., 2001; 7**:** 673.

Kolodziejczyk, Aleksandra A., Kim, Jong K., Tsang, Jason C. H., Ilicic, T., Henriksson, J., Natarajan, Kedar N., Tuck, Alex C., Gao, X., Bühler, M., Liu, P., Marioni, John C. Teichmann, Sarah A. Single Cell RNA-Sequencing of Pluripotent States Unlocks Modular Transcriptional Variation. Cell Stem Cell, 2015; 17**:** 471-485.

Kuzey, C. Veri madenciliğinde destek vektör makineleri ve karar ağaçları yöntemlerini kullanarak bilgi çalışanlarının kurum performansı üzerine etkisinin ölçülmesi ve bir uygulama. Doktora Tezi, İstanbul Üniversitesi, 2012.

Laiho, P., Kokko, A., Vanharanta, S., Salovaara, R., Sammalkorpi, H., Järvinen, H., Mecklin, J., Karttunen, T., Tuppurainen, K. Davalos, V. Serrated carcinomas form a subclass of colorectal cancer with distinct molecular basis. Oncogene, 2007; 26**:** 312.

Lange, E., Tautenhahn, R., Neumann, S. Gröpl, C. Critical assessment of alignment procedures for LC-MS proteomics and metabolomics measurements. BMC Bioinformatics, 2008; 9**:** 375.

Lapointe, J., Li, C., Higgins, J. P., Van De Rijn, M., Bair, E., Montgomery, K., Ferrari, M., Egevad, L., Rayford, W. Bergerheim, U. Gene expression profiling identifies clinically relevant subtypes of prostate cancer. P. Natl. Acad. Sci. USA 2004; 101**:** 811-816.

Le, T. T., Fu, W. Moore, J. H. Scaling tree-based automated machine learning to biomedical big data with a feature set selector. J. Bioinform., 2019.

Leidinger, P., Backes, C., Deutscher, S., Schmitt, K., Mueller, S. C., Frese, K., Haas, J., Ruprecht, K., Paul, F. Stähler, C. A blood based 12-miRNA signature of Alzheimer disease patients. Genome Biol., 2013; 14**:** R78.

Liu, J. J., Cutler, G., Li, W., Pan, Z., Peng, S., Hoey, T., Chen, L. Ling, X. B. Multiclass cancer classification and biomarker discovery using GA-based algorithms. J. Bioinform., 2005; 21**:** 2691-2697.

MacLean, D., Jones, J. D. Studholme, D. J. Application of'next-generation'sequencing technologies to microbial genetics. Nat. Rev. Microbiol., 2009; 7**:** 287.

Madsen, R., Lundstedt, T. Trygg, J. Chemometrics in metabolomics—a review in human disease diagnosis. Anal. Chim. Acta, 2010; 659**:** 23-33.

Mahadevan, S., Shah, S. L., Marrie, T. J. Slupsky, C. M. Analysis of metabolomic data using support vector machines. Anal. Chem., 2008; 80**:** 7562-7570.

Manzoni, C., Kia, D. A., Vandrovcova, J., Hardy, J., Wood, N. W., Lewis, P. A. Ferrari, R. Genome, transcriptome and proteome: the rise of omics data and their integration in biomedical sciences. Brief Bioinform, 2018; 19**:** 286-302.

Marko-Varga, G. Proteomics Principles and Challenges. Pure Appl. Chem., 2004; 76**:** 829-837.

Meng, J., Cui, X., Rao, M. K., Chen, Y. Huang, Y. Exome-based analysis for RNA epigenome sequencing data. J. Bioinform., 2013; 29**:** 1565-1567.

Miyamoto, S., Taylor, S., Barupal, D., Taguchi, A., Wohlgemuth, G., Wikoff, W., Yoneda, K., Gandara, D., Hanash, S. Kim, K. Systemic metabolomic changes in blood samples of lung cancer patients identified by gas chromatography time-of-flight mass spectrometry. Metabolites, 2015; 5**:** 192-210.

Montgomery, S. B., Sammeth, M., Gutierrez-Arcelus, M., Lach, R. P., Ingle, C., Nisbett, J., Guigo, R. Dermitzakis, E. T. Transcriptome genetics using second generation sequencing in a Caucasian population. Nature, 2010; 464**:** 773.

Morgan, J. N. Sonquist, J. A. Problems in the Analysis of Survey Data, and a Proposal, 1963.

Mortazavi, A., Williams, B. A., McCue, K., Schaeffer, L. Wold, B. Mapping and quantifying mammalian transcriptomes by RNA-Seq. Nat. Methods, 2008; 5**:** 621.

