GEN İFADE VERİ SETLERİNDE BOYUT İNDİRGEME YÖNTEMLERİNİN SINIFLAMA PERFORMANSINA ETKİLERİNİN KARŞILAŞTIRILMASI

Fatma Hilal YAĞIN

BİYOİSTATİSTİK ve TIP BİLİŞİMİ ANABİLİM DALI

Tez Danışmanı

Doç. Dr. Harika Gözde GÖZÜKARA BAĞ

Yüksek Lisans Tezi – 2020

T.C.

İNÖNÜ ÜNİVERSİTESİ

SAĞLIK BİLİMLERİ ENSTİTÜSÜ

GEN İFADE VERİ SETLERİNDE BOYUT İNDİRGEME YÖNTEMLERİNİN SINIFLAMA PERFORMANSINA ETKİLERİNİN KARŞILAŞTIRILMASI

Fatma Hilal YAĞIN

Biyoistatistik ve Tıp Bilişimi Anabilim Dalı

Yüksek Lisans Tezi

Tez Danışmanı

Doç. Dr. Harika Gözde GÖZÜKARA BAĞ

MALATYA

2020

İÇİNDEKİLER

TEŞEKKÜR

Gerçekleştirmiş olduğum tez çalışmasının hayata geçirilmesi sürecinde bilgi ve tecrübelerinden faydalandığım, görüşleriyle beni destekleyen, samimiyetini her zaman hissettiren ve beni doğru yönde yönlendiren danışman hocam Sayın Doç. Dr. Harika Gözde GÖZÜKARA BAĞ' a, akademik eğitimim süresince desteklerini esirgemeyen ve birikimleriyle bana yol gösteren değerli hocalarım Prof. Dr. Saim YOLOĞLU, Prof. Dr. Cemil ÇOLAK ve Dr. Öğr. Üyesi Emek Gündoğan'a ve önerileriyle yardımlarını esirgemeyen aynı anabilim dalında görev yaptığım çok değerli asistan arkadaşlarıma sonsuz saygı ve teşekkürlerimi sunarım. Bu zorlu süreçte destekleriyle beni hiçbir zaman yalnız bırakmayan sevgili eşim Burak YAĞIN'a ve hayatımın her döneminde desteklerini daima kalbimde hissettiren aileme sonsuz teşekkürlerimi sunarım.

ÖZET

Gen İfade Veri Setlerinde Boyut İndirgeme Yöntemlerinin Sınıflama Performansına Etkilerinin Karşılaştırılması

Amaç: Bu çalışmanın amacı, yüksek boyutlu Akut Miyeloid Lösemi (AML) hastalığı gen ifade veri setinde boyut indirgeme yöntemlerinin (LASSO, temel bileşenler analizi (PCA) ve bağımsız bileşenler analizi (ICA)) çeşitli destek vektör makinesi sınıflandırma yöntemlerine etkilerinin karşılaştırılmasıdır.

Materyal ve Metot: Bu çalışmada GEO veri deposunda GDS3057 kodu ile yüklenen Akut miyeloid lösemi (AML: Acute myeloid leukemia) gen ifade veri seti kullanılmıştır. Veri setinde 38 sağlıklı donörden alınan normal hematopoietik hücreler ile 26 AML hastasından gelen lösemik blastlar arasındaki gen ifade profilleri bulunmaktadır. AML veri seti 64 kişi ve 22283 gene ait ifade seviyelerini içermektedir. Veri setine filtreleme işlemi yapıldıktan sonra, LASSO, temel bileşenler analizi(PCA), bağımsız bileşenler analizi (ICA) yöntemleri uygulanarak boyut indirgeme analizleri yapılmıştır. Bu yöntemlerden elde edilen boyutu indirgenmiş veri setlerine Doğrusal, Polinomiyal ve Radyal tabanlı çekirdek fonksiyonlu Destek Vektör Makinesi (DVM) yöntemleri uygulanmıştır. Modelleme analizlerinde yeniden örnekleme yöntemi olarak 10 tekrarlı 10 katlı çapraz geçerlik yöntemi kullanılmıştır. Hiperparametre optimizasyonu için rasgele arama yöntemi kullanılmıştır. Oluşturulan modellerin performansını değerlendirmek için doğru sınıflama oranı, duyarlılık, seçicilik, kesinlik ve F ölçütü değerlerinin ortalamaları verilmiştir. Bu ölçütlere ek olarak boyut indirgeme analizlerinin modelleme süresine etkilerini görebilmek için analiz süreleri de saniye olarak verilmiştir.

Bulgular: Filtreleme işlemi yapıldıktan sonra AML veri setinde 6201 gen kalmıştır. PCA/ICA uygulandıktan sonra AML gen ifade veri setinden 10 bileşen çıkarılmıştır. LASSO uygulandıktan sonra ise veri setinden AML hastalığı için biyobelirteç olabilecek 21 gen seçilmiştir. Kurulan modellerin test verileri için doğruluk oranları sonuçlarına göre veri setine PCA uygulandıktan sonra Polinomiyal çekirdek fonksiyon ile kurulan model en yüksek doğruluk oranını vermiştir. Yapılan analizlerin tümü için Polinomiyal çekirdek fonksiyon ile kurulan DVM modelleri en iyi performansı göstermiştir.

Sonuç: Gen ifade veri setleri ile sınıflandırma modelleri oluşturulmadan önce boyut indirgeme yöntemleri kullanılarak yüksek boyutluluk sorunu giderilmeli, modeller daha sonra kurulmalıdır. Bu sayede analiz süresi kısalır ve modellerin tahmin performansı artar. AML gen ifade veri setinde Polinomiyal çekirdek fonksiyon ile kurulan DVM modelleri, Doğrusal ve Radyal tabanlı çekirdek fonksiyonu ile kurulan DVM modellerine göre daha iyi sonuç vermiştir. Ancak birden fazla veri setinde ve/veya simüle veri setinde bu yöntemleri deneyerek sonuçları karşılaştırmak daha kesin sonuçlara ulaşılması açısından önemlidir.

Anahtar Kelimeler: Boyut İndirgeme, Gen İfade Veri Seti, Özellik Çıkarımı, Özellik Seçimi, Sınıflandırma

ABSTRACT

Comparison of the Effect of Dimension Reduction Methods on Classification Performance in Gene Expression Data Sets

Aim: The aim of this study is to compare the effects of size reduction methods (LASSO, principal components analysis (PCA) and independent components analysis (ICA)) on various support vector machine classification methods in the high-dimensional Acute Myeloid Leukemia (AML) disease gene expression data set.

Material and Method: In this study, Acute myeloid leukemia (AML: Acute myeloid leukemia) data set loaded with GDS3057 code was used in the GEO data warehouse. The data set includes gene expression profiles between normal hematopoietic cells from 38 healthy donors and leukemic blasts from 26 AML patients. The AML data set contains expression levels for 64 people and 22283 genes. After filtering the data set, dimention reduction analyzes were performed by applying LASSO, PCA, ICA, methods. Support Vector Machine (DVM) methods with linear, polynamial and radial based kernel functions were applied to the size-reduced data sets obtained from these methods. In modeling analysis, 10-repeated 10-fold cross validity method was used as the resampling method. Random search method was used for hyperparameter optimization. In order to evaluate the performance of the model, the average accuracy rate, sensitivity, spectivity, precision and F criteria values of 500 replicate samples are given. In addition to these criteria, analysis times are given in seconds to see the effects of size reduction analyzes on modeling time.

Results: After filtering, 6201 genes remained in the AML data set. After applying PCA / ICA, 10 components removed from the AML gene expression dataset. After applying LASSO, 21 genes that could be biomarkers for AML disease selected from the data set. According to the results of the accuracy rates for the test data of the created models, the model established with the polynomial kernel function after applying PCA to the data set gave the highest accuracy rate. The best performance for all analyzes obtained from DVM models with polynamial kernel function.

Conclusion: Before creating classification models with gene expression data sets, the problem of high dimensionality should be eliminated by using dimention reduction methods and models should be established later. In this way, the analysis time is shortened and increases the prediction performance of the models . DVM models with polynamial kernel function in the AML gene expression dataset gave better results than DVM models with linear and radial based kernel function. However, comparing the results by trying these methods in more than one dataset and / or simulated dataset is important for achieving more precise results.

Key words: Dimension Reduction, Gene Expression Data Set, Feature Extraction, Feature Selection, Classification

SİMGELER ve KISALTMALAR DİZİNİ

AML : Akut Miyeloid Lösemi

BDLDA : Blok Çapraz Lineer Diskriminant Analizi

BIRS : En İyi Artımlı Sıralı Altküme

BMU : En İyi Eşleştirme Birimi

CARET : Sınıflandırma ve Regresyon Eğitimi

CART : Sınıflandırma ve Regresyon Ağacı

CFS : Korelasyon Tabanlı Özellik Seçimi

CRISP-DM : Veri Madenciliği için Sektörler Arası Standart Süreç

DVM : Destek Vektör Makinesi

EWUSC : Hata Ağırlıklı, İlişkisiz Küçültülmüş Sentroidler

FA : Faktör Analizi

FLDA : Fisher Doğrusal Diskriminant Analizi

GA : Genetik Algoritma

GEO : Gene Expression Omnibus

GLGS : Gradyan Tabanlı Birini Dışarda Bırakarak Gen Seçimi

GPLS : Genelleştirilmiş Kısmi En Küçük Kareler

ICA : Bağımsız Bileşenler Analizi

k-NN : k En Yakın Komşu

LLE : Yerel Doğrusal Gömme

LOOCSFS : Birini Dışarda Bırakarak Hesaplamalı Sıralı İleri Seçim

LOOCVE : Bir Defaya Mahsus Çapraz Doğrulama Hatası

MDA : Karışım Diskriminant Analizi

MDS : Çok Boyutlu Ölçekleme

MOF : Minimum/Maksimum Oto Korelasyon Faktörleri

mRNA : Haberci RNA

PC : Temel Bileşenler

PCA : Temel Bileşenler Analizi

PDA : Cezalandırılmış Diskriminant Analizi

RBF : Radyal Tabanlı Fonksiyon

mRMR : Minimum Fazlalık Maksimum İlişki

R-SVM	: Özyinelemeli Destek Vektör Makinesi
SA	: Simüle Tavlama
SC	: Küçültülmüş Sentroidler
SFS	: Ardışık İleri Yönde Seçim
SOM	: Kendi Kendini Düzenleyen Harita
SPCA	: Denetimli Temel Bileşenler Analizi
SVM-RFE	: Destek Vektör Makineleri-Özyinelemeli Özellik Eliminasyonu
USC	: İlişkisiz Küçültülmüş Sentroidler

ŞEKİLLER DİZİNİ

Şekil No **Sayfa No**

TABLOLAR DİZİNİ

Tablo No **Sayfa No**

1. GİRİŞ

Günümüzde teknolojinin gelişmesiyle birlikte artan veri boyutu multidisipliner çalışmaları ve alanları doğurmaktadır. Bu alanlardan biri olan biyoinformatik; biyomedikal araştırmalarda vazgeçilmez hale gelmektedir. Daha küçük boyutlu veri setleri ile çalışmalar yapan farklı disiplinlerden araştırmacıların çok daha yüksek boyutlu veri setlerinin bulunduğu biyoinformatik gibi araştırma alanlarına yönelmelerinin iki temel sebebi vardır. Bunlardan birincisi, genetik araştırma tarihinin doğal bir sonucu olarak kabul edilen İnsan Genom Projesi, ikincisi ise aynı anda binlerce gen için gen ifade seviyesinin hızlı ve ekonomik bir şekilde ölçülmesini sağlayan mikrodizilim teknolojisidir (1, 2).

İnsan genlerinin yapısının, organizasyonunun ve fonksiyonunun kapsamlı olarak incelendiği uluslararası bir araştırma programı olan İnsan Genom Projesi ile birlikte Şubat 2001'de tüm genomun üç milyar baz çiftinin %90'ı ve Nisan 2003'de ise %100'üne ilişkin sonuçlar tamamlanmıştır (2, 3). Bu proje sayesinde yapılan araştırmalar, dizilenmiş genomlara ait yüksek boyutlu veri setleri içeren geniş ve büyüyen bir organizma kütüphanesine yol açmıştır (1).

Mikrodizilim teknolojisi ise özellikle kanser olmak üzere birçok hastalık ile ilgili binlerce gen ifade profilini aynı anda analiz etme imkânı sağlamıştır. Bu teknolojinin kullanıldığı kanser araştırmalarında veri madenciliği yöntemleri günden güne önem kazanmaktadır. Mikrodizilim veri setleri yardımıyla gen ifade profilinin sınıflandırılması, biyomedikal araştırmalarda ortak bir çalışma haline gelmektedir. Hastalıklar için etkili olabilecek "biyobelirteç" genlerin belirlenmesi ile tanı ve tedavilerde başarı doğrudan sağlanabilmektedir. Belirlenen biyobelirteçler sayesinde hastalıklar ilerlemeden kişiye özel önleyici tedaviler yapılabilmektedir. Bu sebeple, özellikle kanser araştırmalarının oldukça önemli bir kısmını kanser sınıflandırılması, kanser alt sınıflarının keşfedilmesi ve ilgilenilen kanser türü için biyobelirteç olabilecek en önemli genlerin seçilmesi oluşturmaktadır (4).

Mikrodizilim teknolojisi ile elde edilen gen ifade veri setleri genellikle az sayıda hastaya ait çok sayıda gen bilgisi içermektedir. Veri madenciliği için yüksek boyutlu olarak tanımlanabilecek olan bu veri setleri modelleme aşamasında model performansını düşürmektedir. Bu amaçla gen ifade veri setlerinde sınıflandırma analizleri gerçekleştirilmeden önce özellik seçimi ve/veya özellik çıkarımı yöntemleri ile boyut indirgeme analizleri yapılmalıdır. Hastalık için ayırt edici özelliklere sahip olan genler/bileşenler seçildikten sonra sınıflandırma modelleri oluşturularak hem analiz süresinde kısalma hem de oluşturulan modellerin performanslarının arttırılması sağlanabilmektedir.

Bu çalışmada aşağıda belirtilen özellik seçimi ve özellik çıkarımı yöntemlerinin sınıflama modelinin performansına etkileri incelenecektir. Bu amaçla veri setine gerekli olan ön işleme adımları uygulandıktan sonra; özellik çıkarımı yöntemlerinden temel bileşenler analizi (PCA) ve bağımsız bileşenler analizi (ICA) özellik seçimi yöntemlerinden ise LASSO yöntemi veri setine ayrı ayrı uygulanacaktır. Daha sonra boyutu indirgenmiş veri setlerine Doğrusal, Polinomiyal ve Radyal tabanlı çekirdek fonksiyonlu Destek Vektör Makinesi (DVM) algoritması ile sınıflandırma modelleri kurulacaktır.

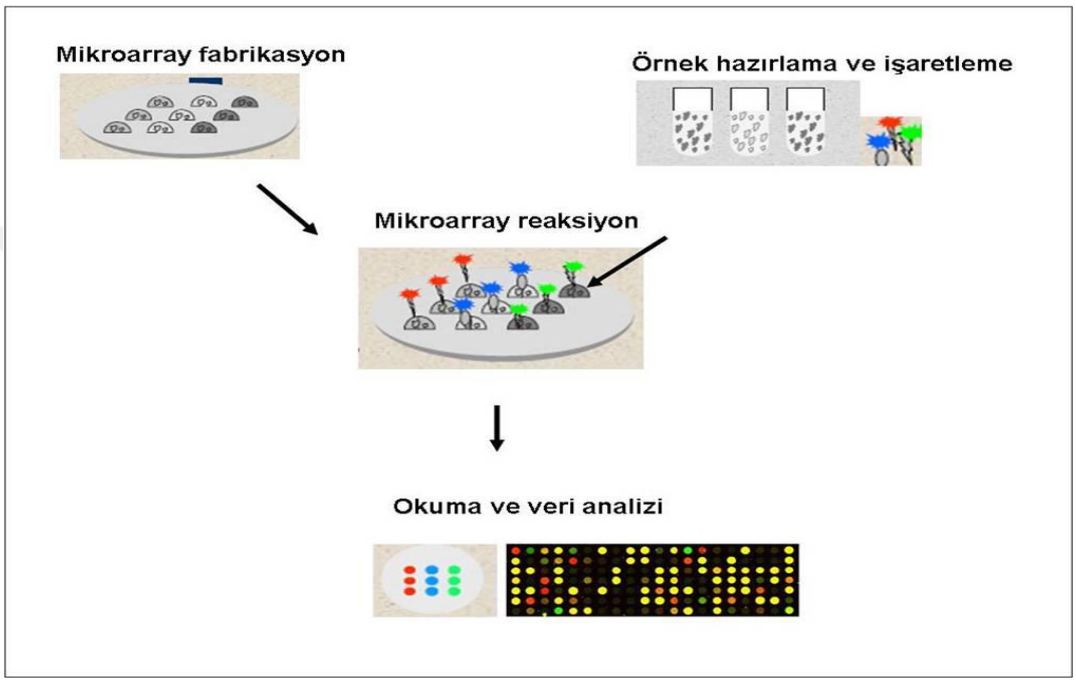
Literatürde gen ifade veri setleri için uygulanan ve iyi sonuçlar veren birçok boyut indirgeme ve sınıflandırma yöntemi bulunmaktadır. Bu yöntemlerden bazıları genel bilgiler kısmında açıklanmaktadır. Ancak, bu çalışmada özellik seçimi olarak LASSO, özellik çıkarımı olarak PCA/ICA ve sınıflandırma algoritması olarak ise Doğrusal, Polinomiyal ve Radyal tabanlı çekirdek fonksiyonlu destek vektör makinesi yöntemleri tercih edildi. Uygulanan bu yöntemler ile AML hastalığı gen ifade veri setinde boyut indirgeme amacıyla özellik çıkarımı ve özellik seçimi yöntemlerinin sınıflandırma modelinin performansını nasıl ve ne derecede etkilediği tartışılacaktır.

2. GENEL BİLGİLER

2.1. Mikrodizilim Teknolojisi

Mikrodizilim teknolojisi özellikle kanser olmak üzere birçok hastalık ile ilgili binlerce gen ifade profilini aynı anda analiz etme imkânı sağlamaktadır. Mikrodizilim teknolojisi ile elde edilen gen ifadesi veri setinde, her bir satır bir hastayı ve her bir sütun belirli bir geni temsil etmektedir. Veri matrisinin her bir girişi belirli bir genin ölçülen ifade düzeyini belirtmektedir. DNA mikrodizilimleri, tek bir mikroskop lamında büyük miktarlarda yüzlerce veya binlerce gen dizisini düzenleyen otomatik sistemlerle yapılmaktadır. Bu amaçla kullanabilecek 40.000'den fazla gene ait veri tabanı bulunmaktadır. Bir gen aktive edildiğinde, hücre o geni transkribe etmeye başlar. Elde edilen ürün, vücudun protein oluşturmaya yönelik şablonu olan haberci RNA (mRNA) olarak bilinir. Bir hücrede hangi genlerin aktive olduğunu hangilerinin kapatıldığını belirlemek için, hücrede bulunan haberci RNA moleküllerinin elde edilmesi gereklidir. Elde edilen mRNA molekülünden cDNA sentezi yapabilmek için revers transkriptaz enzimi kullanılır. Bu işlem sırasında floresan işaretli nükleotitler cDNA'ya bağlanır. Tümörlü ve normal örnekler farklı floresan boyaları ile etiketlenir (Şekil 1.1). Daha sonra, etiketli cDNA bir DNA mikrodizilim slaytına yerleştirilir. Hücrede mRNA'ları temsil eden işaretli cDNA'lar daha sonra, floresan etiketini bırakarak mikrodizilim lamına bağlı sentetik tamamlayıcı DNA'larına hibridize olur veya bağlanır. Daha sonra mikrodizilim slayttaki her nokta/alan için floresan yoğunluğunu ölçmek amacı ile özel bir tarayıcı kullanılır. Belirli bir gen çok aktif ise, birçok haberci RNA molekülü üretir, böylece mikrodizilim slayttaki DNA'ya hibritlenen ve çok parlak bir floresan alanı üreten daha fazla etiketli cDNA'lar üretir. Biraz daha az aktif olan genler daha az mRNA üretir, bu nedenle daha az etiketli cDNA'lar ortaya çıkar ve bu da daha az floresan lekelerle sonuçlanır. Floresan yoksa haberci moleküllerin hiçbiri DNA'ya hibridize edilmemiştir, bu da genin aktif olmadığını gösterir. Bu teknik sıklıkla farklı zamanlardaki çeşitli genlerin aktivitesini incelemek için kullanılır. Tümör numunelerini (kırmızı) ve normal numuneyi (yeşil) birlikte melezleştirirken, bunlar mikrodizilim lamındaki sentetik tamamlayıcı DNA'lar için yarışacaklardır.

