TÜRKİYE CUMHURİYETİ
DİCLE ÜNİVERSİTESİ
SAĞLIK BİLİMLERİ ENSTİTÜSÜ

MAKİNE ÖĞRENİMİ YÖNTEMLERİNİ KULLANARAK EVRE III İNVAZİV DUKTAL KARSİNOMLU HASTA VERİLERİNİN SINIFLANDIRILMASI

Emre DİRİCAN

DOKTORA TEZİ

BİYOİSTATİSTİK ANABİLİM DALI

DANIŞMAN

Prof. Dr. Zeki AKKUŞ

DİYARBAKIR - 2019

TÜRKİYE CUMHURİYETİ
DİCLE ÜNİVERSİTESİ
SAĞLIK BİLİMLERİ ENSTİTÜSÜ

MAKİNE ÖĞRENİMİ YÖNTEMLERİNİ KULLANARAK EVRE III İNVAZİV DUKTAL KARSİNOMLU HASTA VERİLERİNİN SINIFLANDIRILMASI

Emre DİRİCAN

DOKTORA TEZİ

BİYOİSTATİSTİK ANABİLİM DALI

DANIŞMAN

Prof. Dr. Zeki AKKUŞ

DİYARBAKIR - 2019

TÜRKİYE CUMHURİYETİ
DİCLE ÜNİVERSİTESİ
SAĞLIK BİLİMLERİ ENSTİTÜSÜ

ONAY

Dicle Üniversitesi Sağlık Bilimleri Enstitüsü **Biyoistatistik** Anabilim Dalı Doktora öğrencisi **13854301** nolu **Emre DİRİCAN**'ın hazırladığı **"Makine Öğrenimi Yöntemlerini Kullanarak Evre III İnvaziv Duktal Karsinomlu Hasta Verilerinin Sınıflandırılması"** başlıklı tez Dicle Üniversitesi Lisansüstü Eğitim - Öğretim ve Sınav Yönetmeliğinin ilgili maddeleri uyarınca kapsam ve bilimsel kalite yönünden değerlendirilerek Doktora Tezi olarak kabul edilmiştir.

Tarih: 25 / 02 / 2019

Danışman: Prof. Dr. Zeki AKKUŞ

Jüri Üyeleri		İmza
Jüri Başkanı	**Prof.Dr. Saim YOLOĞLU**	
Üye	**Prof.Dr. Sıddık KESKİN**	
Üye	**Prof.Dr. Ömer SATICI**	
Üye	**Prof.Dr. Zeki AKKUŞ**	
Üye	**Doç.Dr. Mahmut BALKAN**	

Bu tez Dicle Üniversitesi Sağlık Bilimleri Enstitüsü Yönetim Kurulu'nun .../.../20.. tarih ve ..... sayılı kararıyla onaylanmıştır.

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

Prof. Dr. Hakkı Murat BİLGİN
Dicle Üniversitesi
Sağlık Bilimleri Enstitüsü Müdürü

TÜRKİYE CUMHURİYETİ
DİCLE ÜNİVERSİTESİ
SAĞLIK BİLİMLERİ ENSTİTÜSÜ

BEYAN

Bu tez çalışmasının kendi çalışmam olduğunu, tezin planlanmasından yazımına kadar bütün safhalarda etik dışı davranışımın olmadığını, bu tezdeki bütün bilgileri akademik ve etik kurallar içinde elde ettiğimi, bu tez çalışmasıyla elde edilmeyen bütün bilgi ve yorumlara kaynak gösterdiğimi ve bu kaynakları da kaynaklar listesine aldığımı, yine bu tezin çalışılması ve yazımı sırasında patent ve telif haklarını ihlal edici bir davranışımın olmadığını ve tezimi Dicle Üniversitesi Sağlık Bilimleri Enstitüsü Tez Yazım Kılavuzu standartlarına uygun bir şekilde hazırladığımı beyan ederim.

11/03/2019

Emre DİRİCAN

TEŞEKKÜR

Tez çalışmalarım ve doktora eğitim sürecimin her aşamasında desteğini hissettiğim; yakın ilgisi ve sabrından dolayı değerli hocam ve danışmanım Sayın Prof. Dr. Zeki AKKUŞ 'a,

ve

bugünlere gelmemde büyük emekleri olan değerli Anneme ve Babama; yüreklendirici desteği ve güveniyle hep varlığını hissettiğim eşim Ayşe Gül 'e ve çalışmalarım nedeniyle zaman zaman ihmal ettiğim biricik oğlum Mehmet İlbey 'e en içten teşekkürlerimi sunarım…

Emre DİRİCAN

İÇİNDEKİLER

KISALTMALAR VE SİMGELER LİSTESİ

A^T A matrisinin transpozu

$K(x_i, x_j)$ Çekirdek fonksiyon

$\varphi(x)$ Aktivasyon fonksiyonu

$\xi(x)$ Tek katmanlı YSA eşik fonksiyonu

$\hat{y}$ Sınıflamada *Y* için olası değerler kümesi

||w||: w'nun öklid normudur.