Mukherjee, S. Sharma, N. Intrusion Detection using Naive Bayes Classifier with Feature Reduction. Proc. Technol., 2012; 4**:** 119–128.

Nagalakshmi, U., Wang, Z., Waern, K., Shou, C., Raha, D., Gerstein, M. Snyder, M. The transcriptional landscape of the yeast genome defined by RNA sequencing. Science, 2008; 320**:** 1344-1349.

Novik, K., Nimmrich, I., Genc, B., Maier, S., Piepenbrock, C., Olek, A. Beck, S. Epigenomics: Genome-Wide Study of Methylation Phenomena. Curr. Issues Mol. Biol., 2002; 4**:** 111-128.

Olson, R. S., Bartley, N., Urbanowicz, R. J. Moore, J. H. Evaluation of a Tree-based Pipeline Optimization Tool for Automating Data Science. CoRR, 2016a; abs/1603.06212.

Olson, R. S., Bartley, N., Urbanowicz, R. J. Moore, J. H. 2016b. Evaluation of a Tree-based Pipeline Optimization Tool for Automating Data Science. Proceedings of the Genetic and Evolutionary Computation Conference 2016. Denver, Colorado, USA: ACM.

Parmigiani, G., Garrett, E. S., Irizarry, R. A. Zeger, S. L. The Analysis of Gene Expression Data: An Overview of Methods and Software. New York, NY: Springer New York, 2003

Patti, G. J., Yanes, O. Siuzdak, G. Innovation: Metabolomics: the Apogee of the Omics Trilogy. Nat. Rev. Mol. Cell Bio., 2012; 13**:** 263.

Rätsch, G. A Brief Introduction into Machine Learning, 2004 Tubingen, Germany.

Samuel, A. L. Some Studies in Machine Learning Using the Game of Checkers. II—Recent Progress, 1969

Schaalje, G. B. Fields, P. J. Open-set Nearest Shrunken Centroid Classification. Commun. Stat.-Theory Methods, 2012; 41**:** 638-652.

Schmidt, C. W. Metabolomics: What's Happening Downstream of DNA. Environ. Health Perspect., 2004; 112**:** A410-A415.

Shalev-Shwartz, S. Ben-David, S. Understanding Machine Learning: From Theory to Algorithms, Cambridge, Cambridge University Press, 2014.

Shannon, C. E. A Mathematical Theory of Communication. Bell Syst. Tech. J., 1948; 27**:** 379-423.

Sibanjan Das Cakmak, U. M. Hands-On Automated Machine Learning, Packt Publishing, 2018

Singh, S. P., Janjuha, S., Chaudhuri, S., Reinhardt, S., Kränkel, A., Dietz, S., Eugster, A., Bilgin, H., Korkmaz, S. Zararsız, G. Machine Learning Based Classification of Cells into Chronological Stages Using Single-Cell Transcriptomics. Sci. Rep., 2018; 8**:** 17156.

Stoop, M. P., Coulier, L., Rosenling, T., Shi, S., Smolinska, A. M., Buydens, L., Ampt, K., Stingl, C., Dane, A. Muilwijk, B. Quantitative Proteomics and Metabolomics Analysis of Normal Human Cerebrospinal Fluid Samples. Mol. Cell. Proteom, 2010; 9**:** 2063-2075.

Tharwat, A., Gaber, T., Ibrahim, A. Hassanien, A. E. Linear Discriminant Analysis: A Detailed Tutorial. Ai Commun., 2017; 30**:** 169-190.

Thornton, C., Hutter, F., Hoos, H. H. Leyton-Brown, K. Auto-WEKA: Combined selection and hyperparameter optimization of classification algorithms, Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2013. ACM, 847-855.

Tibshirani, R., Hastie, T., Narasimhan, B. Chu, G. Diagnosis of Multiple Cancer Types by Shrunken Centroids of Gene Expression. P. Natl. Acad. Sci. USA 2002; 99**:** 6567-6572.

Tibshirani, R., Hastie, T., Narasimhan, B. Chu, G. Class Prediction by Nearest Shrunken Centroids, with Applications to DNA Microarrays. J. Stat. Sci., 2003; 18.

Truong, A., Walters, A., Goodsitt, J., Hines, K., Bayan Bruss, C. Farivar, R.Towards Automated Machine Learning: Evaluation and Comparison of AutoML Approaches and Tools. arXiv e-prints [Online]: Available: https://ui.adsabs.harvard.edu/abs/2019arXiv190805557T. Erişim tarihi: 01.08.2019.