Sonuç olarak, eğer nokta kırmızı ise, bu spesifik genin tümörlü dokuda normalden daha fazla (kanserde yukarı regüle edilen) ifade edildiği anlamına gelir. Bir nokta yeşil ise, bu genin normal dokuda daha fazla ifade edildiği anlamına gelir (kanserde aşağı regüle edilen). Bir nokta sarı ise, bu spesifik genin normal ve tümörlü dokuda eşit şekilde ifade edildiği anlamına gelir (4, 5).



Şekil 1.1: Mikrodizilim Süreci

Gen ifade veri setleri az sayıda hastaya ait (örnek) binlerce gen (değişken) bilgisi içermektedir ve yüksek boyutlu veri setleridir. Gen ifade veri setlerinin yüksek boyutlu olması, gen ifade profilinin sınıflandırılmasında sorunlar oluşturmaktadır. Veri setinde hastalık ile ilişkisiz genlerin bulunması, sınıflandırma algoritmalarının performansını da önemli ölçüde etkilemektedir. Bu amaçla araştırmacılar gen ifade veri setlerini sınıflandırmadan önce hastalıkla ilişkili biyobelirteçleri belirleyebilmek amacıyla özellik çıkarımı ve özellik seçimi yöntemlerini sıklıkla kullanmaktadırlar. Bu yöntemlerin amaçları, sınıflandırma performansını arttırmak ve analiz süresini kısaltmak için gen/bileşen sayısını en aza indirmek yani gen ifade veri setinden hastalık ile ilişkili en önemli genleri/bileşenleri en az bilgi kaybı ile seçmektir (6).

2.2. Veri Madenciliği

Veri madenciliği, veri tabanlarında daha önceden bilinmeyen ilişkileri, kalıpları ve eğilimleri bulma ve bulunan bu bilgilerin tahmin edici modeller oluşturmak için kullanma süreci olarak tanımlanabilir (7). Veri madenciliğinde modeller kurulmadan önce verinin kalitesini arttırmak amacıyla veri setini bir önişleme aşamasından geçirmek gerekmektedir. Veri önişleme, veri madenciliği sürecinde önemli bir aşamadır. Veri toplama aşamasında aralık dışı değerler, imkânsız veri kombinasyonları (Örneğin: Cinsiyet: Erkek, Hamile: Evet) ve eksik değerli gözlemler gibi sorunlar ortaya çıkabilmektedir. Dikkatle taranmayan bu tür sorunlar yanıltıcı sonuçlar doğurabilmektedir. Bu nedenle, bir analiz yapılmadan önce verilerin kalitesinin incelenmesi her şeyden önce gelmektedir (8).

Veri setinde çok fazla ilgisiz ve gereksiz bilgi veya gürültülü ve güvenilir olmayan veriler varsa, eğitim aşamasında bilgi keşfi daha zor hale gelmektedir. Veri ön işleme aşaması; veri temizliği, normalleştirme, dönüşüm, özellik çıkarma ve seçimi gibi yöntemleri içermektedir. Veri önişleme, veri analizinin sonuçlarının yorumlanmasını etkilediği için bu aşama dikkate alınmalıdır (9).

Sektörler Arası Standart Süreç olarak adlandırılan CRISP-DM (www.crisp-dm.org) metodolojisi veri madenciliği için işi anlama, veriyi anlama, veriyi hazırlama, modelleme, değerlendirme ve konuşlandırma aşamalarını kapsamaktadır. İş anlayışı CRISP-DM sürecinin oldukça önemli bir aşamasıdır. Bu aşama, yapılacak olan veri madenciliği projesinin iş hedeflerini ve dolayısıyla başarı kriterlerini belirlemektedir. Veriyi anlama veri setinin elde edilmesi, anlaşılması ve kalitesinin belirlenmesi aşamasıdır. Bu aşamada toplanan verilerin kalitesi ilgili yazılımlarda analiz edilir. Böylece elde edilen verilerin yapılacak proje için yeterliliği, eksik ya da hatalı veri içerip içermediği gibi durumlar incelenir. Veriyi hazırlama aşamasında veriler belirli ön işlemlerden geçirilerek modelleme aşaması için hazır ve daha kaliteli duruma getirilmektedir. Verilerin işlenmesi, temizlenmesi, yapılandırılması ve entegrasyonu gibi aşamaları içermektedir. Modelleme aşamasında ise; veri setine ve planlanan projenin hedeflerine bağlı olarak en uygun model oluşturulmaktadır.

Çoğu veri madenciliği yazılımı kümeleme analizi, diskriminant analizi ve regresyon analizi gibi geleneksel istatistiksel yöntemleri; sinir ağları, karar ağaçları, bağlantı analizi ve ilişkilendirme analizi gibi geleneksel olmayan istatistiksel analizleri de içermektedir. Bu oldukça geniş çeşitlilikteki yöntemlerden dolayı, veri madenciliği yapay zekâ - makine öğrenmesi, veri tabanı yönetimi- istatistik ve bilgisayar olmak üzere üç farklı disiplinin bileşimi olarak görülmektedir. Değerlendirme aşaması; projenin sonuçlarının veya kurulan modelin projenin amaçlarına uygun olup olmadığının belirli ölçütler göz önünde bulundurularak değerlendirilmesidir. Son olarak, konuşlandırma aşaması ise, veri madenciliği modellerinin fiili olarak uygulanması ve kullanılmasıyla ilgilidir. Diğer bir deyişle, bu aşamada kurulan modelin günlük hayatta kullanımına odaklanılır. Bu adım için, değerlendirme aşamasında yer alan sonuçlardan yola çıkılarak yapılması gerekenler planlanır.

Veri madenciliği yöntemleri, geniş bir şekilde sınıflandırılmaktadır. Örneğin; tanımlama ve görselleştirme yöntemleri, özellikle büyük veri setleri olmak üzere bir veri setini anlamak ve karmaşık ya da doğrusal olmayan ilişkileri içeren verilerdeki gizli kalıpları tespit etmek için büyük katkı sağlayabilmektedir. Bu yöntemler genellikle modellemeden önce uygulanmaktadır. CRISP-DM metodolojisinde ise veriyi anlama aşamasını temsil etmektedir.

Birliktelik kurallarında ise amaç, hangi değişkenlerin bir araya geleceğini belirlemektir. “Hastalar A tedavisi görür ise, B semptomu sergileyebilme olasılıkları 0.35’dir.” gibi olasılıksal ifadeler üreten bir yöntemi ifade etmektedir. Sağlık hizmetlerinde ilişkisel yapıları araştırmak için yararlı bir yöntemdir. Kümeleme yöntemlerinde ise amaç, nesneleri gruplamaktır. Aynı kümeye ait nesneler benzerdir ve farklı kümelere ait nesneler birbirine benzememektedir.

Veri madenciliğinde en yaygın kullanılan ve önemli uygulamalar muhtemelen tahmin edici modellerdir. Tahmin edici modelleme için, yaygın olarak kullanılan veri madenciliği yöntemleri; çoklu diskriminant analizi ve lojistik regresyon analizi gibi geleneksel istatistiksel yöntemlerin yanı sıra, yapay zekâ ve makine öğrenmesi alanlarında geliştirilen geleneksel olmayan yöntemleri de içermektedir (10).

2.3. Sınıflandırma

Tahmin edici modellerden biri olan sınıflandırma, kategorik olan bir hedef değişkeninin tahminini ifade etmektedir. Sınıflandırma yöntemlerinde bir giriş vektörü, bir sınıf etiketini tahmin etmek için kullanılmaktadır. Sınıflandırma yöntemlerinin amacı, bir gözlemin birbiriyle örtüşmeyen birkaç gruptan hangisine ait olduğunu tahmin edebilmek olduğunda, kullanılan yöntemler sınıflandırma yöntemleri olarak bilinir (11).

2.3.1. Yüksek Boyutlu Gen İfade Veri Setlerinde Sınıflandırma Sorunu

K sınıf sayısı olmak üzere; 1'den K' ya ($K \geq 2$) kadar sınıflardan birine ait n tane hasta olduğunu düşünelim. Y; sınıfsal rasgele değişkenin "sınıf üyeliğini" temsil etmektedir. Örneğin; kanser çalışmalarında Y, tümör sınıfıdır. X_1'den Xp' ye (p: tahmin ediciler) kadar olmak üzere bunlar gen ifade seviyelerini belirtir ve bunlar sürekli değişkenlerdir. Gen ifade veri seti de, n hasta sayısı p gen sayısı olmak üzere $n \times p$ büyüklüğünde bir matristir.

Finansal ve pratik nedenlerden dolayı, mikrodizilim çalışmaları nadiren 200'den fazla deneyi içermektedir. Çoğu zaman binlerce gen ifade seviyesi ölçüldüğü için, genellikle "$n < p$" sorunu ile karşılaşılmaktadır. Aynı zamanda, bir mikrodizilim çalışmasına dâhil edilen genlerin sayısı, dizi analizi alanındaki ilerlemelerden dolayı giderek artmaktadır. Bu nedenle mikrodizilim çalışmalarda gözlem sayısı n genlerin sayısı p'den oldukça küçüktür. Dolayısıyla sınırlı sayıdaki hastaya ait birçok değişkeni ele almak analiz aşamasında sorunlara yol açmakta ve p boyutunu azaltmak için istatistiksel yöntemlere ihtiyaç duyulmaktadır. Mikrodizilim, DNA, proteomik vb. gibi veri setlerinde boyutsallığın azaltılmasının bazı avantajları vardır:

• Çoğu makine öğrenmesi ve veri madenciliği yöntemi yüksek boyutlu veriler için etkili olmayabilir. Bu nedenle boyutsallık azaltıldığında bu yöntemler ile daha etkin sonuçlar elde edilebilmektedir.

• Veri setleri ne kadar karmaşıksa, hesaplama süresi o kadar yüksek olur ve verilerin analiz edilmesi zorlaşır. Bu nedenle boyutsallığın azaltılması hesaplama maliyetini düşürmek için oldukça etkilidir.

• Aşırı uyum sorunundan kaçınmak için etkilidir (12).

Sınıflandırma çerçevesinde yüksek boyutlu verileri kullanmanın üç temel yaklaşımı vardır. İlk yaklaşım, genlerin bir alt kümesini seçmek ve bu küçük gen alt kümesinde klasik bir sınıflandırma yöntemi uygulamaktır. Mikrodizilim literatüründe bu yaklaşım genellikle gen seçimi, gen taraması, özellik seçimi, alt küme seçimi veya gen filtreleme olarak adlandırılır. Mikrodizilim veri analizinde kullanılan klasik sınıflandırma yöntemlerinden bazıları (en yakın komşuluk gibi) açıkça n > p gerektirmez, ancak ilgisiz değişken sayısı çok büyük olduğunda bu durum sınıflandırma performansının düşmesine sebep olur.

Alternatif bir yaklaşım özellik çıkarımıdır. Özellik çıkarımı yöntemleri küçük bir gen alt kümesini seçmek ve diğerlerini elimine etmek yerine, verileri bir anlamda mümkün olduğu kadar özetleyen yeni bileşenler yaratmaktadırlar. Yeni bileşenler daha sonra klasik bir sınıflandırma yöntemi için tahmin edici değişkenler olarak kullanılmaktadır.

Son olarak, kendi içinde özellik seçimi yapan ve herhangi bir ön değişken seçimi gerektirmeyen bir sınıflandırma yöntemi kullanılabilir. Örneğin, sınıflandırma ağaçlarında (CART) değişken seçimi içseldir.

Ancak, mikrodizilim veri setlerinde CART kullanılması, en az üç nedenden dolayı önerilmez. Bunlardan ilki çok yavaş olmasıdır. İkincisi; elde edilen ağaçların verilerdeki küçük değişikliklere karşı çok hassas olmasıdır. Üçüncü sebep ise az sayıda gözlemle elde edilen ağaçların genellikle az sayıda bölünmeye sahip olmasıdır. Diğer bir deyişle, bu yöntem ile muhtemelen önemli olan birçok gen göz ardı edilir. Sınıflandırma ağaçları kapsamında, bagging (13), boosting (14) veya rastgele ormanlar (15) gibi birleştirme yöntemleri çoğu zaman sınıflandırma doğruluğunun gözle görülür şekilde iyileşmesine yol açmaktadır. Ayrıca değişken seçimini kendi içinde uygulayan yöntemler olarak da görülebilirler. Değişken seçimini kendi içinde yapan başka bir yöntem, özellikle mikrodizilim veri setlerini sınıflandırmak için tasarlanmış en yakın sentroid sınıfıdır. Test setindeki her gözlem, en yakın küçülen sentroid sınıfına atanır. Küçültülmüş sentroidler, d; gürültünün sinyal türünün bir istatistiği olmak üzere, yalnızca yüksek d puanlı genler kullanılarak belirlenir. Analizde yer alan genlerin sayısı, d istatistiği için seçilen eşik değerine bağlıdır. Böylece, kendi içinde bir değişken seçimi gerçekleştirilir (16).

2.4. Boyut İndirgeme

Boyut indirgemenin temel amacı veri setinin boyutunu ilgili veri setini en az bilgi kaybı ile açıklayabilecek değişkenlerine (boyutlarına) yoğunlaştırarak azaltmaktır. Bu işlemi yapmanın bir avantajı, daha küçük boyutlu bir veri seti kullanarak analizi hızlandırmaktır. Ayrıca yüksek boyutlu verileri görselleştirmek zordur. Bu nedenle, bir veri setini yalnızca iki veya üç yüksek derecedeki önemli boyuta indirgemek, orijinal veri setinin açıklayıcı özelliklerini kaybetmeden verilerin kolay sunumuna izin vermektedir. Ek olarak analizde modelleme aşamasına geçmeden önce, bir veri setinde hangi boyutların daha belirgin olduğunu bilmek önemlidir (17). Gen ifade veri setlerinde boyut indirgeme amacıyla kullanılan özellik seçimi ve özellik çıkarımı yöntemleri aşağıda açıklanmıştır.

2.4.1. Özellik Seçimi

Özellik seçimi yöntemleri, veri setinden modelleme için ilgisiz veya gereksiz özellikleri kaldırarak çalışmaktadır. Seçilen özellikler bazı objektif fonksiyonlara göre modelleme için en iyi performansı vermelidir (18). İşlenecek verilerin boyutu son 10 yılda oldukça artmıştır ve bu nedenle sınıflandırma yapılmadan önce özellik seçimi bir gereklilik haline gelmiştir. Özellik çıkarma yöntemlerinin aksine, özellik seçme yöntemleri verilerin orijinal sunumunu değiştirmeyen yöntemlerdir (19).

Hem özellik seçimi hem de özellik çıkarma yöntemlerinin ortak bir amacı, daha etkin analizlerin yapılmasını sağlamak için verilerin aşırı uyum göstermesini önlemektir. Özellik seçme algoritmaları üç kategoriye ayrılmaktadır (20):

- Herhangi bir öğrenmeye gerek kalmadan verilerden özellikler çıkaran filtreleme (filter) yöntemleri.
- Hangi özelliklerin yararlı olduğunu değerlendirmek için öğrenme tekniklerini kullanan sarmal (wrapper) yöntemler.
- Özellik seçme adımını ve sınıflandırıcı yapısını birleştiren gömülü (embedded) yöntemler.

Filtreleme yöntemleri, sınıflandırıcıyı dikkate almadan çalışmaktadır. Bu durum hesaplama açısından filtreleme yöntemlerini çok verimli kılmaktadır. Çok değişkenli ve tek değişkenli yöntemler olarak ayrılmaktadırlar. Çok değişkenli yöntemler özellikler arasında ilişki bulabilirken, tek değişkenli yöntemler her özelliği ayrı ayrı ele almaktadır.

Bu yöntemlerde genlerin önemlerine göre sıralanması popüler bir istatistiksel yaklaşımdır. Genleri önemlerine göre sıralamak için aşağıdaki yöntemler önerilmiştir (21).

Koşulsuz Karışım Modellemesi (Unconditional Mixture Modeling), yöntemi belirli bir ifade seviyesinin marjinal olasılığının, iki bileşenli (tek bir bileşenin dejenere durumunu içeren) tek değişkenli bir karışım olarak modellenmesidir. Genin altta yatan ikili durumu iki sınıf arasında değişmezse, gen sınıflandırma problemi için ayırt edici değildir ve veri setinden atılmalıdır (21).

Bilgi Kazancı Sıralaması (Information Gain Rank) yönteminde, Y özelliğini tanımak için gereken bilgi ve X özelliği de kullanılarak Y özelliğini tanımak için gereken bilgi arasındaki farkı ifade eden Bilgi Kazancı (Information Gain) skorunun hesaplanmasında entropi modelinden yararlanılır. Entropi, bir sistemdeki belirsizliğin veya tahmin edilemezliğin ölçüsüdür. Bilgi kazancı simetrik bir ölçüttür. X gözlemlendikten sonra Y hakkında kazanılmış bilgi ile Y gözlemlendikten sonra X hakkında kazanılmış bilgi birbirine eşittir. Bilgi kazancı yönteminin zayıf yanı, daha fazla bilgiye sahip olmasa da çok çeşitli değerlere sahip özellikler lehine önyargılı sonuçlar vermesidir (22).

Saklı Markov Filtreleme (Markov Blanket Filtered) yöntemi sınıf etiketinden bağımsız değişkenleri bulur ve bu değişkenlerin veri setinden kaldırılmasının sınıflama doğruluğunu etkilemediğini varsayar. Çok değişkenli yöntemlerde çift-skorlar; daha iyi bir sınıflandırma sağlamak için birlikte çalışan genleri tanımlamak amacıyla iki sınıfı ne kadar iyi ayırabileceklerine bağlı olarak gen çiftlerini değerlendirmek için kullanılır (23).

Hata Ağırlıklı, İlişkisiz Küçültülmüş Sentroidler (Error-Weighted Uncorrelated Shrunken Centroid - EWUSC) yöntemi ilişkisiz küçültülmüş sentroidler (USC) ve küçültülmüş sentroidler (SC) tabanlıdır. Küçültülmüş sentroidler, her sınıftaki her bir gen için ortalama gen ifadesinin, aynı sınıftaki o gen için standart sapmaya bölünmesiyle bulunur. Bu şekilde, aynı sınıftaki farklı örnekler arasında ifadesi aynı olan genlere daha yüksek ağırlık verilir. Yeni örnekler etikete en yakın ortalama desenle (kare mesafe kullanılarak) atanır.

İlişkisiz küçültülmüş sentroidler yaklaşımı, SC tarafından hâlihazırda bulunan genler kümesinde yüksek derecede korelasyona sahip genler bularak gereksiz genleri veri setinden kaldırır. EWUSC, bu adımların her ikisini de kullanır ve buna ek olarak, (sınıf içi değişkenliğe dayalı olarak) hata ağırlıkları ekler, böylece gürültülü genler indirgenir ve gereksiz genler kaldırılır (24).