$R_{den}[f]$ f fonksiyonu için deneysel risk

R(f) Risk fonksiyonu

$\phi(h, m, \delta)$ Risk hesabı için güvenirlik terimi

$\varepsilon_i (\geq 0)$ Gevşek (slack) değişken

m_{try} Dallara ayırıcı karar değişkeni sayısı

ntree Modelleme için kullanılacak karar ağacı sayısı

OOB Her bir eğitim örneğinde ortalama tahmin hatası

ML Makine öğrenimi

KKT Karush- Kuhn-Tucker koşulları

DVM Destek vektör makineleri

RF Rasgele orman

YSA Yapay sinir ağları

MLP Çok katmanlı algılayıcılar

VC Vapnik-Chervonenkis boyutu

YRM Yapısal risk minimizasyonu

DRM Deneysel risk minimizasyonu

ŞEKİL VE TABLOLAR LİSTESİ

1. ÖZET

Makine Öğrenimi Yöntemlerini Kullanarak Evre III İnvaziv Duktal Karsinomlu Hasta Verilerinin Sınıflandırılması

Öğrencinin Adı ve Soyadı: Emre DİRİCAN

Danışman: Prof. Dr. Zeki AKKUŞ

Anabilim Dalı: Biyoistatistik

1.1. Türkçe Özet

Amaç: Tez çalışmasında, danışmanlı makine öğrenimi yöntemleri sınıflama performansına göre kıyaslanarak çalışmada kullanılan yöntemlerin içerisinden, sınıflama başarısı yüksek olan yöntemin bulunması amaçlandı.

Gereç ve Yöntem: Çalışmamızda, destek vektör makineleri, rasgele orman ve yapay sinir ağları yöntemlerinin sınıflama performanslarını kıyaslamak için meme kanseri türlerinden biri olan invaziv duktal karsinomlu 302 hastanın sağkalım bilgilerini içeren veri seti ile birlikte simülasyonla elde edilen 24 farklı veri seti kullanıldı. Kullanılan yöntemlerin sınıflama başarıları meme kanseri verilerinde genel doğruluk, duyarlılık, seçicilik, F-ölçütü, Matthews korelasyon kastsayısı, AUC ve ayırsama gücüne göre kıyaslandı. Simülasyon verilerinde ise eğitim-test doğrulukları farkı ve bu farkın anlamlılığı değerlendirildi.

Bulgular: İnvaziv duktal karsinom evre III hastalarının test seti için en yüksek sağkalım sınıflama doğruluğu (%80) ve yüksek performans değerlendirme kriteri değerleri radyal çekirdekli destek vektör makinelerinden elde edildi. Simülasyon verilerinde de yüksek sınıflama doğruluğu ve doğruluklar arasındaki farkın küçüklüğü genel olarak destek vektör makinelerinden elde edildi.

Sonuç: Destek vektör makineleri hem gerçek veri setinde hem de simülasyon verilerinde, rasgele orman ve yapay sinir ağlarına göre daha yüksek doğruluk oranına sahiptir.

Anahtar Sözcükler: Makine öğrenimi, sınıflama, meme kanseri, destek vektör makineleri, rastgele orman.

Classification of Patients with Stage III Invasive Ductal Carcinoma Using Machine Learning Methods

Student's Surname and Name: DİRİCAN Emre

Adviser of Thesis: Prof. Dr. Zeki AKKUŞ

Department: Biostatistics

1.2. Abstract

Aim: In the thesis study, it was aimed to find the method with high classification success among the methods used in the study by comparing the machine learning methods according to the classification performance.

Material and Method: In our study, the data set of 302 patients with invasive ductal carcinoma, one of the breast cancer types, and 24 different data sets obtained by simulation were used to compare the classification performances of support vector machines, random forest and artificial neural network methods. The success of classifications of the methods used were compared according to the general accuracy, sensitivity, specificity, F-measure, Matthews correlation coefficient, AUC and discriminant power in breast cancer data. In the simulation data, the difference between train-test accuracy and the significance of this difference were evaluated.