Truong, Y., Lin, X. Beecher, C. 2004. Learning a Complex Metabolomic Dataset Using Random Forests and Support Vector Machines. Proceedings of the Tenth ACM SIGKDD İnternational Conference on Knowledge Discovery and Data Mining. Seattle, WA, USA: ACM.

Viant, M. R., Applications of Metabolomics to the Environmental Sciences. Springer, 2009.

Wang, Z., Gerstein, M. Snyder, M. RNA-Seq: A Revolutionary Tool For Transcriptomics. Nat. Rev. Genet., 2009; 10**:** 57.

West, M., Blanchette, C., Dressman, H., Huang, E., Ishida, S., Spang, R., Zuzan, H., Olson, J. A., Marks, J. R. Nevins, J. R. Predicting the Clinical Status of Human Breast Cancer by Using Gene Expression Profiles. P. Natl. Acad. Sci. USA 2001; 98**:** 11462-11467.

Wikoff, W. R., Grapov, D., Fahrmann, J. F., DeFelice, B., Rom, W. N., Pass, H. I., Kim, K., Nguyen, U., Taylor, S. L. Gandara, D. R. Metabolomic Markers of Altered Nucleotide Metabolism in Early Stage Adenocarcinoma. J. Cancer Prev., 2015; 8**:** 410-418.

Witten, D. Classification and Clustering of Sequencing Data Using a Poisson Model. Ann. Appl. Stat., 2012; 5.

Witten, D., Tibshirani, R., Gu, S. G., Fire, A. Lui, W.-O. Ultra-high throughput sequencing-based small RNA discovery and discrete statistical biomarker analysis in a collection of cervical tumours and matched controls. BMC biol., 2010; 8**:** 58.

Witten, D. M. Tibshirani, R. Penalized Classification Using Fisher's Linear Discriminant. J. Royal Stat. Soc., 2011; 73**:** 753-772.

www.tdk.gov.tr. Erişim tarihi: 10.11.2019.

Xie, G., Zhou, B., Zhao, A., Qiu, Y., Zhao, X., Garmire, L., Shvetsov, Y. B., Yu, H., Yen, Y. Jia, W. Lowered Circulating Aspartate is a Metabolic Feature of Human Breast Cancer. Oncotarget, 2015; 6**:** 33369.

Yi, L., Dong, N., Yun, Y., Deng, B., Ren, D., Liu, S. Liang, Y. Chemometric Methods in Data Processing of Mass Spectrometry-Based Metabolomics: A Review. Anal. Chim. Acta, 2016; 914**:** 17-34.

Zararsız, G. Development and Application of Novel Machine Learning Approaches for Rna-Seq Data Classification. PhD thesis, Hacettepe University, 2015.

Zararsız, G., Goksuluk, D., Korkmaz, S., Eldem, V., Zararsiz, G. E., Duru, I. P. Ozturk, A. A Comprehensive Simulation Study on Classification of RNA-Seq Data. PloS one, 2017; 12**:** e0182507.

Zhao, Y.-Y. Lin, R.-C. UPLC–MSE Application in Disease Biomarker Discovery: The Discoveries in Proteomics to Metabolomics. Chem.-Biol. Interact, 2014; 215**:** 7-16.

Zheng, H.-Q., Chiang-Hsieh, Y.-F., Chien, C.-H., Hsu, B.-K. J., Liu, T.-L., Chen, C.-N. N. Chang, W.-C. AlgaePath: comprehensive analysis of metabolic pathways using transcript abundance data from next-generation sequencing in green algae. BMC genomics, 2014; 15**:** 196.

meltem

ORIJINALLIK RAPORU

%4	%3	%1	%
BENZERLIK ENDEKSI	İNTERNET KAYNAKLARI	YAYINLAR	ÖĞRENCI ÖDEVLERI

BIRINCIL KAYNAKLAR

1	www.biyoistatistikkongresi.org İnternet Kaynağı	%2
2	epistasislab.github.io İnternet Kaynağı	%1
3	www.journalagent.com İnternet Kaynağı	%1
4	uludagtipdergisi.org İnternet Kaynağı	<%1
5	www.slideserve.com İnternet Kaynağı	<%1
6	edergi.sdu.edu.tr İnternet Kaynağı	<%1
7	A. Narin, M. Ozer, Y. Isler. "Comparision of classifier performances in diagnosing congestive heart failure using heart rate variability", 2013 21st Signal Processing and Communications Applications Conference (SIU), 2013 Yayın	<%1