Minimum Fazlalık Maksimum İlişki (Minimum Redundancy Maximum Relevance - mRMR) yöntemi sınıf (çıktı) değişkeni ile yüksek korelasyon ve kendi aralarında düşük korelasyon gösteren özellikleri seçme eğiliminde olan bir özellik seçim yaklaşımıdır. Sürekli özellikler için, F-istatistiği sınıf değişkeni ile korelasyonu (ilişki düzeyini), Pearson korelasyon katsayısı ise değişkenler arasındaki korelasyonu hesaplamak için kullanılmaktadır. mRMR, her sınıftaki fazlalığı en aza indirirken, sınıf etiketi ile genlerin ilişki düzeyini en üst düzeye çıkaran bir yöntemdir. Bunu yapmak için çeşitli istatistiksel önlemler kullanır. Bunlardan biri olan karşılıklı bilgi (MI) rastgele bir değişkenin diğeri hakkında (özellikle de gen aktivitesi ve sınıf etiketi) verebileceği bilgileri ölçer. Yöntem hem kategorik hem de sürekli değişkenlere uygulanabilmektedir (25).

Korelasyon Tabanlı Özellik Seçimi (Correlation-based feature selection - CFS) "İyi bir özellik altkümesi, sınıf değişkeni ile yüksek düzeyde ilişkili ancak birbiriyle ilişkisiz olan özelliklerdir" ilkesini izleyen bir yöntemdir. CFS bir alt kümeyi değişkenlerin her birinin ayrı ayrı tahmin yeteneğini ve fazlalık derecelerini dikkate alarak değerlendirir. CFS ve diğer yöntemler arasındaki fark, CFS'de her değişkenin bağımsız olarak bir özellik altkümesi için "sezgisel değer" sağlamasıdır (26). Bundan dolayı bir işlev (sezgisel) verildiğinde algoritma bu işlevin çıktısını en üst düzeye çıkaran seçeneği seçerek sonraki adımlarına karar verebilmektedir.

ReliefF yöntemi de kanser mikrodizilim veri setlerinde yaygın olarak kullanılmaktadır (27). Farklı sınıflar arasında en belirgin olan özellikleri seçen çok değişkenli bir yöntemdir. Tekrar tekrar bir örnek çizer ve bir özelliğin komşularına bakarak o özelliği farklı bir sınıfın komşularından ayırt etmeye yardımcı olan özelliklere en fazla ağırlığı verir (28, 29).

Gen ifade veri setlerinde özellik seçim yöntemi olarak iki adımda bağımsız lojistik regresyon kullanan bir yöntem de önerilmiştir (30). İlk adım, genlerin Pearson korelasyon katsayılarına göre sıralandığı tek değişkenli bir yöntemdir. Üst genler, aşamalı değişken seçim olan ikinci adımda dikkate alınır. Bu, zaten dâhil edilen değişkenlere bağlı olarak, bir seferde tek bir genin dâhil edilmesine (veya hariç tutulmasına) dayanan tek değişkenli bir yöntemdir. Filtreleme yöntemleri genellikle sarmal yöntemlerden daha hızlıdır. Filtreleme yöntemleri bilginin sınıflandırılmasında oldukça etkili olsalar da, bu yöntemler tarafından belirlenen genlerin kanıtlanmış biyolojik bir önemi yoktur. Diğer bir deyişle, yukarıda bahsedilen yöntemlerin hiçbiri, sonuçların gerçekten biyolojik olarak önemli olup olmadığı hakkında bilgi sağlamamıştır. Ek olarak bu yöntemlerin başka bir dezavantajı sınıflandırıcıyı dikkate almamalarıdır. Sınıflandırıcının spesifik tahmin edici yönünü ve sapmalarını göz ardı etmek sınıflandırma doğruluğunu düşürebilmektedir.

Sarmal yöntemler, sınıflandırma doğruluğuna bağlı olarak en iyi tahmin performansını gösteren özellikleri seçtiği için genellikle daha iyi performans gösterirler. Sarmal yöntemlerin dezavantajı, özellik alanı büyüdükçe hesaplama verimsizliğine yol açmalarıdır. Filtreleme yöntemlerinin aksine, özellik bağımlılıklarını algılayabilirler. Sarmal yöntemler deterministik ve rasgele olmak üzere iki kategoriye ayrılmaktadırlar.

Deterministik sarmal yöntemlerden olan ve Whitney tarafından önerilen ardışık ileri yönde seçim (Sequential Forward Selection- SFS) algoritması, temel ve etkin bir özellik seçimi yöntemidir (31). SFS algoritması, boş bir gen alt kümesinden başlar ve değerlendirme fonksiyonunda daha fazla gelişme sağlanamayana kadar genleri sırayla seçer. Veri setindeki herhangi bir özelliğin alt kümeye seçilip seçilmemesinde sınıflandırma performansına olan katkısını dikkate alır. SFS algoritması her tekrarda yalnızca bir özelliği alt kümeye dâhil eder ve sınıflandırma performansında artış olmayana kadar bu işlem devam eder. Bu tekrarların herhangi birinde seçilen özellik sonrasında alt kümeden çıkartılamamaktadır (32).

Kanserli ve kanserli olmayan örnekler arasında sınıflandırma modelleri oluşturmak için destek vektör makineleri ile aşağıdaki özellik seçim yöntemleri mikrodizilim veri setlerinde sıklıkla kullanılmaktadır:

- Gradyan-tabanlı birini dışarda bırakarak gen seçimi (Gradient-based-leave-one-out gene selection (GLGS)) yöntemi önerilmiştir. Bu yöntemde veri setine PCA uygulayarak başlanır. Yeni düşük boyutlu uzayın ölçeklendirme faktörlerine sahip bir vektör, gradyan tabanlı bir algoritma kullanılarak hesaplanır ve optimize edilir. Genler, bir korelasyon faktörüne göre sırayla seçilir (33).

- Birini dışarda bırakarak hesaplamalı sıralı ileri seçim (LOOCSFS) yöntemi ardışık ileri yönde seçim (SFS) yöntemine dayalı ve kanser verileri için çok yaygın kullanılan bir özellik seçim yöntemidir. Başlangıçta boş bir kümeye özellikleri ekler ve bir defaya mahsus bırakılan çapraz doğrulama hatasını hesaplar (34). Destek vektör makineleri (SVM'ler) ve C sınırı kullanan genelleme hatasının neredeyse yansız bir tahmin edicisidir. C sınırı, karar sınırıdır ve alt kümedeki farklı özelliklerin, aynı bir defaya mahsus çapraz doğrulama hatasına (LOOCVE) sahip olması durumunda ek bir kriter olarak kullanılır (35, 36). SFS ayrıca seçilecek alt kümenin boyutuna kısıtlamalar ekleyebilir (37). LOOCSFS'nin genelleme hatasının doğru bir tahmincisi olması beklenirken, GLGS'nin ise yüksek boyutlu veri setleriyle çok iyi ölçeklenmesi beklenir. Hem LOOCSFS hem de GLGS yöntemi için özellik alt kümesindeki genlerin sayısı önceden verilmelidir. Literatürde GLGS'nin LOOCSFS'den daha iyi performans gösterdiği yapılan çalışmalarda belirtilmiştir.

- Literatürde önemli genleri veya biyobelirteçleri seçmek için özyinelemeli destek vektör makinesi (R-SVM) algoritması kullanılmıştır (38). Bu yöntemde her bir genin destek vektör makinesinin minimum hatası için katkı faktörü hesaplanır ve sıralanır. En üst sırada yer alan genler alt küme için seçilir.

Rasgele sarmal yöntemlerden genetik algoritma (GA) ve simüle tavlama (simulated annealing (SA)) algoritması en sık kullanılanlarıdır. Rasgele sarmal yöntemlerden biri olan en iyi artımlı sıralı altküme (BIRS) yöntemi genleri önemlerine ve sınıf etiketlerine göre puanlayan ve daha sonra gereksiz genleri tanımlamak için artımlı sıralama ölçütünü (saklı markov yöntemine dayalı) kullanan bir yöntemdir (39). Genetik algoritmalarla birlikte lineer diskriminant analizi yöntemi de literatürde kullanılmıştır. Bu yöntemde gen alt kümeleri kromozom olarak kullanılır ve her neslin en iyi %10'u öncekilerle birleştirilir. Kromozomun bir kısmı, bir sınıf etiketi için bir genin önemini gösteren ayırt edici katsayıdır (40, 41)

Simüle tavlama yöntemi, mevcut çözümün bazı bölümlerinin daha iyi bir bölüme ait olduğunu varsayarak çalışır. Bundan dolayı objektif işlevi en aza indiren ve küresel optimum'dan kaçınan çözümler arayan komşuları araştırmaya devam eder. Simüle tavlama ve genetik algoritmalara dayalı hibrit yöntemler de kullanılmıştır (42). Bu yöntemlerde genetik bir algoritma, en uygun değişkenleri elde etmek için simüle tavlama yönteminden önce ilk adım olarak çalıştırılır. Her çözüm Fuzzy -Means (bir özelliğin bir kümeyle ne kadar alakalı olduğunu tanımlamak için katsayılar kullanan bir kümeleme algoritması) kullanılarak değerlendirilir (43).

Gömülü yöntemler, sarmal yöntemlerden daha iyi hesaplama yapma eğilimindedir, ancak sınıflandırıcıya bağlı seçimler yaparlar. Bunun nedeni, sınıflandırıcı oluşturulduğunda ve seçim sınıflandırıcının yaptığı hipotezlerden etkilendiğinde optimum gen kümesinin oluşturulmasıdır. İyi bilinen bir gömülü yöntem rastgele ormanlardır. Rasgele ormanlar topluluk sınıflandırıcısıdır. Rastgele ormanlar, en düşük öneme sahip genlerin küçük bir kısmının atılmasıyla yinelemeli olarak oluşturulur (44). En az sayıda özelliğe ve en düşük hataya sahip orman özellik alt kümesi olarak seçilir.

Blok çapraz lineer diskriminant analizi (BDLDA) yöntemi bir hastalıkla sadece az sayıda genin ilişkili olduğunu ve bu nedenle sınıflandırmanın doğru olması için sadece az sayıda genin gerekli olduğunu varsayar. Özellik sayısını sınırlamak için kovaryans matrisi üzerine bir blok çapraz yapı uygular (45).

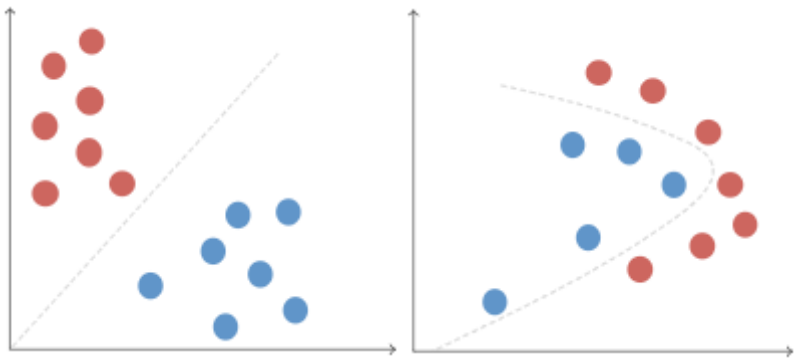
Ek olarak SVM'ler hem özellik seçimi hem de sınıflandırma için kullanılabilir. Sınıflandırmaya katkıda bulunmayan özellikler, sınıflandırmada daha fazla ilerleme sağlanamayana kadar her turda elimine edilir (46). Destek vektör makineleri-özyinelemeli özellik eliminasyonu (SVM-RFE) tüm özelliklerle başlar ve farklı sınıflardaki ayırma örneklerini tanımlamayan özellikleri yavaş yavaş veri setinden dışlar. Bir özellik, mevcut özellik grubu ile SVM'lerin eğitiminden kaynaklanan ağırlığına göre yararlı kabul edilir. Yalnızca “en iyi” özelliklerin seçilme olasılığını artırmak için özelliklerin kaldırılması aşamalı olarak ilerler ve çapraz doğrulama adımlarını içerir (47, 48). SVM-RFE yönteminin en büyük avantajı, belirli bir sınıflandırıcı için yüksek kaliteli özellik alt kümeleri seçebilmesidir. Ancak bu yöntemde, tüm özellikler tek tek geçtiği ve özelliklerin sahip olabileceği herhangi bir korelasyonu hesaba katmadığı için hesaplama açısından zordur (36).

Literatürde SVM-RFE yöntemi sarmal yöntemler olan bir defaya mahsus hesaplama sıralı ileri seçimi (LOOCSFS) ve gradyan esaslı-bir defaya mahsus (GLGS) yöntemleri ile karşılaştırılmıştır. Bu yöntemlerin üçü de Hepatoselüler Karsinom veri kümesine (7129 gen ve 60 örnek) uygulanmış ve benzer hesaplama sürelerine sahip olduğu görülmüştür. GLGS yöntemi, benzer performans hatalarına sahip olan LOOCSFS ve SVM-RFE yöntemlerinden daha iyi performans göstermiştir (49).

2.4.2. Özellik Çıkarma

Özellik çıkarma, ham veri setinin işlenmek üzere daha yönetilebilir gruplara indirgendiği boyut indirgeme sürecidir. Özellik çıkarma, özellikleri birleştiren yöntemlerin adıdır ve bu yöntemler işlenmesi gereken veri miktarını etkili bir şekilde azaltırken, orijinal veri setini hala doğru ve etkin bir şekilde açıklamaktadır. Özellik çıkarma yöntemleri önemli veya ilgili bilgileri kaybetmeden işlem için gereken kaynak sayısını azaltmak gerektiğinde oldukça yararlıdır. Bu yöntemler, belirli bir analiz için gereksiz veri miktarını da azaltabilmektedir. Çok sayıda değişkenle yapılan analizler genellikle büyük miktarda bellek ve hesaplama gücü gerektirir. Ek olarak değişken sayısının çok fazla olması bir sınıflandırma algoritmasının eğitim örneklerine uymasına ve yeni örnekleri zayıf genellemesine neden olabilir. Özellik çıkarma, verileri yeterli doğrulukla açıklarken bu sorunların üstesinden gelmek amacıyla değişkenlerin kombinasyonlarını oluşturma yöntemleri için genel bir terimdir (50).

Özellik çıkarma algoritmaları için doğrusal ve doğrusal olmayan olarak iki temel kategori vardır (51). Doğrusal ve doğrusal olmayan problemler arasındaki fark Şekil 2.1'de gösterilmiştir.



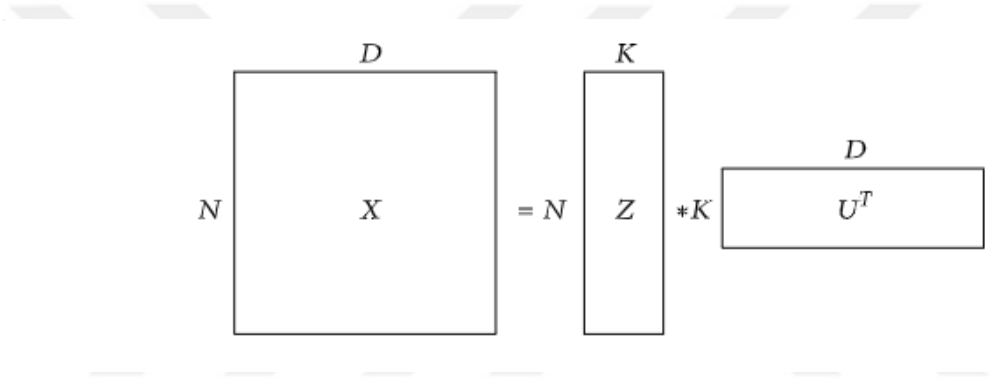
Şekil 2.2: Doğrusal ve Doğrusal Olmayan Sınıflandırma

Doğrusal özellik çıkarma yöntemleri, verilerin daha düşük boyutlu bir doğrusal alt alanda olduğunu varsayar. Matris çarpanlarına ayırma yöntemini kullanarak verileri bu alt alana yansıtır. Doğrusal özellik çıkarma yöntemlerinde;

X: N * D olan veri seti verilsin.

Burada Z: N * K olan bir projeksiyon, Z= X * U ve U: D * K olan bir projeksiyon matrisi vardır.

$U * U^T = I$ kullanıldığında (öz vektörlerin dik özellikleri) $X=Z * U^T$ elde edilir. Grafiksel gösterim Şekil 2.2'de gösterilmektedir.



Şekil 2.3: Doğrusal Matris Çarpanlarına Ayırarak Boyut İndirgeme

En yaygın kullanılan doğrusal özellik çıkarma yöntemlerinden biri temel bileşenler analizidir (PCA). Literatürde PCA ve PCA'nın birçok çeşidi, kanser mikrodizilim verilerindeki boyutsallığı azaltmak için uygulanmıştır (52, 53). Kovaryans matrisi, öz değerleri ve öz vektörleri yardımıyla PCA veri setinde, her biri verilerdeki bir miktar varyasyonu temsil eden ilişkisiz öz vektörler olan "temel bileşenleri" bulur. Bir veri setinin temel bileşenlerini (PC'leri) hesaplarken PC'lerin sınıf değişkeniyle ilişkili olacağının garantisi yoktur (54, 55). Bu nedenle, yapılan bir çalışmada özellikleri sınıf değişkenlerine göre çıkaran denetimli temel bileşenler analizi (SPCA) önerilmiştir. Bu ekstra adıma gen tarama adımı adı verilmektedir.

PCA'nın denetimli versiyonu denetimsiz olandan daha iyi performans gösterse de, denetimli PCA özellikle karmaşık biyolojik sistemlerde yer alan veri setlerinde sıklıkla bulunan doğrusal olmayan ilişkileri yakalayamamaktadır. SPCA aşağıdaki gibi çalışmaktadır.

- Doğrusal, lojistik veya oransal risk modellerini kullanarak bir gen ile sınıf değişkeni arasındaki ilişki ölçüsünü hesaplanır.
- 1. Adımdaki modellerin çapraz geçerliliğini kullanarak sınıf değişkeni ile en çok ilişkili genleri seçilir.
- Sadece seçilen genleri kullanarak temel bileşen puanlarını tahmin edilir.
- 1. Adımdaki modeli kullanarak regresyonu sonuçla uyumlu hale getirilir.

Klasik çok boyutlu ölçekleme yöntemi de (Ana Koordinatlar Analizi olarak da bilinir.) doğrusal özellik çıkarma yöntemlerindendir (56). Çok boyutlu ölçekleme (MDS) yöntemi çok boyutlu verilerdeki “gizli” yapıları keşfetmek için kullanılan grafiksel bir yöntemdir.

PCA da ve diğer koordinasyon tekniklerinde olduğu gibi, MDS de veri setindeki değişkenliği özetlemek için bir dizi korelasyonsuz (dikey) eksen üretir. Her eksen, büyüklüğü o eksende açıklanan varyasyon miktarını gösteren bir özdeğer içerir. Belirli bir öz değerin tüm öz değerlerin toplamına oranı, her eksenin göreceli 'önemini' gösterir. Başarılı bir MDS, giriş verilerindeki varyasyonun % 50'sinden fazlasını açıklayan birkaç (2-3) eksen üretir ve diğer tüm eksenler küçük özdeğerlere sahiptir. Her nesnenin her eksen boyunca bir 'puanı' vardır. Nesne puanları, koordinasyon grafiğindeki nesne koordinatlarını sağlar. MDS grafiğinin yorumlanması basittir. Grafikte birbirine yakın sıralanan nesneler, uzak sıralananlardan daha benzerdir.

MDS çok çeşitli verileri işlemek için uygun olsa da, MDS sonucunda orijinal değişkenlerle ilgili bilgiler kurtarılamaz. Bunun nedeni, MDS'nin orijinal verileri değil, orijinal verilerden türetilen benzerlik matrisini dikkate almasıdır (57).

Doğrusal olmayan özellik çıkarma yöntemleri farklı şekillerde çalışmaktadır. Örneğin, düşük boyutlu bir alana sahip veri seti yüksek boyutlu bir alan üzerine eşlenebilir. Teorik olarak, bir çekirdek fonksiyon özellikleri daha yüksek boyutlu bir alana eşlemek için kullanılmaktadır.