Results: The highest survival classifying accuracy (80%) and high performance evaluation criterion values of the test set of invasive ductal carcinoma stage III patients were obtained from radial kernel support vector machines. In the simulation data, the high classification accuracy and the small difference between accuracies were obtained from the support vector machines in general.

Conclusion: Support vector machines have higher accuracy in both the real data set and simulation data than random forest and artificial neural networks.

Key Words: Machine learning, classification, breast cancer, support vector machines, random forest.

2. GİRİŞ ve AMAÇ

Günlük hayatta gerçekleştirilen işlemlerin büyük bir kısmının bilgisayarla yapılması büyük boyutlarda veri kaydı oluşturmaktadır. Küresel telekomünikasyon şebekeleri her gün onlarca petabayt veriyi taşımaktadır. (1 petabyte=1024 terrabyte). Tıp ve sağlık endüstrisi, tıbbi kayıtlar, hasta izleme ve tıbbi görüntüleme yöntemleriyle çok büyük miktarda veri üretmektedir. Arama motorları tarafından desteklenen milyarlarca web araması günde on petabyte veri işlemektedir. Sosyal medya, dijital resimler ve videolar, bloglar, web toplulukları ve çeşitli sosyal ağlar önemli veri kaynakları haline gelmiş bulunmaktadır. Büyük miktarda veriyi faydalı bilgiye dönüştürmek için yardımcı olabilecek yeni teknikler ve otomatik modelleme araçları acil bir ihtiyaç alanı haline gelmiştir. Bu durum, veri madenciliği adı verilen bilgisayar biliminde umut verici aynı zamanda sürekli gelişen yöntemler topluluğunun ve onun çeşitli uygulamalarının ortaya çıkmasına yol açmıştır. Yakın bir tarihte ortaya çıkan ve 1990'lı yıllarda büyük ilerleme kaydederek günümüzde gelişmeye devam eden veri madenciliği Guidici tarafından; "*Büyük miktardaki veriden, veri tabanı sahibine açık ve faydalı bilgi sağlamak amacı ile başlangıçta bilinmeyen örüntü ve ilişkileri keşfetmek için kullanılan seçme, keşfetme ve modelleme süreci"* olarak tanımlanmıştır (1).

Veri madenciliği tekniklerinden biri olan makine öğrenimi yaklaşımı, bilgisayarların nasıl öğrenebildiğini ya da performansının nasıl geliştirilebileceğini incelemektedir (2). Bu amaçla makine öğrenimi pek çok kullanışlı yönteme ve uygulama alanına ayrılmıştır. Özellikle modelleme ve bu modellemeye göre sınıflama yapma; son yıllarda tıp, fen ve mühendislik alanında yaygın olarak kullanılmaktadır. (3). Tıp alanında son yıllarda yayınlanan birçok çalışmanın analizinde makine öğrenimi yöntemlerinin kullanıyor olması ve görsel yaklaşımlarıyla daha anlamlı sonuçların edinilebilirliği, bu alanda tez çalışması yapmayı gerekli hale getirmiştir.

Meme kanserli hastaların sağkalım analizi ve bu verileri değerlendirmek için kurgulanan danışmanlı makine öğrenimi yöntemleri birçok çalışmada işlenmiştir. Bu çalışmalarda meme kanserinin birkaç türü ve evreleri birlikte değerlendirilmiştir.

Ancak tez çalışmasının birinci bölümünde işlenecek olan invaziv duktal karsinomlu* üçüncü evre hastalarının sağkalım verilerinden oluşan analize herhangi bir çalışmada rastlanamamıştır. Dolayısıyla çalışma bu alanda yeni bir yaklaşım olarak değerlendirilmektedir. Veri analizinin ikinci bölümünde bağımlı değişkendeki farklı risk görülme oranlarına göre simüle edilen veriler üzerinden model kıyaslamaları yapılmıştır. Simülasyon verileriyle model karşılaştırmaları, herhangi bir sınıflama güdümü olmayan çok değişkenli normal dağılımdan türetilmiş benzer bağımsız değişkenlerle yapılan sınıflamaya göre yöntemlerin performanslarını değerlendirme amacıyla yapılmıştır. Bu amaçla makine öğrenimi yöntemlerinin doğru sınıflama yüzdelerine ve bazı başarı değerlendirme kriterlerine yer verilmiştir. Simüle verilerin analizinde ön plana çıkan yaklaşım, eğitim ve test seti doğruluk uyumunun model performansı ile ilişkili olabileceğidir.