ÖZGEÇMİŞ

KİŞİSEL BİLGİLER

Adı, Soyadı: Meltem Ünlüsavuran

Medeni Durumu: Bekar

e-Mail: meltemunlusavuran@gmail.com

Cep Telefon: 05533466131

Adres: Buğdaylı Mah. Susuz Sok. No:7 Kocasinan/Kayseri

EĞİTİM BİLGİLERİ

Yüksek lisans: Erciyes Üniversitesi Sağlık Bilimleri Enstitüsü Biyoistatistik ABD-Biyoistatistik Bölümü

Tez konusu: Omik Verilerinde Otomatik Makine Öğrenimi Algoritmalarının Performansının Değerlendirilmesi

Lisans: Fırat Üniversitesi Biyomühendislik Bölümü

SINAV BİLGİLERİ

ALES (05.05.2019): 75.6238

YÖKDİL (05.11.2017): 65.000

KAZANDIĞI ÖDÜL VE BAŞARILAR

Erciyes Üniversitesi Tıp Fakültesi Gevher Nesibe Araştırma Teşvik Ödülü: 14 Mart 2017 tarihinde yapılan Geleneksel Gevher Nesibe Araştırma Teşvik Ödülü yarışmasında öğrenci kategorisinde "Galaktozemi için Terapotik Yöntemlerin Araştırılması: Arjinin" isimli çalışma ile ikincilik ödülü.

KULLANABİLDİĞİ PROGRAM VE YAZILIMLAR

- R (Programlama Dili) : İyi düzeyde
- Python (Programlama Dili): Orta düzeyde
- C++ (Programlama Dili): Başlangıç düzeyinde
- IBM SPSS: İyi düzeyde
- MedCalc: İyi düzeyde
- Minitab: İyi düzeyde

- NCSS: İyi düzeyde
- Weka: İyi düzeyde
- SQL: Orta düzeyde

ARAŞTIRMA ALANLARI

- Omik Veri Analizi
- Metabolomik
- Biyoistatistik
- Biyoinformatik
- Veri Madenciliği
- Makine Öğrenmesi
- Yapay Zeka

ULUSLARARASI BİLİMSEL TOPLANTILARDA SUNDUĞU BİLDİRİLER

- XVIII. Ulusal ve I. Uluslararası Biyoistatistik Kongresi, "A Statistical Pipeline in Derivation of Composite Reference Ranges in The Clinical Laboratory" 26-29 Ekim 2016 Belek / ANTALYA (Poster Bildiri)
- XVIII. Ulusal ve I. Uluslararası Biyoistatistik Kongresi, "Metabolomik Verilerinin Kümelenmesinde Danışmansız Random Forest Yaklaşımı" 26-29 Ekim 2016 Belek / ANTALYA (Sözlü Bildiri)
- XVIII. Ulusal ve I. Uluslararası Biyoistatistik Kongresi, "Independent Hypothesis Weighting In Differential Expression Analysis Of RNA-Seq Data" 26-29 Ekim 2016 Belek / ANTALYA (Sözlü Bildiri)
- XVIII. Ulusal ve I. Uluslararası Biyoistatistik Kongresi, "Kayseri'de 6-17 Yaş Arasındaki Çocukların Vücut Bileşenlerinin Değerlendirmesinde Hattori Grafiği Uygulaması" 26-29 Ekim 2016 Belek / ANTALYA (Sözlü Bildiri)
- XVIII. Ulusal ve I. Uluslararası Biyoistatistik Kongresi, "fMRI Veri Analizlerinin İstatistiksel Değerlendirmesi ve Dikkat Eksikliği Verisi Uygulaması" 26-29 Ekim 2016 Belek / ANTALYA (Sözlü Bildiri)
- XIX. Ulusal ve II. Uluslararası Biyoistatistik Kongresi, "Metabolomik Verilerin Sınıflandırılmasında Kısmi En Küçük Kareler Ayırma Analizi Yaklaşımı" 25-28 Ekim 2017 Belek / ANTALYA (Sözlü Bildiri)
- XIX. Ulusal ve II. Uluslararası Biyoistatistik Kongresi, "Metabolomik biyobelirteçlerinin tespitinde ANOVA-PCA yaklaşımı" 25-28 Ekim 2017 Belek / ANTALYA (Sözlü Bildiri)
- 20. Ulusal ve 3. Uluslararası Biyoistatistik Kongresi, "H2O Otomatik Makine Öğrenme Algoritmasının Metabolomik Verilerinde Performansının Araştırılması" 26-29 Ekim 2018 GAZİANTEP (Tam Metin Sözlü Bildiri)
- Fourth International Jubilee Congress of General/Family Medicine, "Determining the Effect of Nutrition on Pre- and Academic Achievement in Medical Students" 22-24 Kasım Plovdiv/BULGARİSTAN (Sözlü Bildiri)