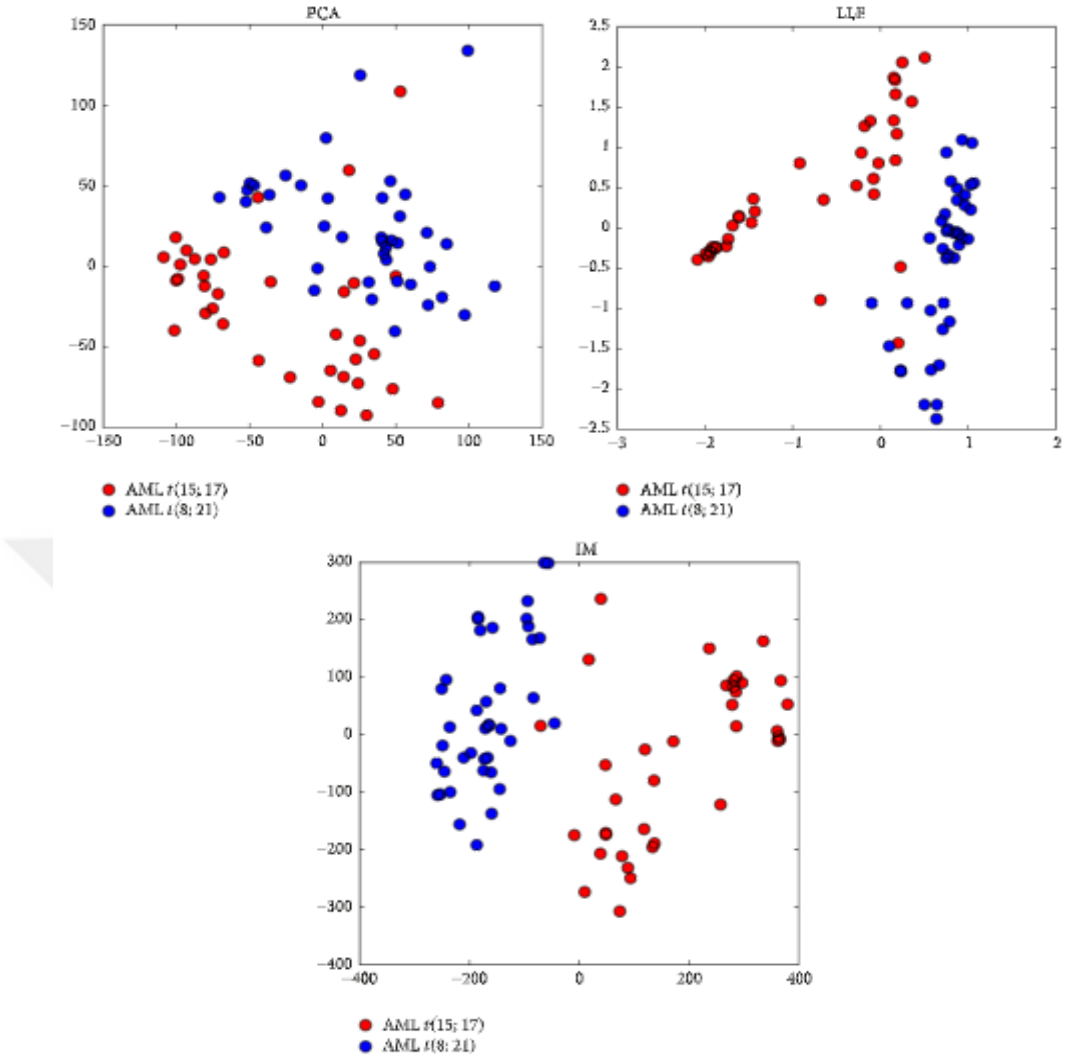
Daha yüksek boyutlu bir alanda özellikler arasındaki ilişki doğrusal olarak görülür ve bu sayede ilişkiler kolayca tespit edilir. Daha sonra veri seti alt boyutlu uzayda haritalandırılır. Uygulamada, çekirdek fonksiyonları, kaldırma fonksiyonunu açıkça hesaplamaya gerek kalmadan aynı etkiyi yaratacak şekilde tasarlanır.

Doğrusal olmayan özellik çıkarma için başka bir yaklaşım manifoldlar kullanmaktır. Bu yaklaşım verilerin (ilgilenilen genler) ham veri uzayından daha düşük bir boyuta sahip olan ve içinde yer alan gömülü doğrusal olmayan bir manifold üzerinde olduğu varsayımına dayanmaktadır. Manifold uzayında çalışan ve mikrodizilim veri setlerine uygulanan çeşitli algoritmalar vardır. Uygun bir manifold bulmak için yaygın olarak kullanılan bir yöntem olan Isomap manifold'u her noktayı yalnızca en yakın komşularına birleştirerek oluşturur (58). Noktalar arasındaki mesafeler daha sonra elde edilen grafikte jeodezik mesafeler olarak alınır.

Literatürde Isomap yönteminin birçok çeşidi kullanılmıştır. Örneğin grafiğin oluşturulma biçiminde farklılık gösteren, ağaca bağlı bir Isomap yöntemi önerilmiştir (59). Bu yöntemde en yakın noktalar, yarı hiper bir küre yardımıyla minimum yayılan ağaç oluşturarak bulunur.

Veri setinde gürültülü ve aykırı değerler söz konusu olduğunda Isomap algoritması oldukça sağlamdır (60). Isomap yöntemi, mikrodizilim veri setlerinde çok iyi sonuçlar vermiştir (60, 61). PCA ile karşılaştırıldığında, Isomap veriler hakkında daha fazla yapısal bilgi elde edilmesini sağlamaktadır.

Ek olarak, mikrodizilim veri setleriyle Yerel Doğrusal Gömme (LLE) (62) ve Laplacian Eigenmaps (63, 64) gibi başka manifold algoritmaları da kullanılmıştır. Şekil 2.3'te gösterildiği gibi veri görselleştirme için PCA ve benzeri manifold yöntemleri de kullanılmaktadır. Veri setleri genellikle manifold LLE ve Isomap kullanılarak daha iyi ayrılabilir, ancak PCA bu iki yöntemden çok daha hızlıdır.



Şekil 2.4: PCA, LLE ve Isomap ile Lösemi veri setinin görüntülenmesi

Özellik çıkarmak için bir diğer doğrusal olmayan yöntem Kernel PCA'dır. Sonuçların yorumlanmasına yardımcı olduğu için yaygın olarak kullanılmaktadır (65, 66). Bu yöntemde, eğitim veri seti matrisin boyutu, veri noktalarının sayısı ile kuadratik olarak arttığı için, alan karmaşıklığı açısından önemli bir sınırlaması vardır (67).

Nöral yöntemler, kendi kendini düzenleyen Haritalar (SOM)(68) veya girdi özelliklerini koruyarak daha düşük boyutlu bir harita oluşturan Kohonen haritaları gibi yöntemlerde boyutsallığı azaltmak için kullanılmaktadır.

Nöral yöntemler düğümler veya nöronlardan oluşurlar ve her düğüm kendi ağırlık vektörü ile ilişkilidir. Bir eğitim örneği ağa bağlandığında tüm düğümlerle öklid mesafesi hesaplandığı ve en küçük mesafe (En İyi Eşleştirme Birimi (BMU)) o düğüme atandığı için SOM eğitimi "rekabetçi" olarak kabul edilir. Bu düğümün komşu düğümleri ile birlikte ağırlığı, girişe uyacak şekilde ayarlanır.

Boyut indirgeme (veya boyut genişletme) amacıyla kullanılan başka bir sinir ağı yöntemi otomatik kodlayıcılardır (otoenkoder). Otomatik kodlayıcılar, verilerin sınıflandırılabileceği bir işleve yaklaşmak üzere eğitilmiş ileri beslemeli sinir ağlarıdır. Her katman için giriş ve çıkış arasındaki fark ölçülür (kare hatası kullanılarak) ve farklı katmanlara ağırlık güncellemeleri yapmak için sinir ağı üzerinden geri yayılır. Ek olarak yığılmış otomatik kodlayıcılar yöntemini PCA yöntemi ile 13 gen ifade veri setinde karşılaştıran bir makalede, otomatik kodlayıcılar veri setlerinin çoğunda daha iyi performans göstermiştir (69). Otomatik kodlayıcılar, parametrelerini ayarlamak için bir geri yayılma yöntemi kullanır. Geri yayılma olmadan, oto-kodlayıcılar çok düşük doğruluklara sahiptir. Yığılmış otomatik kodlayıcılar yöntemiyle ilgili genel bir sorun, çok sayıda dâhili katmanın eğitim verilerini kolayca "ezberlemesi" ve yeni örnek verilerini sınıflandıramayacak bir model oluşturmasıdır. SOM'lar gen ifade verileri için boyutsal azaltma yöntemi olarak kullanılmıştır (64, 70). Ancak iyi performans göstermesi için yeterli miktarda veriye ihtiyaç duyduğundan geniş çapta kabul edilmemiştir.

Bağımsız bileşenler analizi (ICA) mikrodizilim veri setlerin de yaygın olarak kullanılmaktadır (71, 72). Bağımsız bileşenler analizi veriler arasındaki korelasyonu bulur ve kontrast bilgisini en üst düzeye çıkararak veya en aza indirerek verileri düzenler. PCA ile kombinasyon halinde uygulanabilir. Veriler PCA ile analiz edildikten sonra ICA ile analiz edilirse daha iyi sonuçlar alınır (73). Bu durum sadece yüksek boyutun neden olduğu hesaplama yükündeki azalmadan kaynaklanabilir.

2.5. Özellik Çıkarımı ve Özellik Seçimi Arasındaki Fark

Özellik seçimi algoritmaları yalnızca verilerde bulunan boyutlarla çalışır, bu nedenle boyutları sıralaması muhtemeldir ve modelleme için en az yararlı olan özellikleri veri setinden çıkarır. Aksine özellik çıkarma algoritmaları, modelleme için daha yararlı olabilecek yeni boyutları oluşturmak amacıyla mevcut boyutları harmanlamaktadır.

Örneğin, geleneksel PCA, yeni boyutlar kümesinin verilen ilk boyutların doğrusal bir kombinasyonu olması için yeniden yönlendirir ve her boyutun daha az kullanışlı kısımlarını atlar. Bu nedenle özellik çıkarımı özellik seçiminden daha karmaşık olabilir.

Daha somut bir ifadeyle örneğin, çeşitli yerler için aylık ortalama sıcaklıklardan oluşan bir veri setini düşünelim. Tüm veri setine bakarak herhangi bir ayın diğerinden daha önemli olup olmayacağını tahmin etmek oldukça zordur. Bu nedenle bir özellik seçimi algoritması, hangi boyutların tutulacağını seçmekte zorlanırken, bir özellik çıkarımı algoritması her bir boyutun nispeten eşit bir kombinasyonunu alabilir.

Sonuç olarak; özellik çıkarımı ve özellik seçim yöntemleri benzer görevleri farklı şekilde yapmaktadırlar. Özellik çıkarımı orijinal veri setindeki bilginin çoğunu kapsayan ve bu veri setindeki özelliklerin birleşiminden daha az sayıda yeni özellik oluşturan yöntemlerdir (74). Özellik seçimi ise, orijinal veri setindeki özelliklerden en az bilgi kaybı ile bu özelliklerin bir alt kümesini seçen yöntemlerdir (75). Tablo1.1'de özellik seçimi ve özellik çıkarma yöntemlerinin avantajları ve dezavantajları verilmiştir.

Tablo 1.1: Özellik Seçimi ve Özellik Çıkarma Yöntemlerinin Avantajları/Dezavantajları

Yöntem	**Avantajları**	**Dezavantajları**
Özellik Seçimi	Yorumlanabilirlik için veri özellikleri korunur. Aşırı öğrenmeyi azaltır	Ayrım gücü daha düşüktür. Eğitim süresi daha kısadır.
Özellik Çıkarma	Daha yüksek ayrım gücü vardır. Denetimsiz olduğunda aşırı öğrenme kontrol edilebilir.	Veri yorumlanabilirliği kaybı vardır. Dönüşüm işlemleri karmaşıktır.

Teorik temellerin tartışılması, performansa katkıda bulunabilecek veri seti özelliklerinin zenginliğini aydınlatmaktadır. Veri setinin hangi yöntem için daha uygun olduğunun kestirilmesinde; veri setinin boyut sayısı, veri setindeki boyutların sürekliliği ve veri setinin çok değişkenli dağılımını incelemek oldukça faydalıdır.

Bir veri setinin kaç boyuta sahip olduğunu belirlemek kolay değildir, aynı zamanda bu durumun farklı algoritmalarda nasıl farklı performans düzeylerine neden olabileceği kolayca görülebilmektedir. Özellikle, özellik çıkarımı ve özellik seçimi yöntemleri arasında yukarıda belirtilen farkları göz önünde bulunduracak olursak, kurulacak olan modellerin performansı ile bir veri setindeki boyutların sayısı önemli bir etkileşim oluşturabilir. Örneğin, yalnızca 3 boyutlu bir veri seti düşünelim. Bu boyutların her birinde ilgi düzeyi açısından çok az fark varsa, bu boyutların en kötüsünü bile kaldırmak, verilerin sunduğu bilgilerin neredeyse üçte birini kaldırabilir. Ancak, yeni boyutlar oluşturmak için mevcut her boyutun en iyi kısımları kullanılırsa, muhtemelen verilerde yer alan bilgiler üzerinde büyük bir eksiklik yaratmadan bir boyut kaldırılabilir. Şimdi bunun yerine 3000 boyutlu bir veri seti düşünelim. Bu veri setinde eşit derecede ilgili boyutlara sahip olma sorunu olabilir. Ek olarak veri setinden bir boyutun kaldırılması, verilerde yer alan bilgiler üzerindeki bir etkiyi daha az etkiler.

Daha da önemlisi büyük sayılar kanunu, verideki boyutların sayısı arttıkça bu boyutların eldeki görevle eşit şekilde ilgili olduğunu düşündürmektedir. Diğer bir ifadeyle, bu durum özellik çıkarmanın düşük boyutlu veri setleri ile bir avantaja sahip olduğunu, ancak boyutluluğun artmasıyla bu avantajın azalacağını düşündürmektedir.

Bir boyutun sürekliliği, değerlendirilmesi önemli olan bir diğer konudur. Her boyutun kendi bağımsız sürekliliği olduğu için tüm veri setinin genel sürekliliğini değerlendirmek daha zor hale gelir. Boyut sürekliliği için daha spesifik tanımlayıcılar vardır, ancak sadece kesikli ve sürekli boyutlar arasındaki ayrım göz önünde bulundurularak, hem özellik seçimi hem de özellik çıkarma algoritmaları için ilginç çıkarımlar ortaya çıkabilir.

Özellik seçim algoritmaları için, hangi boyutların yüksek ilgi ile sıralandığı ve bu nedenle seçildiği dikkate alınabilir. Bir veri setinin çok sayıda sürekli boyutu varsa, yüksek olasılıkla seçilen boyutlar sürekli olacaktır, ancak yine de kesikli boyutların seçilmesi mümkündür. Bu nedenle, boyutu indirgenmiş veri seti, girdi olarak kullanılan orijinal veri setinden tamamen farklı olabilir. Özellik çıkarma algoritmaları ile bu durum daha karmaşık hale gelebilir.

Özellik çıkarmada yeni boyutlar eski boyutların bir kombinasyonudur. Bu nedenle orijinal veri setindeki her bir boyuttan ne kadarının kullanıldığına ve her bir orijinal boyutun yeni boyut üzerinde hangi düzeyde bir etkiye sahip olduğuna bağlı olarak yeni boyutlar kesikli ve sürekli bir karışımdan oluşabilir.

Özetle, özellik seçim algoritmaları, farklı süreklilik oranlarına sahip çıktı veri setleri oluşturabilirken, özellik çıkarma algoritmaları, tüm girdi veri setindeki herhangi bir girdi boyutundan farklı bir sürekliliğe sahip çıktı boyutları oluşturabilir. Bu potansiyel çıktı aralığı, büyük ölçüde değişen performansa yol açabilir, çünkü herhangi bir karışık veri seti bu şekilde işlendikten sonra çok farklı özelliklere sahip olabilir.

Yukarıda bahsedildiği gibi, PCA yöntemi dikkate alınan verilerin çok değişkenli normal dağılım sergilediğini varsayar. ICA yöntemi ise bu varsayımdan yoksundur. Bu ve benzeri algoritmaların sağladığı performans kazançları, bu varsayımlar sağlanmazsa oldukça olumsuz etkilenirler. Bundan dolayı bir veri seti için en iyi boyut indirgeme yöntemi seçilirken bu varsayımlar dikkate alınmalıdır.

2.6. Sınıflandırma Yöntemleri

Sınıflandırma, kategorik olan bir hedef değişkeninin tahminini ifade etmektedir. Sınıflandırma yöntemlerinde bir girdi vektörü bir sınıf etiketini tahmin etmek için kullanılmaktadır. Sınıflandırma yöntemlerinin amacı, bir gözlemin birbiriyle örtüşmeyen birkaç gruptan hangisine ait olduğunu tahmin edebilmek olduğunda kullanılan yöntemler, sınıflandırma yöntemleri olarak bilinir (11). Mikrodizilim deneyinde gen ifade profilleri için birçok sınıflandırma yöntemi bulunmaktadır. Doğrusal diskriminant analizi, k-en yakın komşu, lojistik regresyon ve logit modellerine uyan genelleştirilmiş kısmi en küçük kare yöntemi, sınıflandırma ve regresyon ağacı, torbalama, artırma, logit artırma ve rasgele orman gibi sınıflandırma ağaçları, topluluk sınıflandırıcılar, destek vektör makinesi yaklaşımları, cezalandırılmış diskriminant analizi ve karışım diskriminant analizi gibi bazı genelleştirilmiş algoritmalar gen ifade veri setlerinde sıklıkla kullanılmaktadır. Bu yöntemlerin her biri aşağıda açıklanmaktadır.

2.6.1. Fisher Doğrusal Diskriminant Analizi (FLDA)

Yüksek boyutlu verileri bir hatta yansıtan ve bu tek boyutlu alanda sınıflandırma yapan bir sınıflandırma yöntemidir. İzdüşüm, iki sınıfın arasındaki mesafeyi en üst düzeye çıkarırken, her sınıftaki varyansı en aza indirir. Bu kriteri en üst düzeye çıkarmak, kovaryans benzeri bir matrisin tersini içeren kapalı bir form verir.

FLDA yöntemi için gözlemlerin normal (Gauss) dağılması ve "eşit grup kovaryansı" varsayımı bulunmaktadır. Ayrıca, değişkenler birbirlerinin doğrusal ifadelerini oluşturamazlar. Bir başka ifadeyle, mükemmel bir şekilde ilişkilendirilemeyebilirler (76, 77).

2.6.2. Lojistik Regresyon

Lojistik regresyon, iki veya çok sınıflı sınıflandırma problemi için denetimli bir yöntemdir. Çeşitli tahmin edici değişkenler (kategorik veya sürekli) ve iki ya da çok sınıflı bir sonuç değişkeni arasındaki ilişkiyi değerlendirmek için kullanılan istatistiksel bir yöntemdir (78).

2.6.3. Genelleştirilmiş Kısmi En Küçük Kareler (GPLS)

Genelleştirilmiş kısmi en küçük kareler yaklaşımlarını Ding ve Gentleman ilk olarak 2003'te uygulamışlardır. İki ve çok gruplu sınıflandırma problemleri için yanlılık azaltma prosedürü seçeneğiyle ayrı ayrı tüm C sınıflarına ve taban çizgisi sınıfına karşı logit modelleri ayarlayan bir yöntemdir. GPLS, bir tür ağırlıklı kısmi en küçük kareler yöntemi olarak kabul edilebilir (79).

2.6.4. k En Yakın Komşu (k-NN)

k-NN, basit bir yöntem olmakla birlikte birçok durumda etkili bir sınıflandırma yöntemidir. Bir veri kaydının sınıflandırılması için en yakın komşuları alınır ve bu, t'nin bir mahallesini oluşturur. Mahalledeki veri kayıtları arasında çoğunluk oylaması genellikle mesafeye dayalı ağırlıklandırma olsun veya olmasın t sınıflamasına karar vermek için kullanılır. Bununla birlikte, k-NN'in uygulamasında k için uygun bir değer seçilmesi gerekir ve sınıflandırmanın başarısı bu değere çok bağlıdır. Bir anlamda, k-NN yöntemi k değeri açısından yanlıdır. K değerini seçmenin birçok yolu vardır, ancak en basit yöntem, algoritmayı farklı k değerleri ile birçok kez çalıştırmak ve en iyi performansı veren k değerini seçmektir (80).