- Fourth International Jubilee Congress of General/Family Medicine, "Factors Affecting Susceptibility of University Students for Depressive Disorders" 22-24 Kasım Plovdiv/BULGARİSTAN (Sözlü Bildiri)
- Fourth International Jubilee Congress of General/Family Medicine, "Pubis to Sole Growth Reference Charts for Turkish Children Aged 0-84 Months" 22-24 Kasım Plovdiv/BULGARİSTAN (Sözlü Bildiri)

YAYINLAR

- Büyüktuna SA, Doğan HO, Unlusavuran M, Bakir M, An Evaluation Of The Different Biomarkers To Discriminate Bleeding In Crimean-Congo Hemorrhagic Fever, Ticks and Tick-borne Diseases (2019), https://doi.org/10.1016/j.ttbdis.2019.05.008
- Tekin T, Çiçek B, Konyalıgil N, Gunturk I, Yazici C, Karaca Z, Unlusavuran M, Increased Hip Circumference in Individuals with Metabolic Syndrome Affects Serum Nesfatin-1 Levels, Postgraduate Medical Journal (2019). doi: 10.1136/postgradmedj-2019-136887

KATILDIĞI ULUSAL VE ULUSLARARASI TOPLANTILAR

- 1. International Advisory Board Meeting on "The Current Status and Future Prospective in Scientific Research", 01–03 Eylül 2014, Kayseri
- 1. Yaşam Bilimleri Sempozyumu, Şubat 2016 AGÜ Kayseri
- XVIII. Ulusal ve I. Uluslararası Biyoistatistik Kongresi, 26-29 Ekim 2016 Belek / ANTALYA
- 2. Yaşam Bilimleri Kongresi, Şubat 2017 AGÜ Kayseri
- XIX. Ulusal ve II. Uluslararası Biyoistatistik Kongresi, 25-28 Ekim 2017 Belek / ANTALYA
- Proje sunum çalıştayı, Erciyes Üniversitesi Tıp Fakültesi, 17 Şubat 2018, Kayseri
- Yapay Zeka ve Bulut Bilişim Konferansı, Abdullah Gül Üniversitesi, Mart 2018, Kayseri
- 20. Ulusal ve 3. Uluslararası Biyoistatistik Kongresi, 26-29 Ekim 2018, GAZİANTEP
- Fourth International Jubilee Congress of General/Family Medicine, 22-24 Kasım 2018, Plovdiv/BULGARİSTAN

KATILDIĞI KURSLAR

- Sağlık Bilimlerinde Bilimsel Proje Hazırlama Eğitimi Kursu: THED-ERCİYES-2014
- Statistical Methods for Meta-Analysis, Ege Üni. Biyoistatistik ve Tıbbi Bilişim AD, Çeşme/İZMİR, 2016
- R ile Biyoinformatik Analizler, Belek / ANTALYA, 2016

- R ile Kestirime Yönelik İstatistiksel Modellerin Oluşturulması, Belek/ANTALYA, 2017
- Araştırıcılar İçin Deney Hayvanları Kullanım Sertifikası, Erciyes Üniversitesi Hakan Çetinsaya Deneysel ve Klinik Araştırma Merkezi, 2017
- Deep Learning ile Biyolojik Veri Analizi Nasıl Yapılır?, GAZİANTEP, 20
- Evrimsel Genombilim Uygulamalı Eğitim 2019, Ege Üniversitesi, İzmi Şubat 2019
- Machine Learning by Stanford University on Coursera. Certificate earned at Saturday, April 20, 2019 (https://www.coursera.org/account/accomplishments/verify/TH4ZXLQYSLM9)
- Python ile Derin Öğrenme Eğitimi, Biyoistatistik Derneği Bilişim Yaz Okulu, Erciyes Üniversitesi, Kayseri, 15-16 Haziran 2019
- SQL ile Veritabanı Yönetimi ve Sorguları eğitimi, Biyoistatistik Derneği Bilişim Yaz Okulu, Erciyes Üniversitesi, Kayseri, 17-18 Haziran 2019

VERMİŞ OLDUĞU DERSLER

- Kapadokya Üniversitesi, Tıbbi Dökümantasyon ve Sekreterlik Bölümü, Biyoistatistik dersi (2019-2020 Güz Dönemi)