2.6.5. CART ve Topluluk Sınıflandırma Algoritmaları

Sınıflandırma ve regresyon ağacı (CART), geleneksel sınıflandırma yöntemlerinden farklı olarak bir ağaç oluşturma tekniğidir (81). CART analizi ikili yinelemeli bölümleme şeklinde ağaç oluşturur. Topluluk öğrenme ise, sınıf tahmininin doğruluğunu artırmak için sınıflandırıcıları birleştirmek anlamına gelmektedir.

Topluluk öğrenme yöntemlerinden olan rastgele ormanlar, her bir ağacın bağımsız olarak örneklenen rastgele bir vektörün değerlerine bağlı olacağı ve ormandaki tüm ağaçlar için dağılımların aynı olacağı şekilde tahmin edicilerin bir kombinasyonudur. Rastgele ormanlar aykırı değerlere ve gürültülere karşı oldukça sağlamdır ve torbalama veya artırma yöntemlerinden daha hızlıdır (81).

2.6.6. Cezalandırılmış Diskriminant Analizi (PDA)

PDA, bir tür cezalandırılmış lineer diskriminant analizidir (LDA). Yüksek derecede korelasyonlu birçok tahmin edici bulunan durumlar için tasarlanmıştır (82).

2.6.7. Karışım Diskriminant Analizi (MDA)

MDA, gözlemlenen her sınıfın gözlemlenmemiş alt sınıfların bir karışımı olduğu varsayılarak genelleştirilmiş bir LDA'dır. MDA, sınıflandırma problemlerine kümelenmeyi genelleyen öğrenme vektörü nicelemenin (LVQ) düzgün bir versiyonu olarak görülebilir (83).

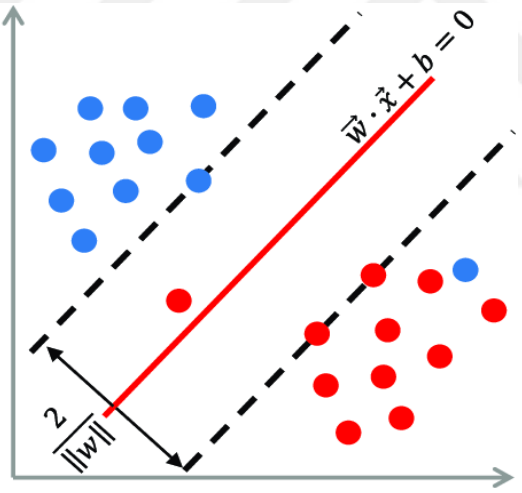
2.6.8. Destek Vektör Makinesi (DVM)

Destek vektör makinesi, istatistiksel öğrenme teorisinden yapısal risk minimizasyon prensibine dayanmaktadır. Regresyon, sınıflandırma ve yoğunluk tahmin problemlerinde uygulanabilmektedir. Yapısal riski en aza indirme fikri, en düşük hata olasılığını garanti edebileceği bir hipotez bulmaktır. Doğrusal olarak ayrılabilen iki sınıflı öğrenme görevi için, DVM'nin amacı verilen örneklerin iki sınıfı maksimal bir marj ile ayırabilen bir hiperdüzlem bulmaktır, bu tür bir hiperdüzlemin en iyi genelleştirme performansını sunabildiği ispatlanmıştır (84). DVM, veri setinin doğrusal ya da doğrusal olmayan şekilde sınıflandırılmasına göre iki durumda incelenmektedir (18).

Doğrusal Destek Vektör Makineleriyle sınıflandırmada, iki sınıfa ait örnekler doğrusal olarak dağılmaktadır. Bu durumda bu iki sınıfın, eğitim veri seti kullanılarak elde edilen bir karar fonksiyonu yardımıyla birbirinden ayrılması amaçlanır. Burada veri setini ikiye ayıran doğru karar doğrusu olarak adlandırılmaktadır.

Sonsuz sayıda karar doğrusu çizebilme imkanı olsa bile önemli olan optimal yani en uygun karar doğrusunu belirlemektir. Karar doğrusunun yeni elde edilecek olan veriye karşı güçlü olabilmesi için sınır çizgisinin, iki sınıfın sınır çizgilerine en yakın mesafede olması gerekmektedir.

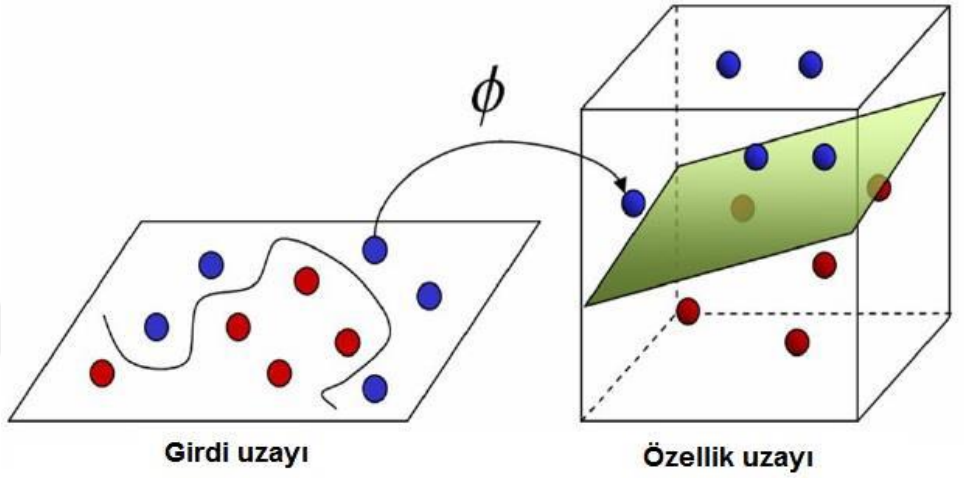
Bu sınır çizgisine en yakın noktalar, destek noktaları olarak adlandırılmaktadır. Destek vektör makineleriyle sınıflandırmada genellikle (-1,+1) şeklinde sınıf etiketleri kullanılmaktadır (85).



Şekil 2.5: Doğrusal DVM modeline ilişkin grafiksel gösterim

Doğrusal olmayan DVM yönteminde, değişik yapıdaki çekirdek fonksiyonları maksimum aralıklı hiperdüzleme uygulanarak doğrusal olmayan sınıflandırıcılar elde edilmektedir. Ortaya çıkan algoritma, doğrusal DVM ile benzerdir; ancak, her iç çarpımı doğrusal olmayan bir çekirdek fonksiyonu ile değiştirilmiştir. Böylelikle algoritma, maksimum aralıklı hiperdüzlemin dönüştürülmüş örnek uzayına yerleşmesine izin vermektedir. Söz konusu dönüşüm, doğrusal yapıda olmayabilir ve dönüştürülmüş örnek uzayı ise yüksek boyutlu olabilir.

İncelenen sınıflandırıcı dönüştürülmüş örnek uzayında bir hiperdüzlem olmasına rağmen, orijinal girdi uzayında doğrusal olmayabilir (86). Doğrusal olarak sınıflandırılamayan girdi uzayının bir üst boyuta çekirdek fonksiyonu ile haritalanarak doğrusal olarak sınıflandırılması Şekil 2.5'de gösterilmektedir (87).



Şekil 2.6: Doğrusal olarak sınıflandırılamayan girdi uzayının bir üst boyuta çekirdek fonksiyonu ile haritalanması

Bu araştırmada değişik destek vektör makinesi modellerini oluşturmak için aşağıda tanımlanan çekirdek fonksiyonlardan, Doğrusal, Radyal tabanlı ve Polinomiyal çekirdek fonksiyonları kullanılmıştır. Çekirdek fonksiyonları sınıflama ve regresyon problemlerinde bir benzerlik fonksiyonu olarak görev yapmaktadır. Bu fonksiyonlar veri madenciliği algoritmalarına çeşitli esnek özellikler kazandırır. Sınıflama ve regresyon analizlerinde bu esnek özellikleri nedeni ile daha yüksek performans sonuçları elde edilebilmektedir (88). Bir örnek ile çekirdek fonksiyonu açıklanacak olursa;

$$X = (x_1, x_2), Y = (y_1, y_2)\ ve\ k(x, y) = x^T + c$$

çekirdek fonksiyonu olduğunda,

$$k(x, y) = x^T + c = \left(\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} [y_1 \quad y_2]\right) + c = (x_1 y_1 + x_2 y_2) + c = \emptyset(x)^T \emptyset(y)$$

olarak elde edilir.

Doğrusal çekirdek, en basit çekirdek fonksiyonudur. <x, y> iç çarpımı ve bir c sabiti ile belirtilir. İlgili çekirdek fonksiyonu aşağıdaki formül ile ifade edilir (88).

$$k(x, y) = x^T + c$$

Polinomiyal çekirdek fonksiyonu durağan olmayan bir çekirdektir. Polinomiyal çekirdeği, tüm eğitim verilerinin normalleştirildiği problemler için çok uygundur. İlgili çekirdek fonksiyonu aşağıdaki formül ile ifade edilir (88).

$$k(x, y) = (\alpha x^T + c)^d$$

Ayarlanabilir parametrelerden, alfa eğimi, c sabit terimi ve d polinom derecesini ifade etmektedir.

Gauss çekirdeği Radyal taban fonksiyonu çekirdeğine bir örnektir. İlgili çekirdek fonksiyonu aşağıdaki formül ile ifade edilir.

$$k(x, y) = exp(-\gamma \|x - y\|^2)$$

Ayarlanabilir parametre sigma, çekirdeğin performansında önemli bir rol oynar ve eldeki soruna göre dikkatle seçilmelidir. Aşırı tahmin söz konusu olduğunda, üstel ifade yaklaşık olarak doğrusallaşır ve yüksek boyutlu projeksiyon doğrusal olmayan yapısını kaybetmeye başlar. Öte yandan, eğer gerçek değerden daha düşük tahmin gerçekleşiyorsa, fonksiyon düzgünleştirmeyecek ve karar sınırı ise eğitim verisindeki gürültü değerlere karşı son derece duyarlı olacaktır (88).

Laplace çekirdeği bütünüyle üstel çekirdek fonksiyonuna eşdeğerdir; ancak sigma parametresindeki olası değişikliklere karşı daha az bir duyarlılık söz konusudur. Ayrıca, Laplace çekirdek fonksiyonu, Radyal tabanlı fonksiyon çekirdeği ile eşdeğerdir. İlgili çekirdek fonksiyonu aşağıdaki formül ile ifade edilir:

$$k(x, y) = exp\left(-\frac{\|x-y\|}{\sigma}\right)$$

Gauss çekirdeği için sigma parametresi hakkında yapılan gözlemlerin üstel ve Laplace çekirdeği için de geçerli olduğu gözlemlenmiştir (88).

ANOVA Radyal tabanlı çekirdek fonksiyonu, Gauss ve Laplace çekirdeği gibi Radyal tabanlı fonksiyon çekirdeğidir. Çok boyutlu regresyon problemlerinde iyi performans gösterdiği bildirilmektedir. İlgili çekirdek fonksiyonu aşağıdaki formül ile ifade edilir (88).

$$k(x,y) = \sum_{k=1}^{n} exp\,(-\sigma(x^k - y^k)^2)^d$$

Bessel çekirdeği ise, aşağıdaki denklem ile ifade edilir:

$$k(x,y) = \frac{J_{v+1}(\sigma\|x-y\|)}{\|x-y\|^{-n(v+1)}}$$

Bu denklemde J birinci tip Bessel fonksiyonudur. Bununla birlikte, kernlab R kaynağında, Bessel çekirdeğinin şu şekilde gösterildiği bildirilmektedir(88).

$$k(x,x^l) = -Bessel^{n}_{(nu+1)}(\sigma|x - x^t|^2)$$

2.7. Hiperparametre Optimizasyonu

Hiperparametreler, model veriler üzerinde eğitilmeden veya test edilmeden önce 'sabit' olan modele özgü özelliklerdir. Örneğin: rastgele bir orman söz konusu olduğunda, hiperparametreler ormandaki karar ağaçlarının sayısı, sinir ağı için öğrenme oranı, gizli katmanların sayısı, her katmandaki birimlerin sayısı vb. birkaç parametreyi içermektedir. Hiperparametre optimizasyonu, oluşturulan modellerin performanslarını arttırmak için doğru hiperparametre setini aramaktır. Hiperparametreleri optimize etmek, veri madenciliği modellerinin oluşturulmasında en zor kısımlardan biridir. Hiperparametre optimizasyonunun temel amacı, daha iyi bir performans elde edebilmek amacıyla model parametreleri için en iyi noktayı bulmaktır. Manuel, Izgara ve Rasgele arama yöntemleri bu amaçla sıklıkla kullanılmaktadır (89).

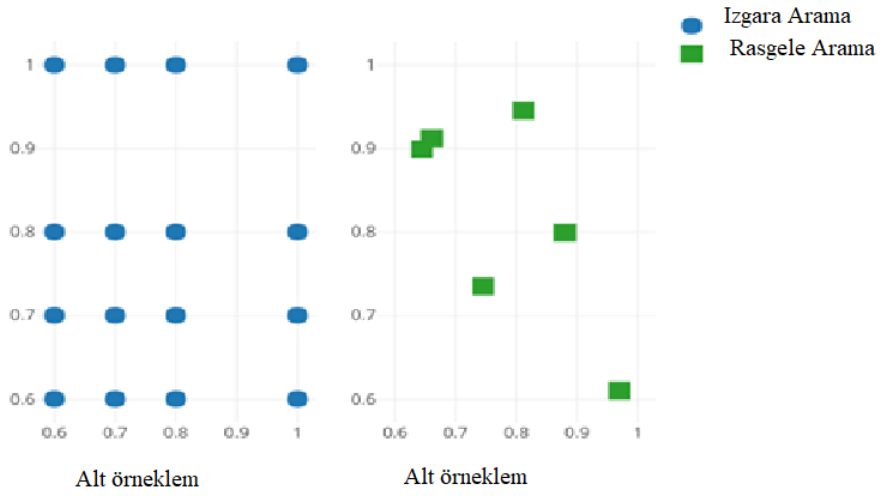
Manuel Arama (Manual Search) yönteminde, araştırıcı deneyimlerine dayanarak model için bazı hiperparametreleri seçer. Daha sonra modeli eğitir, modelin performansını değerlendirir ve süreci tekrar başlatır. Bu döngü, tatmin edici bir performans elde edilene kadar tekrarlanır (89).

Izgara Arama (Grid Search) yöntemi, hiperparametrelerin önceden ayarlanmış bir değerler listesinin her bir kombinasyonunu dener ve her kombinasyon için modeli değerlendirir. Burada izlenen desen, tüm değerlerin bir matris şeklinde yerleştirildiği ızgaraya benzemektedir. Her parametre seti dikkate alınır ve elde edilen performans ölçütleri not edilir. Tüm kombinasyonlar değerlendirildikten sonra, en yüksek performans ölçütünü veren parametre setine sahip model en iyi model olarak kabul edilir. Izgara aramanın en büyük dezavantajlarından biri, boyutsallık söz konusu olduğunda, hiperparametre sayısının katlanarak artmasıdır (89).

Rastgele Arama (Random Search), Izgara Arama'ya benzer bir yaklaşımdır, ancak tüm olası kombinasyonları test etmek yerine, rastgele parametre grupları seçilir. Yani rastgele arama yönteminde, kurulan model için en iyi çözümü bulmak amacıyla hiperparametrelerin rastgele kombinasyonları kullanılır. Rastgele arama yönteminin bazı özellikleri aşağıdaki gibidir (89).

- Izgara arama yöntemine göre daha hızlıdır. Çünkü tüm kombinasyonlar yerine rastgele seçilen alt parametre grupları üzerinde çalışır, bu da yöntemin hızını artırır.
- Her zaman başarılı sonucu vermeyi garanti etmez.
- Izgara aramaya göre daha geniş aralıklarda parametreler belirlenebileceği için bazı durumlarda daha iyi başarım verebilir.

Şekil 2.6’da Izgara Arama ve Rastgele Arama yöntemleri arasındaki fark gösterilmektedir.



Şekil 2.7: Rasgele ve Izgara Arama Arasındaki Fark

3. MATERYAL VE METOT

3.1. Çalışmada Kullanılan Veri Seti

Bu çalışmada GEO veri deposunda GDS3057 kodu ile yüklenen Akut miyeloid lösemi (AML: Acute myeloid leukemia) veri seti kullanılmıştır. AML hematopoetik malignitelerin en sık görülen ve ölümcül şekillerinden biridir. Stirewalt ve ark. (2007), AML'de aşırı eksprese edilen bazı genlerin rolü üzerine mikrodizilim ile yalnızca AML blastlarında ortaya çıkan, daha önce fark edilmemiş ifade değişikliklerinin tanımlanabileceği hipotezine dayanan bir araştırma yapmışlardır. Bu hipotezi test etmek için, 38 sağlıklı donörden alınan normal hematopoietik hücreler ile 26 AML hastasından gelen lösemik blastlar arasındaki gen ifade profilleri karşılaştırılmıştır. AML veri seti 64 kişiye ait 22283 gene ait ifade seviyelerini içermektedir (90).

3.2. Kullanılan Yöntemler

Bu çalışmadaki tüm analizler R 3.6.3 programında yapılmıştır. R programı, http://cran.rproject.org/ adresinden ücretsiz bir şekilde indirilebilmektedir. R programının tercih edilmesindeki sebepler ise; gen ifade verileri gibi yüksek boyutlu veri setlerinde hızlı sonuç veren bir yapıda olması ve kullanıcı dostu olmasıdır. Çalışmada normalizasyon işlemleri için affyPLM(91) paketindeki normalizasyon fonksiyonu, temel bileşenler analizi için mixOmics(92) paketindeki PCA fonksiyonu, bağımsız bileşenler analizi için fastICA (93) paketindeki fastICA fonksiyonu ve LASSO özellik seçimi için glmnet (94) paketindeki glmnet fonksiyonu kullanılmıştır. Çalışmada, PCA, ICA, LASSO özellik seçimi ve özellik çıkarımı yöntemleri AML gen ifade veri setinin boyutunu indirgemek amacıyla uygulanmıştır. Çalışmada kullanılan AML veri setine PCA yöntemi uygulanmadan önce normalizasyon işlemi yapılmıştır. Özellik seçim ve özellik çıkarım yöntemleri uygulanmadan önce gen filtreleme işlemi için R programında geneFilter paketindeki nsFilter fonksiyonu kullanılmıştır. Bu fonksiyonda filtreleme sonucunda örnekler arasında az değişiklik gösteren veya sürekli düşük sinyal sergileyen özelliklerin filtrelenerek atılması yoluna gidilmektedir (95). Modelleme aşamasında yeniden örnekleme yöntemi olarak tekrarlı 10 katlı çapraz geçerlik yöntemi kullanılmıştır.

Boyutu indirgenmiş veri setlerine caret (Classification And REgression Training) (96) paketindeki Tablo 3.1'de belirtilen çekirdek fonksiyonlu DVM'ler uygulanmıştır. Hiperparametre optimizasyonu için rasgele arama (random search) yöntemi kullanılmıştır. Model performansını değerlendirmek için oluşturulan sınıflama modellerinin doğru sınıflama oranları verilmiştir. Performans ölçütleri olarak doğru sınıflama oranına ek olarak gerçekte hasta olanlar arasında testin pozitif çıkma oranını ifade eden duyarlılık, gerçekte sağlam olanlar arasında testin negatif çıkma oranını ifade eden seçicilik, sınıfı 1 olarak tahmin edilmiş doğru pozitif (true positive) örnek sayısının sınıfı 1 olarak tahminlenmiş tüm örnek sayısına oranını ifade eden kesinlik (precision) ölçütleri hesaplanmıştır. Ancak kesinlik ve duyarlılık ölçütleri tek başına anlamlı bir karşılaştırma sonucu çıkarmamıza yeterli değildir. Bu sebeple her iki ölçütü beraber değerlendirmek daha doğrudur. Doğru sınıflama oranı, duyarlılık, seçicilik ve kesinlik ölçütlerine ek olarak duyarlılık ve kesinlik ölçütlerinin harmonik ortalaması olarak hesaplanan F ölçütü de verilmiştir. Tüm bu ölçütlere ek olarak boyut indirgeme analizlerinin modelleme süresine etkilerini görebilmek için analiz süreleri de saniye olarak verilmiştir.

3.2.1. LASSO Özellik Seçimi

LASSO yöntemi ilk olarak 1996'da Robert Tibshirani tarafından kullanılmıştır. Yöntemin temel iki görevi düzenlileştirme ve özellik seçimidir. LASSO yöntemi, model parametrelerinin mutlak değerlerinin toplamına bir kısıt koyar, toplamın sabit bir değerden (üst sınır) daha az olması gerekir. Bunu yapmak için yöntem, bir kısmını sıfıra düşüren regresyon değişkenlerinin katsayılarını cezalandırdığı bir daraltma (düzenlileştirme) süreci uygular. Özellik seçme işlemi sırasında, daraltma işleminden sonra hala sıfır olmayan bir katsayısı olan değişkenler modele seçilir. Bu işlemin amacı tahmin hatasını en aza indirmektir. Uygulamada, cezanın gücünü kontrol eden parametre olan λ büyük önem taşır. λ yeterince büyük olduğunda, katsayılar tam olarak sıfıra eşit olmaya zorlanır, bu şekilde boyutsallık azaltılabilir. λ parametresi ne kadar büyük olursa, o kadar fazla katsayı sıfıra indirilir. LASSO yöntemini kullanmanın birçok avantajı vardır, her şeyden önce çok iyi bir tahmin doğruluğu sağlayabilir, çünkü katsayıların küçültülmesi ve çıkarılması, sapmanın önemli bir artışı olmadan varyansı azaltabilir. Özellikle veri setinde az sayıda gözlem ve çok sayıda değişken olduğunda kullanışlıdır.

Ayrıca LASSO, yanıt değişkeni ile ilişkili olmayan alakasız değişkenleri ortadan kaldırarak modelin yorumlanabilirliğini artırmaya yardımcı olur, bu şekilde aşırı öğrenme sorunu da giderilebilir (97).

3.2.2. Temel Bileşenler Analizi (PCA)

Temel bileşenler analizi, yeni bir koordinat sistemi (yeni boyutlar kümesiyle, D′) oluşturmak için bir veri setinde (D boyutlarına sahip) dönüşümler gerçekleştirir. Verilerdeki varyans; D′1 D′2'den daha fazla değişkenlik gösterirken, D′2 ise D′3'ten daha fazla değişkenlik gösterir. Daha basit bir ifadeyle, veriler hakkında mümkün olan en önemli bilgiler bir boyuta sığar, daha sonra ikinci bir boyut için aynı işlemi yapar ve böylece verilerde orijinal olarak mevcut olan aynı sayıda yaratılmış boyut bulunana kadar devam eder. Bu, birinci boyutun verilerin orijinal yapısını ikinciden daha fazla tanımlayabildiği anlamına gelir. Bir boyutluluk azaltma yöntemi olarak PCA'da, verileri tanımlamak için daha az önemli olan boyutlar ihmal edilebilir. Örneğin, kullanıcı % 95 oranında değişimin korunması gerektiğini belirtirse, D ′ orijinal varyansın en az % 95'ini koruyan en küçük boyut setini içerecektir. Özellikle bir algoritmanın teorik temelinin performansı nasıl etkilediğine odaklanırken, boyutların belirlendiği metrik (yani varyans) göz önünde bulundurmalıdır. Veriler sürekli, çok değişkenli normalliğe sahip ve bağımsız olduğunda varyans yararlı bir metrik olacaktır. Ek olarak bu varsayımları sağlamayan veri setleri, PCA ile analiz edildiğinde kötü sonuçlar verecektir. Bu nedenle varsayımların sağlanıp sağlanmadığı PCA uygulanmadan önce mutlaka kontrol edilmelidir. Ayrıca PCA veri içindeki doğrusal bileşenleri bulan bir yöntemdir. Ancak doğrusal olmayan bileşenlerde verilerde bulunabilir. PCA'nın bu tür veriler üzerinde çalıştırılması böyle bir varsayımda bulunmayan algoritmalardan daha kötü sonuçlar verebilir (98).

3.2.3. Bağımsız Bileşenler Analizi (ICA)

Bağımsız bileşenler analizinde amaç normal dağılımlı olmayan verinin istatistiksel olarak bağımsız ya da olabildiğince bağımsız olmasını sağlayacak şekilde bir doğrusal dönüşümünü elde etmektir. ICA ilk olarak, kas büzülmesindeki hareketin basitleştirilmiş bir modelinin geliştirilmesi konusunda yapılan bir çalışmada Herault vd. tarafından ortaya atılmıştır. Bağımsız bileşenler adı ise ilk kez Comon tarafından yazılan bir makalede dile getirilmiştir. Bağımsız bileşenler analizi günümüzde genetik, görüntü işleme, beyin tomografisi, iletişim, finans, sismoloji vb. gibi farklı disiplinlerde oldukça geniş bir uygulama alanı bulmaktadır.

Bağımsız bileşenler analizinde çok değişkenli verilerin bir dizi bağımsız bileşenin (faktörün) doğrusal birleşiminden oluştuğu varsayılır. Bileşen sayısı genel olarak değişken sayısına eşit alınmaktadır.

Her biri n adet noktada örneklenmiş p adet değişkenden oluşan veri setini Z matrisi ile gösterecek olursak, ICA modelinde Z matrisi

$$Z = AY \qquad (2.1)$$

matris çarpımı ile ifade edilir. Bu eşitlikte A: karışım matrisini, Y: bağımsız bileşenleri içeren kaynak matrisini göstermektedir. Burada hem karışım matrisi, hem de kaynak matrisi bilinmemektedir.

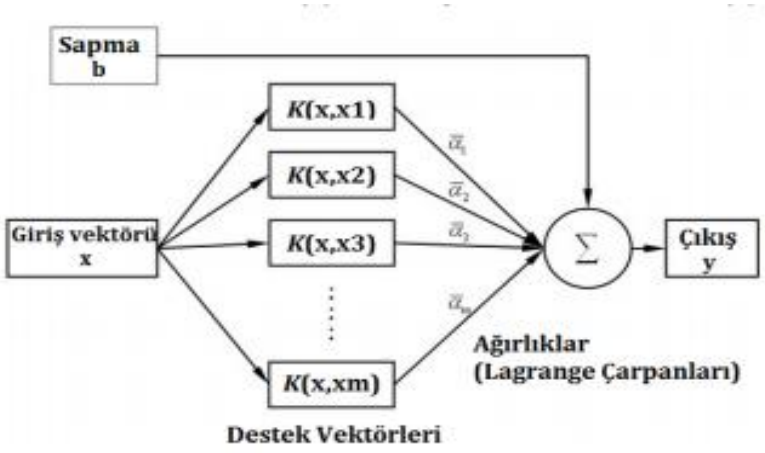
Bağımsız bileşenler analizi altında her iki matris, yalnızca orijinal veri matrisi kullanılarak kestirilir. Önce karışım matrisi kestirilir. Daha sonra A'nın tersi ile Z veri matrisi çarpılarak bağımsız bileşenleri içeren matris elde edilir. ICA modelinin tanımlı olabilmesi için bağımsız bileşenlerin normal dağılmaması gerekmektedir. Ayrıca karışım sayısının bağımsız bileşen sayısına eşit olduğu varsayılır. Bu varsayım ayrıştırma için gerekli olan işlemleri kolaylaştırmaktadır.

ICA, temel bileşenler analizi (PCA), faktör analizi (FA) ve minimum/maksimum oto korelasyon faktörleri (MOF) analizi gibi çok değişkenli yöntemlere benzemektedir. Çok değişkenli veri setleri bu yöntemlerin hepsinde faktörlerin doğrusal bir birleşimi olarak ifade edilir. PCA, FA ve MOF yöntemleri ile ilişkisiz ve normal dağılım gösteren faktörler elde edilirken, veri seti ICA ile analiz edildiğinde birbirinden bağımsız ve aynı zamanda normal dağılım göstermeyen faktörler elde edilmektedir (62).

Bağımsızlık, ilişkisizlikten daha kuvvetli bir kavramdır. İlişkisiz faktörlerin elde edilmesinde kovaryans gibi ikinci derece istatistiklerden yararlanılmaktadır. ICA ile faktörlerin kestiriminde ise daha yüksek dereceden istatistiklere gereksinim duyulur. ICA ile diğer çok değişkenli yöntemler arasındaki bir fark elde edilen faktörlerin yorumudur. ICA'da elde edilen bileşenler için büyüklük sıralaması yapılmamaktadır. Diğer bir ifade ile kötü ya da iyi bileşen yoktur. Ayrıca elde edilen bileşenler kaynağın işaretine göre değişiklik göstermez. Görüntü işleme analizlerinden örnek verecek olursak; kırmızı bir zemin üzerindeki beyaz bir harf, beyaz bir zemin üzerindeki kırmızı bir harf ile aynıdır (63).

3.2.4. Oluşturulan Destek Vektör Makinesi Modelleri

DVM son yıllarda sıklıkla kullanılan ve oldukça iyi sonuçlar veren sınıflandırma yöntemlerinden biridir. Bu yöntem, hastalıkların tanısı, farklı mühendislik uygulamaları vb. farklı alanlarda kullanılmaktadır (99). Cortes ve Vapnik tarafından önerilen DVM yöntemi yapısal olarak çalışan riski minimize etme ilkesini kullanmaktadır. Bu teknikte, iki sınıfın/grubun birbirlerine en yakın gözlemlere ilişkin uzaklıklarını en üst düzeye çıkarıldığı bir düzlem incelenir. Doğrusal yöntemlerle sınıflandırılmayan veriler, DVM yöntemiyle giriş vektörü daha büyük boyutlu bir uzaya lineer olmayan bir işlev aracılığıyla eşleştirilir. DVM'nin eğitilmesinde 2. dereceden bir optimizasyon denklemi kullanılabilir (100).Aynı zamanda, makine öğrenme yöntemlerinde, destek vektör makineleri (destek vektör ağları olarak da adlandırılır) sınıflandırma ve regresyon analizi için kullanılan öğrenme algoritmalarıyla denetlenen modellerdendir. Her biri, her iki kategoriden birine ya da diğerine ait olarak işaretlenmiş bir dizi eğitim örneği verildiğinde, bir DVM eğitim algoritması, ikili bir sınıflayıcıdan yararlanarak incelenen kategorilerden birisine ya da diğerine yeni örnekler atayan bir model oluşturur. Bir DVM modelinde, örnekler uzaydaki noktalar olarak gösterilir ve ayrı kategorilerin örnekleri mümkün olduğunca açık bir boşluk ile bölünür. Yeni örnekler daha sonra aynı alana atanır ve yeni örneklerin boşluğun hangi tarafına dayanan bir kategoriye ait olduğu öngörülür. Doğrusal sınıflandırmaya ek olarak, DVM'ler doğrusal olmayan sınıflandırma problemlerini de başarılı bir şekilde gerçekleştirebilir ve girişlerini yüksek boyutlu özellik alanlarına örtülü olarak eşler. Verilerin atanabilecekleri olası gruplar belli olmadığında, denetimli/danışmanlı öğrenme mümkün değildir. Bu durumda verilerin gruplara doğal olarak kümelenmesini ve daha sonra bu gruplara yeni verilerin atanmasını gerçekleştiren denetimsiz/danışmansız bir öğrenme yaklaşımı gereklidir (101). DVM'nin yapısı Şekil 3.1'de verilmiştir (102).



Şekil 3.1: DVM'nin yapısı

Şekil 3.1'de $K(x, x1)$, $K(x, x2)$, $K(x, x3)$, … , $K(x, x\text{m})$ ile gösterilen fonksiyonlar çekirdek fonksiyonlarını; $\bar{\alpha}1$, $\bar{\alpha}2$, $\bar{\alpha}3$, … , $\bar{\alpha}\text{m}$ ile ifade edilen Lagrange çarpanları ise ağırlık katsayılarıdır. Bu çalışmada çeşitli DVM modelleri, Tablo 3.1'de ayrıntılı şekilde tanımlanan değişik çekirdek fonksiyonları ile oluşturulmuş ve PCA, ICA, LASSO yöntemleri ile boyutu indirgenmiş veri setlerine uygulanmıştır.

3.2.5. Hiperparametre Optimizasyonu

Çalışmada hiperparametre optimizasyonu için literatürde sıklıkla kullanılan yöntemlerden biri olan rasgele arama (random search) yöntemi kullanıldı. Yöntem ilk olarak Bengio tarafından 2012 yılında yayınlanan makalede önerilmiştir. Rasgele arama yönteminde probleme dair ön bilgiler kullanılarak hiperparametre aralıkları belirlenir. Daha sonra bu aralıkta ki değerlerin her birini denemek yerine rastgele değerler seçilerek hiperparametre grupları oluşturulur. En iyi performansı sağlayan parametreler bulunana kadar rastgele farklı parametre gruplarıyla model eğitilir (103).

Tablo 3.1: Çekirdek Tipleri ve Fonksiyonları

Çekirdek Tipi	Fonksiyon
Doğrusal	$k(x,y)= x^T y + c$
Polinomiyal	$k(x,y)=(ax^T y + c)^d$
Radyal Tabanlı	$k(x,y)= \exp(-\gamma \|\|x-y\|\|^2$

4. BULGULAR

22283 gen ifade verisi içeren AML veri setine Nsfilter ile filtreleme işlemi yapıldıktan sonra veri setinde 6201 gen kalmıştır. Çalışmanın amacına bağlı olarak filtrelenmiş veri setine PCA, ICA, LASSO yöntemleri uygulandıktan sonra Doğrusal, Polinomiyal ve Radyal tabanlı çekirdek fonksiyonlu DVM yöntemleri uygulanmıştır. Kullanılan her bir yönteme ilişkin eğitim veri seti için sonuçlar Tablo 4.1'de ve test veri seti sonuçlar için Tablo 4.2'de verilmiştir. PCA ile oluşturulan 10 temel bileşen veri setindeki toplam varyansın %89'unu açıklamaktadır.

Tablo 4.1: Eğitim Veri Seti için Modellerin Sonuçları

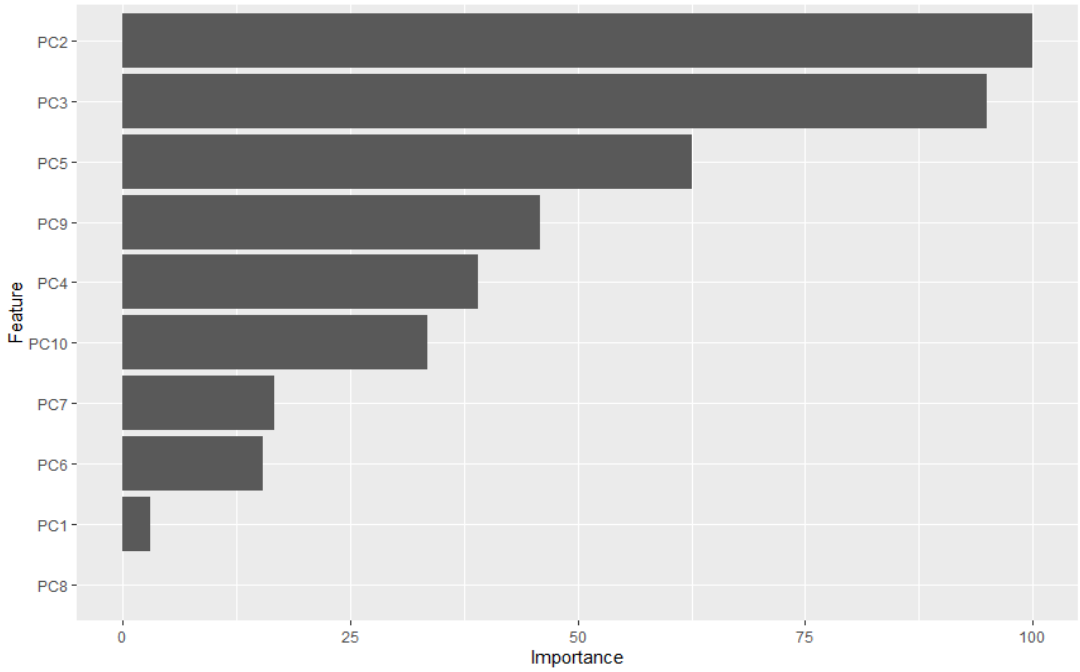
Çekirdek Fonksiyonu/Boyut İndirgeme Yöntemi	Optimizasyon Parametreleri	Parametre Değerleri	Doğruluk (%)
Doğrusal	C	4	98.52
Doğrusal/PCA	C	323.29	99.69
Doğrusal/ICA	C	0.66	99.83
Doğrusal/LASSO	C	0.052	1
Polinomiyal	C Derece Ölçek	188.94 1 0.0001	99.03
Polinomiyal/PCA	C Derece Ölçek	1.804 2 0.0048	99.85
Polinomiyal/ICA	C Derece Ölçek	0.38 3 1.89	1
Polinomiyal/LASSO	C Derece Ölçek	0.067 1 0.067	1
Radyal Tabanlı	C sigma	2.01 7.85	99.13
Radyal Tabanlı/PCA	C sigma	0.20 0.022	99.23
Radyal Tabanlı/ICA	C sigma	0.72 0.02	1
Radyal Tabanlı/LASSO	C sigma	0.03 0.031	1

Eğitim veri seti için kurulan modellerden LASSO özellik seçimi sonrası seçilen 21 gen ile oluşturulan Doğrusal, Polinomiyal ve Radyal tabanlı DVM modellerinin eğitim seti için doğru sınıflama oranları 1 olarak bulunmuştur. ICA sonrası seçilen 10 bileşen ile oluşturulan Polinomiyal ve Radyal tabanlı DVM modellerinin de eğitim seti için doğruluk oranları 1 olarak bulunmuştur.

Tablo 4.2: Test Veri Seti için Modellerin Sonuçları

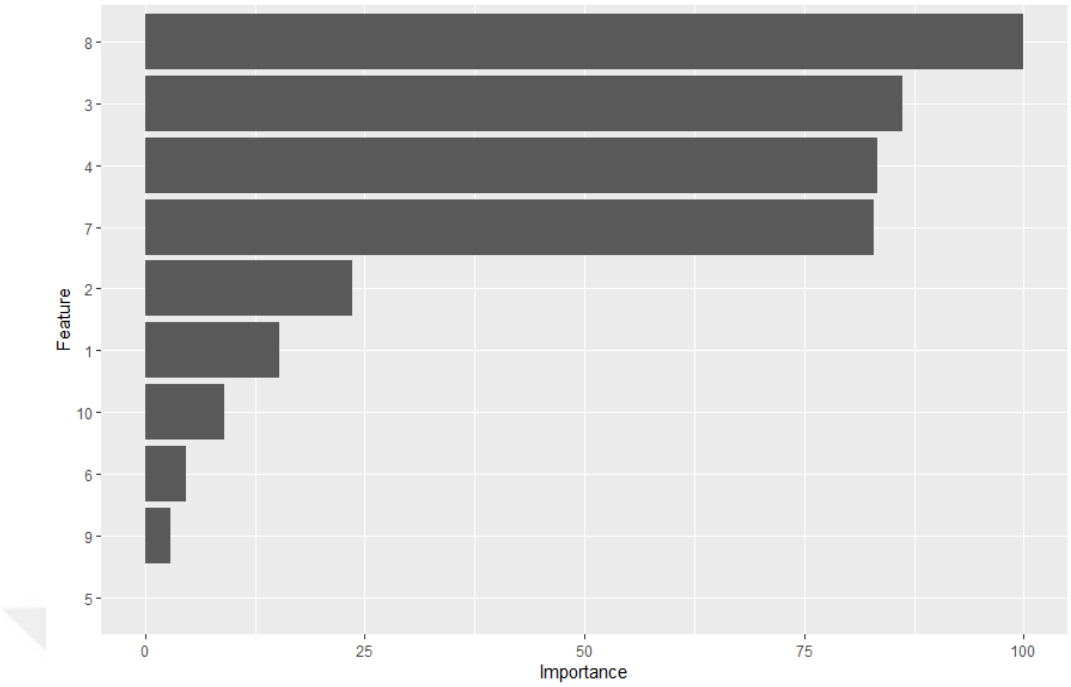
Özellik Çıkarımı/Seçimi Yöntemi	DVM Çekirdek Fonksiyonu	Doğruluk (%)	Duyarlılık (%)	Seçicilik (%)	Kesinlik (%)	F-ölçütü (%)	Analiz Süresi (sn.)
	Doğrusal	94.10	93.22	95.25	94.79	93.31	2155.8
	Polinomiyal	95.05	94.24	96.04	95.08	94.08	2174.77
	Radyal Tabanlı	95.01	94.35	95.56	94.58	93.92	2252.36
PCA	Doğrusal	95.20	93.72	96.22	95.28	93.97	35.78
	Polinomiyal	95.64	94.53	96.46	95.48	94.99	36.33
	Radyal Tabanlı	94.94	93.54	95.90	94.91	93.65	36.07
ICA	Doğrusal	95.15	93.83	96.05	95.09	93.91	34.24
	Polinomiyal	95.41	94.02	96.03	95.1	94.05	38.97
	Radyal Tabanlı	95.07	93.84	95.91	94.94	93.86	35.74
LASSO	Doğrusal	94.98	93.71	95.85	94.90	93.75	32.67
	Polinomiyal	95.38	93.36	96.27	95.39	93.84	33.54
	Radyal Tabanlı	95.23	94.26	95.89	94.90	94.05	35.53

Doğruluk değeri dikkate alındığında, AML gen ifade veri setini sınıflamada en iyi performansı Polinomial DVM ile kurulan modeller vermiştir. Doğruluk oranları sırasıyla; PCA + Polinomial (%95.64), ICA + Polinomial (%95.41), LASSO + Polinomial (%95.38) olarak bulunmuştur. Doğruluk, Duyarlılık, Seçicilik, Kesinlik ve F ölçütleri için sonuçlar incelendiğinde kurulan modellerden en ayırt edici olanının PCA sonrası Polinomiyal DVM ile kurulan model olduğu görülmektedir.



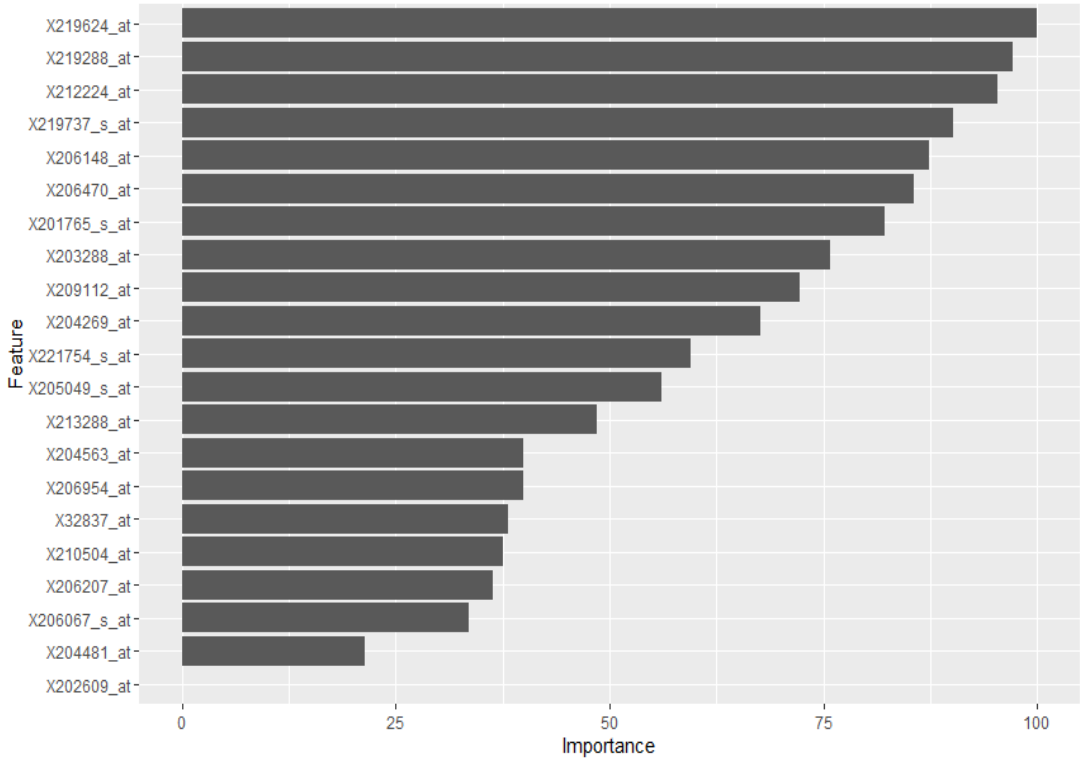
Şekil 4.1: PCA sonrası modele eklenen değişkenlerin önem sırası

Şekil 4.1 incelendiğinde PCA sonrası modele dâhil edilen değişkenlerden en önemlisi 2. temel bileşendir. Diğer bileşenlerin önemi giderek azalırken, 6. ve 7. temel bileşenlerin önemi yaklaşık olarak eşittir. En az öneme sahip bileşenin ise 1. temel bileşen olduğu gözlenmiştir. Ek olarak 8. temel bileşenin modele bir katkısı olmadığı grafikten görülmektedir.



Şekil 4.2 ICA sonrası modele eklenen değişkenlerin önem sırası

ICA sonrası modele dâhil edilen değişkenlerin Şekil 4.2 ile verilen önem sıralaması incelendiğinde en önemli bileşenin 8. bağımsız bileşen, diğer önemli bileşenlerin ise sırasıyla 3., 4. ve 7. bağımsız bileşenler olduğu bulunmuştur. En az öneme sahip bileşenin ise 9. bileşen olduğu gözlenmiştir. 5. bağımsız bileşenin ise modele herhangi bir katkısının olmadığı görülmektedir.



Şekil 4.3: LASSO sonrası modele eklenen genlerin önem sırası

Şekil 4.3 incelendiğinde LASSO özellik seçimi yöntemi ile elde edilen 21 genden modele en çok katkı yapan gen prob numarası 219624_at olan gendir. LASSO özellik seçimi yöntemi ile seçilen 21 genden bazılarına ait bilgiler BIOGPS veri tabanından elde edilmiş olup Tablo 4.3'te verilmiştir. Tablo 4.3 incelendiğinde LASSO yöntemi ile seçilen genler AML hastalığının tanı ve tedavisinde etkili genler olup, AML hastalığında biyobelirteç olabilecek genlerdir.

Tablo 4.3: LASSO özellik seçimi yöntemi ile seçilen genlerin bazıları için açıklamalar

Prob Set İsmi	Tanımlayıcı	Gen Sembolü	Fonksiyonu
219624_at	BAG cochaperone 4	BAG4	Bu gen tarafından kodlanan protein, BAG gen ailesi ile ilişkili protein ailesinin bir üyesi olup, anti-apoptotik bir proteindir(anti-apoptotik protein:hücreleri çoğaltan protein). BAG gen ailesi normal kan hücrelerinde düşük seviyelerde eksprese edilirken, hem primer lösemide hem de lösemili hastaların yerleşik hücre hatlarında yüksek oranda eksprese edilir. BAG gen ifade seviyeleri kemoterapiye duyarlı hastalar ile karşılaştırıldığında, ilaca dirençli hastalarda daha yüksektir (104).
219288_at	chromosome 3 open reading frame 14	C3orf14	Washington Üniversitesi Akut Miyeloid Lösemi Programı Genomikleri ifade verilerinin keşfi ve doğrulanması için yapılan çalışmada bulunan C3orf14 geni AML hastalığı için biyobelirteç olabilecek bir gen olarak bulunmuştur (105).
212224_at	aldehyde dehydrogenase 1 family member A1	ALDH1A1	Toksik ALDH1A1 substratlarına sahip lösemik hücreler, akut miyeloid lösemiler için yeni bir hedefe yönelik terapötik strateji olabilir (106).
219737_s_at	protocadherin 9	PCDH9	PCDH9 upregülasyonu Akut lenfoblastik lösemi'de (ALL) zayıf bir prognostik faktördür. Yani lösemide çok ifade edilir (107).
206148_at	interleukin 3 receptor subunit alpha	IL3RA	IL3RA geni AML numunelerinde normallere kıyasla artan bir ifade göstermektedir (90).
206470_at	plexin C1	PLXNC1	AML'de PLXNC1 geni ifadesi azalma göstermektedir (90).
201765_s_at	hexosaminidase subunit alpha	HEXA	Literatürde Tay-Sachs hastalığına heksozaminidazın alfa alt birimindeki mutasyondan kaynaklandığı ile ilgili bilgiler yer almaktadır (108).
209112_at	cyclin dependent kinase inhibitor 1B	CDKN1B	CDKN1B, akut miyeloid lösemide prognoz için potansiyel biyobelirteç olmaya aday bir gendir (109).
204269_at	Pim-2 proto-oncogene, serine/threonine kinase	PIM2	Akut Miyeloid Lösemi Hastaları için tedavi hedefi olabilir (110).
206207_at	Charcot-Leyden crystal galectin	CLC	Miyeloid lösemiler ile ilişkili olabilir (111).

5. TARTIŞMA

Mikrodizilim deneylerinden elde edilen gen ifade veri setlerinin boyutları oldukça yüksektir. Yüksek boyutlu olması nedeniyle gen ifade veri setleri ile modelleme analizleri uzun sürmekte ve bundan dolayı bu veri setleri analizlerde hesaplama verimsizliğine yol açabilmektedir. Yüksek boyutluluk sorunu model performansının da düşmesine sebep olabilmektedir. Ek olarak gen ifade veri setlerinde gen sayılarının çok fazla olması bir sınıflandırma algoritmasının eğitim örneklerine uymasına ve yeni örnekleri zayıf genellemesine neden olabilmektedir. Bu sorunları giderebilmek amacıyla bu çalışmada boyut indirgemenin iki farklı şekli olan; özellik seçimi ve özellik çıkarımı yöntemlerinden PCA, ICA, LASSO yöntemlerinin sonuçları karşılaştırmalı olarak incelendi.

Tablo 4.1'deki sonuçlara göre boyut indirgeme yöntemleri uygulanmayan ham veri seti ile oluşturulan modeller, AML gen ifade veri seti için hesaplama verimsizliğine yol açmıştır. Tablo 4.1'deki analiz süreleri incelendiğinde PCA ile kurulan model en kısa (35.78 sn.), veri setine hiçbir işlem yapılmadan Radyal tabanlı DVM ile kurulan model ise en uzun (2252.36 sn.) sürede tamamlanmıştır.

Ayrıca hem eğitim hem de test veri seti için boyut indirgeme yapılmadan kurulan sınıflandırma modelleri (Doğrusal, Polinomiyal ve Radyal tabanlı DVM), boyut indirgeme analizleri yapıldıktan sonra kurulan modellere göre daha düşük performans göstermiştir. Diğer bir ifadeyle, AML gen ifade veri setini sınıflandırmadan önce uygulanan özellik seçimi ve özellik çıkarımı yöntemleri sınıflama performansını arttırmıştır. PCA ve ICA özellik çıkarımı yöntemleri karşılaştırılacak olursa test veri seti için doğrusal DVM ve Polinomiyal çekirdek fonksiyonlu DVM ile kurulan modellerde PCA, Radyal tabanlı fonksiyon ile kurulan DVM modelinde ise ICA yöntemi daha iyi performans göstermiştir.

LASSO özellik seçimi ve PCA/ICA özellik çıkarımı yöntemleri sonucunda oluşturulan modeller incelendiğinde; LASSO özellik seçimi yöntemi sonrasında Radyal tabanlı DVM ile kurulan model PCA/ICA özellik çıkarımı yöntemleri sonrasında Radyal tabanlı DVM ile kurulan modellere göre daha iyi performans göstermiştir.

PCA, ICA ve LASSO yöntemlerinin sonucunda boyutu indirgenmiş veri setlerinin tümü ile kurulan DVM modellerinden en iyisi Polinomiyal çekirdek fonksiyon ile kurulan modellerdir. Doğruluk, Duyarlılık, Seçicilik, Kesinlik ve F ölçütleri için elde edilen sonuçlara göre PCA sonrası Polinomiyal DVM ile kurulan model en iyi performansı vermiştir.

Ek olarak LASSO özellik seçimi sonucunda AML hastalığı ile en çok ilgili olan 21 adet biyobelirteç gen seçilmiştir. Bu seçilen genlere ait prob numaraları BIOGPS veri tabanından ve literatürden incelenmiştir. Seçilen bu genlerin AML hastalığı için tanı ve tedavide biyobelirteç gen olarak kullanılabileceği ile ilgili bilgiler literatürde yer almaktadır. LASSO yöntemi ile seçilen bu genler, AML hastalığına yol açan başka genlerin tespit edilmesine yardımcı olacak ağlar için kullanılabilir. LASSO özellik seçimi yöntemi ile seçilen genlerden PIM2 ve CDKN1B geni Akut Miyeloid Lösemi hastaları için tedavi hedefi olabilecek bir gendir. Bu geni hedef alarak geliştirilecek olan ilaçlar AML hastalığının tedavisinde etkili olabilirler.

6. SONUÇ VE ÖNERİLER

Bu çalışma sonucunda mikrodizilim deneyleri ile elde edilen AML gen ifade veri setinde boyut indirgeme yöntemlerinin sınıflandırma modelleme performansına etkileri incelenmiştir. Gen ifade veri setleri oldukça yüksek boyutlu veri setleridir. Uygulamada analiz aşamasında bu yüksek boyutluluk sorunu analizlerin uzun zaman almasına ve model performanslarının düşmesine sebep olmaktadır. Bu nedenle gen ifade veri setleri ile tahmin modelleri oluşturulmadan önce boyut indirgeme analizleri yapılmalıdır. Boyut indirgeme amacıyla kullanılan özellik seçim ve özellik çıkarımı yöntemlerinin her ikisi de model performansını arttırmaya yardımcı yöntemlerdir. Bu yöntemlerden özellik seçimi yöntemleri ile biyobelirteç gen bulunurken, özellik çıkarımı yöntemleri ile model performansını arttıran yeni bileşenler elde edilir. Amaç biyobelirteç gen bulmak ise özellik seçimi yöntemleri, verinin boyutunu indirgeyerek veriyi en iyi şekilde temsil eden daha küçük boyutlu veri setine indirgemek ise özellik çıkarımı yöntemleri tercih edilmelidir. Çalışmada kullanılan LASSO yöntemi ile seçilen genlerin birçoğu AML hastalığı için biyobelirteç olabilecek genlerdir. Ancak seçilen genlerden HEXA geninin Tay-Sachs hastalığına neden olan bir gen olduğu literatürden ve biyolojik veri tabanlarından bulunmuş olup, gelecek çalışmalarda HEXA geninin AML hastalığı için biyobelirteç olup olmayacağıyla ilgili araştırmalar yapılabilir. AML gen ifade veri seti için kurulan modellerin tümünde Polinomiyal çekirdek fonksiyonlu DVM modelleri en iyi tahmin performansını göstermişlerdir. Ancak tek bir veri setinde elde edilen sonuçlara bakarak Polinomiyal çekirdek fonksiyonlu DVM yönteminin en iyi yöntem olduğunu söylemek doğru değildir. Gelecek çalışmalarda farklı gen ifade veri setleri üzerinde ve/veya simüle gen ifade veri setinde yöntemler uygulanarak daha kesin sonuçlara ulaşılabilir.

KAYNAKLAR

1. Dziuda DM. *Data mining for genomics and proteomics: analysis of gene and protein expression data*, John Wiley & Sons 2010.

2. Apitz JC. *A statistical method for selection, classification, and network construction in genetic systems*, California State University, Long Beach 2016.

3. Başaran E, Aras S, Cansaran-Duman D. General outlook and applications of genomics, proteomics and metabolomics. *Turk Hij Den Biyol Derg* 2010, 67(2): 85-96.

4. Cosgun E, Karaağaoğlu E. Veri madenciliği yöntemleriyle mikrodizilim gen ifade analizi. *Hacettepe Tıp Dergisi* 2011, 42180-9.

5. Singh RK, Sivabalakrishnan M. Feature selection of gene expression data for cancer classification: a review. *Procedia Comput Sci* 2015, 5052-7.

6. Lotfi E, Keshavarz A. Gene expression microarray classification using PCA–BEL. *Comput Biol Med* 2014, 54180-7.

7. Perçın İ, Yağin FH, Güldoğan E, Yoloğlu S, editors. ARM: An Interactive Web Software for Association Rules Mining and an Application in Medicine. 2019 International Artificial Intelligence and Data Processing Symposium (IDAP); 2019: IEEE.

8. Pyle D. *Data preparation for data mining*, morgan kaufmann 1999.

9. Oliveri P, Malegori C, Simonetti R, Casale M. The impact of signal pre-processing on the final interpretation of analytical outcomes–A tutorial. *Anal Chim Acta* 2019, 10589-17.

10. Koh HC, Tan G. Data mining applications in healthcare. *J Healthc Inf Manag* 2011, 19(2): 65.

11. Perçin İ, Yağın FH, Arslan AK, Çolak C, editors. An Interactive Web Tool for Classification Problems Based on Machine Learning Algorithms Using Java Programming Language: Data Classification Software. 2019 3rd International Symposium on Multidisciplinary Studies and Innovative Technologies (ISMSIT); 2019: IEEE.

12. Fodor IK. A survey of dimension reduction techniques. *Center for Applied Scientific Computing* 2002.

13. Breiman L. Bagging predictors. *J Mach Learn Res* 1996, 24(2): 123-40.

14. Freund Y. Boosting a weak learning algorithm by majority. *Inf Comput* 1995, 121(2): 256-85.

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

16. Boulesteix A-L. Dimension reduction and classification with high-dimensional microarray data. 2005.

17. Roweis ST, Saul LK. Nonlinear dimensionality reduction by locally linear embedding. *science* 2000, 290(5500): 2323-6.

18. Blum AL, Rivest RL. Training a 3-node neural network is NP-complete. *Neural Networks* 1992, 5(1): 117-27.

19. Saeys Y, Inza I, Larrañaga P. A review of feature selection techniques in bioinformatics. *bioinformatics* 2007, 23(19): 2507-17.

20. BLUM Avrim L LP. Selection of Relevant Features and Examples in Machine Learning. *Artificial intelligence*.

21. Xing EP, Jordan MI, Karp RM, editors. Feature selection for high-dimensional genomic microarray data. Icml; 2001: Citeseer.

22. Novaković J. Toward optimal feature selection using ranking methods and classification algorithms. *Operations research* 2016, 21(1).
23. Bø TH, Jonassen I. New feature subset selection procedures for classification of expression profiles. *Genome Biol* 2002, 3(4).
24. Yeung KY, Bumgarner RE, Raftery AE. Bayesian model averaging: development of an improved multi-class, gene selection and classification tool for microarray data. *Bioinformatics* 2005, 21(10): 2394-402.
25. Ding C, Peng H. Minimum redundancy feature selection from microarray gene expression data. *J Bioinform Comput Biol* 2005, 3(02): 185-205.
26. Wang Y, Tetko IV, Hall MA, Frank E, Facius A, Mayer KF, Mewes HW. Gene selection from microarray data for cancer classification—a machine learning approach. *Comput Biol Chem* 2005, 29(1): 37-46.
27. Hall MA, Smith LA. Practical feature subset selection for machine learning. Computer Science '98 Proceedings of the 21st Australasian Computer Science Conference ACSC'98; Berlin: Springer1998.
28. Mercier G, Berthault N, Mary J, Peyre J, Antoniadis A, Comet JP, Cornuejols A, Froidevaux C, Dutreix M. Biological detection of low radiation doses by combining results of two microarray analysis methods. *Nucleic Acids Res* 2004, 32(1): e12-e.
29. Wang Y, Makedon F, editors. Application of Relief-F feature filtering algorithm to selecting informative genes for cancer classification using microarray data. Proceedings. 2004 IEEE Computational Systems Bioinformatics Conference, 2004. CSB 2004.; 2004: IEEE.
30. Weber G, Vinterbo S, Ohno-Machado L. Multivariate selection of genetic markers in diagnostic classification. *Artificial Intelligence in Medicine* 2004, 31(2): 155-67.
31. Whitney AW. A direct method of nonparametric measurement selection. *IEEE Trans Comput* 1971, 100(9): 1100-3.
32. Hüseyin B. Özellik Seçim Yöntemleri ve Yeni Bir Yaklaşım. *Süleyman Demirel Üniversitesi Fen Bilimleri Enstitüsü Dergisi* 2018, 2221-31.
33. Chapelle O, Vapnik V, Bousquet O, Mukherjee S. Choosing multiple parameters for support vector machines. *J Mach Learn Res* 2002, 46(1-3): 131-59.
34. Ambroise C, McLachlan GJ. Selection bias in gene extraction on the basis of microarray gene-expression data. *Proc Natl Acad Sci USA* 2002, 99(10): 6562-6.
35. Liu Q, Sung AH, Chen Z, Liu J, Chen L, Qiao M, Wang Z, Huang X, Deng Y. Gene selection and classification for cancer microarray data based on machine learning and similarity measures. *BMC Genomics* 2011, 12(S5): S1.
36. Guyon I, Weston J, Barnhill S, Vapnik V. Gene selection for cancer classification using support vector machines. *J Mach Learn Res* 2002, 46(1-3): 389-422.
37. Gutlein M, Frank E, Hall M, Karwath A, editors. Large-scale attribute selection using wrappers. 2009 IEEE symposium on computational intelligence and data mining; 2009: IEEE.
38. Jirapech-Umpai T, Aitken S. Feature selection and classification for microarray data analysis: Evolutionary methods for identifying predictive genes. *BMC Bioinformatics* 2005, 6(1): 148.
39. Ruiz R, Riquelme JC, Aguilar-Ruiz JS. Incremental wrapper-based gene selection from microarray data for cancer classification. *Pattern Recognit Lett* 2006, 39(12): 2383-92.

40. Huerta EB, Duval B, Hao J-K, editors. Gene selection for microarray data by a LDA-based genetic algorithm. IAPR international conference on pattern recognition in bioinformatics; 2008: Springer.
41. Perez M, Marwala T, editors. Microarray data feature selection using hybrid genetic algorithm simulated annealing. 2012 IEEE 27th Convention of Electrical and Electronics Engineers in Israel; 2012: IEEE.
42. Revathy N, Balasubramanian R. GA-SVM wrapper approach for gene ranking and classification using expressions of very few genes. *J Theor Appl Inf Technol* 2012, 40(2): 113-9.
43. Dunn JC. A fuzzy relative of the ISODATA process and its use in detecting compact well-separated clusters. *Cybern Syst* 1973.
44. Díaz-Uriarte R, De Andres SA. Gene selection and classification of microarray data using random forest. *BMC Bioinform* 2006, 7(1): 3.
45. Sheng L, Pique-Regi R, Asgharzadeh S, Ortega A, editors. Microarray classification using block diagonal linear discriminant analysis with embedded feature selection. 2009 IEEE International Conference on Acoustics, Speech and Signal Processing; 2009: IEEE.
46. Maldonado S, Weber R, Basak J. Simultaneous feature selection and classification using kernel-penalized support vector machines. *Inf Sci* 2011, 181(1): 115-28.
47. Tang EK, Suganthan PN, Yao X, editors. Feature selection for microarray data using least squares svm and particle swarm optimization. 2005 IEEE Symposium on Computational Intelligence in Bioinformatics and Computational Biology; 2005: IEEE.
48. Zhang X, Lu X, Shi Q, Xu X-q, Hon-chiu EL, Harris LN, Iglehart JD, Miron A, Liu JS, Wong WH. Recursive SVM feature selection and sample classification for mass-spectrometry and microarray data. *BMC bioinformatics* 2006, 7(1): 197.
49. Tang EK, Suganthan PN, Yao X. Gene selection algorithms for microarray data based on least squares support vector machine. *BMC bioinformatics* 2006, 7(1): 95.
50. Li G, Li J, Ju Z, Sun Y, Kong J, Applications. A novel feature extraction method for machine learning based on surface electromyography from healthy brain. *Neural Comput Appl* 2019, 31(12): 9013-22.
51. Jiménez AA, Márquez FPG, Moraleda VB, Muñoz CQG. Linear and nonlinear features and machine learning for wind turbine blade ice detection and diagnosis. *Renewable Energy* 2019, 1321034-48.
52. Wang A, Gehan EA. Gene selection for microarray data analysis using principal component analysis. *Statistics in medicine* 2005, 24(13): 2069-87.
53. Evangelista PF, Bonissone PP, Embrechts MJ, Szymanski BK, editors. unsupervised fuzzy ensembles and their use in intrusion detection. ESANN; 2005.
54. Bair E, Tibshirani R. Semi-supervised methods to predict patient survival from gene expression data. *PLoS Comput Biol* 2004, 2(4): e108.
55. Bair E, Hastie T, Paul D, Tibshirani R. Prediction by supervised principal components. *J Am Stat Assoc* 2006, 101(473): 119-37.
56. Borg P. Groenen. Modern multidimensional scaling. New York: Springer; 2005.
57. Gower JC. Some distance properties of latent root and vector methods used in multivariate analysis. *Biometrika* 1966, 53(3-4): 325-38.
58. Tenenbaum JB, De Silva V, Langford JC. A global geometric framework for nonlinear dimensionality reduction. *science* 2000, 290(5500): 2319-23.
59. Balasubramanian M, Schwartz EL. The isomap algorithm and topological stability. *Science* 2002, 295(5552): 7-.

60. Orsenigo C, Vercellis C. An effective double-bounded tree-connected Isomap algorithm for microarray data classification. *Pattern Recognit Lett* 2012, 33(1): 9-16.
61. Dawson K, Rodriguez RL, Malyj W. Sample phenotype clusters in high-density oligonucleotide microarray data sets are revealed using Isomap, a nonlinear algorithm. *BMC bioinformatics* 2005, 6(1): 195.
62. Chao S, Lihui C, editors. Feature dimension reduction for microarray data analysis using locally linear embedding. Proceedings Of The 3rd Asia-Pacific Bioinformatics Conference; 2005: World Scientific.
63. Ehler M, Rajapakse VN, Zeeberg BR, Brooks BP, Brown J, Czaja W, Bonner RF, editors. Nonlinear gene cluster analysis with labeling for microarray gene expression data in organ development. BMC proceedings; 2011: BioMed Central.
64. Kotani M, Sugiyama A, Ozawa S, editors. Analysis of DNA microarray data using self-organizing map and kernel based clustering. Proceedings of the 9th International Conference on Neural Information Processing, 2002. ICONIP'02.; 2002: IEEE.
65. Liu Z, Chen D, Bensmail H. Gene expression data classification with kernel principal component analysis. *J Biomed Biotechnol* 2005, 2005(2): 155-9.
66. Reverter F, Vegas E, Oller JM. Kernel-PCA data integration with enhanced interpretability. *BMC Syst Biol* 2014, 8(S2): S6.
67. Liu X, Yang C, editors. Greedy kernel PCA for training data reduction and nonlinear feature extraction in classification. MIPPR 2009: Automatic Target Recognition and Image Analysis; 2009: International Society for Optics and Photonics.
68. Kohonen T. Self-organized formation of topologically correct feature maps. *Biol Cybern* 1982, 43(1): 59-69.
69. Fakoor R, Ladhak F, Nazi A, Huber M, editors. Using deep learning to enhance cancer diagnosis and classification. Proceedings of the international conference on machine learning; 2013: ACM New York, USA.
70. Kaski S, Nikkilä J, Törönen P, Castren E, Wong G. Analysis and visualization of gene expression data using self-organizing maps. *Neural networks* 2001.
71. Engreitz JM, Daigle Jr BJ, Marshall JJ, Altman RB. Independent component analysis: mining microarray data for fundamental human gene expression modules. *J Biomed Inform* 2010, 43(6): 932-44.
72. Lee S-I, Batzoglou S. Application of independent component analysis to microarrays. *Genome biology* 2003, 4(11): R76.
73. Cao L, Chua KS, Chong W, Lee H, Gu Q. A comparison of PCA, KPCA and ICA for dimensionality reduction in support vector machine. *Neurocomputing* 2003, 55(1-2): 321-36.
74. Wang JJ-Y, Bensmail H, Gao X. Multiple graph regularized nonnegative matrix factorization. *Pattern Recognit Lett* 2013, 46(10): 2840-7.
75. Arslan S, Ozturk C. Feature Selection for Classification with Artificial Bee Colony Programming. Swarm Intelligence-Recent Advances, New Perspectives and Applications: IntechOpen; 2019.
76. Alizadeh AA, Eisen MB, Davis RE, Ma C, Lossos IS, Rosenwald A, Boldrick JC, Sabet H, Tran T, Yu X. Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. *Nature* 2000, 403(6769): 503-11.
77. Alon U, Barkai N, Notterman DA, Gish K, Ybarra S, Mack D, Levine AJ. Broad patterns of gene expression revealed by clustering analysis of tumor and normal

colon tissues probed by oligonucleotide arrays. *Proc Natl Acad Sci USA* 1999, 96(12): 6745-50.
78. Hosmer DW, Lemeshow S. *Applied logistic regression*1989.
79. Ding B, Gentleman R. Classification using generalized partial least squares. *J Comput Graph Stat* 2005, 14(2): 280-98.
80. Zhang S, Li X, Zong M, Zhu X, Cheng D. Learning k for knn classification. *ACM Trans Intell Syst Technol* 2017, 8(3): 1-19.
81. Breiman L, Friedman J, Stone CJ, Olshen RA. *Classification and regression trees*, CRC press 1984.
82. Hastie T, Buja A, Tibshirani R. Penalized discriminant analysis. *Ann Stat* 1995, 73-102.
83. Hastie T, Tibshirani R. Discriminant analysis by Gaussian mixtures. *J Royal Stat Soc* 1996, 58(1): 155-76.
84. Vapnik V. *The nature of statistical learning theory*, Springer science & business media 2013.
85. Kavzoğlu T, Çölkesen İ. Destek vektör makineleri ile uydu görüntülerinin sınıflandırılmasında kernel fonksiyonlarının etkilerinin incelenmesi. *Harita Dergisi* 2010, 144(7): 73-82.
86. Aizerman MA. Theoretical foundations of the potential function method in pattern recognition learning. *First Annu IEEE Conf Control Technol Appl* 1964, 25821-37.
87. Yu W, Liu T, Valdez R, Gwinn M, Khoury MJ. Application of support vector machine modeling for prediction of common diseases: the case of diabetes and pre-diabetes. *BMC Med Inform Decis Mak* 2010, 10(1): 16.
88. Güldoğan E, Arslan AK, Yağmur J. Çeşitli Çekirdek Fonksiyonları ile Oluşturulan Destek Vektör Makinesi Modellerinin Performanslarının İncelenmesi: Bir Klinik Uygulama. *Firat Tip Dergisi* 2017, 22(3).
89. Bergstra JS, Bardenet R, Bengio Y, Kégl B, editors. Algorithms for hyper-parameter optimization. Advances in neural information processing systems; 2011.
90. Stirewalt DL, Meshinchi S, Kopecky KJ, Fan W, Pogosova-Agadjanyan EL, Engel JH, Cronk MR, Dorcy KS, McQuary AR, Hockenbery D. Identification of genes with abnormal expression changes in acute myeloid leukemia. *Genes Chromosomes Cancer* 2008, 47(1): 8-20.
91. Bolstad B, Bolstad MB, BiocGenerics I, biocViews Microarray O, Preprocessing Q. Package 'affyPLM'. 2013.
92. Le Cao K-A, Rohart F, Gonzalez I, Le Cao MK-A. Package 'mixOmics'. 2018.
93. Marchini J, Heaton C, Ripley B, Ripley MB. The fastICA Package. 2007.
94. Friedman J, Hastie T, Tibshirani R. glmnet: Lasso and elastic-net regularized generalized linear models. *R package version* 2009, 1(4).
95. Gentleman R, Carey V, Huber W, Hahne F. Genefilter: Methods for filtering genes from microarray experiments. *R package version* 2011, 1(0).
96. Kuhn MJRFfSC, Vienna, Austria. URL https://cran. r-project. org/package= caret. The caret package. *R J* 2012.
97. Fonti V, Belitser E. Feature selection using lasso. *Amsterdam Research Paper in Business Analytics* 2017, 301-25.
98. Wold S, Esbensen K, Geladi P. Principal component analysis. *Chemometr Intell Lab Syst* 1987, 2(1-3): 37-52.
99. Amsterdam EA, Wenger NK, Brindis RG, Casey DE, Ganiats TG, Holmes DR, Jaffe AS, Jneid H, Kelly RF, Kontos MC. 2014 AHA/ACC guideline for the

management of patients with non–ST-elevation acute coronary syndromes: a report of the American College of Cardiology/American Heart Association Task Force on Practice Guidelines. *JACC CardioOncol* 2014, 64(24): e139-e228.

100. Eskidere Ö. A Comparison Of Feature Selection Methods For Diagnosis Of Parkinson's Disease From Vocal Measurements. *Sigma* 2012, 30402-14.

101. Ben-Hur A, Horn D, Siegelmann HT, Vapnik V. Support vector clustering. *J Mach Learn Res* 2001, 2(Dec): 125-37.

102. Arslan A, Şen B, editors. Detection of non-coding RNA's with optimized support vector machines. 2015 23nd Signal Processing and Communications Applications Conference (SIU); 2015: IEEE.

103. Bergstra J, Bengio Y. Random search for hyper-parameter optimization. *J Mach Learn Res* 2012, 13(1): 281-305.

104. Dyer JO, Dutta A, Gogol M, Weake VM, Dialynas G, Wu X, Seidel C, Zhang Y, Florens L, Washburn MP. Myeloid Leukemia Factor acts in a chaperone complex to regulate transcription factor stability and gene expression. *J Mol Biol* 2017, 429(13): 2093-107.

105. Tomasson MH, Xiang Z, Walgren R, Zhao Y, Kasai Y, Miner T, Ries RE, Lubman O, Fremont DH, McLellan MD. Somatic mutations and germline sequence variants in the expressed tyrosine kinase genes of patients with de novo acute myeloid leukemia. *Blood* 2008, 111(9): 4797-808.

106. Gasparetto M, Pei S, Minhajuddin M, Khan N, Pollyea DA, Myers JR, Ashton JM, Becker MW, Vasiliou V, Humphries KR. Targeted therapy for a subset of acute myeloid leukemias that lack expression of aldehyde dehydrogenase 1A1. *Haematologica* 2017, 102(6): 1054-65.

107. Silveira VS, Scrideli CA, Moreno DA, Yunes JA, Queiroz RG, Toledo SC, Lee MLM, Petrilli AS, Brandalise SR, Tone LG. Gene expression pattern contributing to prognostic factors in childhood acute lymphoblastic leukemia. *Leuk Lymphoma* 2013, 54(2): 310-4.

108. Beutler E, Kuhl W, Comings D. Hexosaminidase isozyme in type O Gm2 gangliosidosis (Sandhoff-Jatzkewitz disease). *American Journal of Human Genetics* 1975, 27(5): 628.

109. Haferlach C, Bacher U, Kohlmann A, Schindela S, Alpermann T, Kern W, Schnittger S, Haferlach T. CDKN1B, encoding the cyclin-dependent kinase inhibitor 1B (p27), is located in the minimally deleted region of 12p abnormalities in myeloid malignancies and its low expression is a favorable prognostic marker in acute myeloid leukemia. *Haematologica* 2011, 96(6): 829-36.

110. Kapelko-Slowik K, Owczarek TB, Grzymajlo K, Urbaniak-Kujda D, Jazwiec B, Slowik M, Kuliczkowski K, Ugorski M. Elevated PIM2 gene expression is associated with poor survival of patients with acute myeloid leukemia. *Leuk Lymphoma* 2016, 57(9): 2140-9.

111. Dvorak A, Letourneau L, Weller P, Ackerman S. Ultrastructural localization of Charcot-Leyden crystal protein (lysophospholipase) to intracytoplasmic crystals in tumor cells of primary solid and papillary epithelial neoplasm of the pancreas. *Lab Invest* 1990, 62(5): 608-15.

EKLER

EK-1. Özgeçmiş

ÖZGEÇMİŞ

Adı Soyadı: Fatma Hilal YAĞIN

Doğum Tarihi: 1990

Öğrenim Durumu: Yüksek Lisans

Derece	**Bölüm/Program**	**Üniversite**	**Yıl**
Lisans	İstatistik	Gazi Üniversitesi	2017
Y. Lisans	Biyoistatistik ve Tıp Bilişimi AD	İnönü Üniversitesi	2020

Yüksek Lisans Tez Başlığı ve Tez Danışman(lar)ı:

Gen İfade Veri Setlerinde Boyut İndirgeme Yöntemlerinin Sınıflama Performansına Etkilerinin Karşılaştırılması, Doç. Dr. Harika Gözde GÖZÜKARA BAĞ

Görevler:

Görev Unvanı	**Görev Yeri**	**Yıl**
Arş. Gör.	İnönü Üniversitesi	2019 – Devam ediyor

EK-2. Etik Kurul Almama Gerekçesi

Evrak Tarih ve Sayısı: 06/09/2018-E.66849

T.C.
İNÖNÜ ÜNİVERSİTESİ REKTÖRLÜĞÜ

Tıp Fakültesi Dekanlığı
Biyoistatistik Anabilim Dalı Başkanlığı

Sayı : 58137357-020
Konu : Fatma Hilal UZUNOĞLU Tez Öneri Formu

SAĞLIK BİLİMLERİ ENSTİTÜSÜ MÜDÜRLÜĞÜNE

Enstitünüz 38173842004 numaralı yüksek lisans öğrencisi Fatma Hilal UZUNOĞLU'nun Yüksek Lisans Tez Öneri Formu ektedir.
Gereğini arz ederim.

e-imzalıdır
Prof.Dr. Saim YOLOĞLU
Anabilim Dalı Başkanı

Ek:1 Adet Tez Öneri Formu

Tıp Fakültesi Dekanlığı
Telefon No: 3410660 Faks No: 3410036
E-Posta: biyoistatistik@inonu.edu.tr İnternet Adresi:
https://www.inonu.edu.tr/tr/cms/biyoistatistik

Bilgi İçin: Şeyma YAŞAR
Unvan: Öğretim Elemanı
Telefon No: 3410660

EK-2. Etik Kurul Almama Formu (Devamı)

SAĞLIK BİLİMLERİ ENSTİTÜSÜ MÜDÜRLÜĞÜNE

13 Nisan 2013 tarih ve 28617 sayı ile T.C. Resmi Gazetede yayınlanan " Klinik Araştırmalar Hakkında Yönetmelik'in birinci bölümünün 2. maddesinin 1. fıkrası (Bu yönetmelik biyoyararlanım ve biyoeşdeğerlik çalışmaları dahil, ruhsat veya izin alınmış olsa dahi insanlar üzerinde yapılacak olan ilaç, tıbbi ve biyolojik ürünler ile bitkisel ürünlerin klinik araştırmaları, klinik araştırma yerlerini ve bu araştırmaları gerçekleştirecek gerçek veya tüzel kişileri kapsar.) gereğince yüksek lisans öğrencisi Fatma Hilal UZUNOĞLU' nun tezinin klinik bir çalışma olmaması, kullanılacak olan verinin web sitesinde yayınlanmış veri tabanından elde edilecek olması sebebiyle Etik Kurul kararı alınmamıştır.