Çalışmamızda üç makine öğrenimi yöntemi kullanılmıştır. Meme kanseri verilerinde dört faklı çekirdek fonksiyon ile destek vektör makineleri, farklı m_{try} ve ağaç sayısıyla rasgele orman ve dört faklı aktivasyon fonksiyonu ile ileri beslemeli geri yayılımlı yapay sinir ağları kullanılmıştır. Simülasyon verilerinde ise radyal tabanlı destek vektör makineleri, değişken sayısına göre belirlenmiş m_{try} ile rasgele orman ve sigmoid aktivasyon fonksiyon ile yapay sinir ağları kullanılmıştır. Söz konusu yöntemlerin doğru sınıflama oranlarına ve model başarı ölçütleri değerlendirmelerine göre sınıflama başarısı yüksek yöntem bulunmaya çalışılmıştır.

* Meme kanseri türlerinden biri olan invaziv duktal karsinom, meme kanseri vakalarının %80 'nini teşkil etmektedir.

3. GENEL BİLGİLER

Büyük miktarda ve boyutta veri üreten sonsuz sayıda kaynak listesi bulunmaktadır. Bu kadar verinin kullanılabilir olması, akıllı veri analizini teknolojik ilerleme için gerekli bir bileşen haline getirmektedir. Her kurum-kuruluş, verilerini toplama ve analiz etme avantajından yararlanabilmektedir. Hastaneler, hasta kayıtlarındaki eğilimleri ve anormallikleri bulabilmekte, arama motorları daha iyi sıralama ve reklam yerleşimini yapabilmektedir (2). Bir arama motoruna sorgu gönderme işleminde, önce sorguyla alakalı web sayfaları bulunur sonra onlar ilişki düzeyine sıralanır. Bu amaca ulaşmak için arama motorunun bilmesi gerekenler, hangi sayfaların ilişkili olduğu ve hangi sayfaların sorguyla eşleştiğidir. Bir diğer örnek birçok güvenlik kontrolünde kullanılan yüz tanıma sistemi ve erişim kontrolüdür. Bir kişinin fotoğrafı (veya video kaydı) eldeki veri ise o kişinin kim olduğunu tanımak için kullanılan uygulamalar düşünüldüğünde, sistemin yüzleri pek çok kategoriden birine sınıflandırması veya bilinmeyen bir yüz olup olmadığına karar vermesi gerekmektedir. Burada evet/hayır sorusuna yanıt aranır. Makine öğreniminden beklenen; kişinin gözlük takması-saç modelini değiştirmesi durumunda, farklı aydınlatma koşulları ve ilişkisiz yüz ifadeleri ile bir kişinin tanımlanmasında hangi özelliklerin bulunduğunu öğrenen bir sistemi kurgulamasıdır (3). Makine öğrenimi, veri madenciliğinin bir uygulama alanı teknolojisi olmasına rağmen kullanılan yöntemler bakımından veri madenciliğine benzerlik göstermektedir. Yöntemlerin birbirinden ayrıldığı temel unsur; makine öğrenimi bilinen özelliklerden sonuca ulaşırken veri madenciliği bilinmeyen ilişki ve örüntülerin keşfini yapmaktadır.

Veri madenciliği süreci aşağıdaki adımlarla gerçekleştirilir:

Problemin belirlenmesi; çalışmanın amacı belirlenir ve planlaması yapılır.

Verinin hazırlanması; bu aşamada araştırıcıya düşen pek çok görev vardır. Bu görevler;

- Farklı kaynaklardan alınan verilerin birleştirilmesi

- Verinin temizlenmesi (aykırı gözlemlerin ve kayıp değerlerin düzenlenmesi)
- Veriye gerekli dönüşümlerin yapılmasıdır.

Modelleme; uygun modelleme algoritması seçilir ve uygulanır. Optimum sonuç için model parametreleri düzenlenir.

Değerlendirme; oluşturulan modelin etkinliği ve performansı değerlendirilir.

Raporlama; elde edilen sonuçlar görselleme teknikleri ve verilmesi gereken argümanlarla sunulur.

Genel olarak veri madenciliği, belirlenen veriler bir hedef için anlamlı olduğu sürece her türlü veriye uygulanabilir. Veri tabanlarından (Microsoft Access ve MySQL) alınabilen veriler, veri ambarlarından edinilebilen ETL süreçli veriler (ETL: *Extract;* veriyi kaynak sistemden alma, *Transform;* gerekli dönüşümlerin yapılması, *Load;* verilerin hedef sisteme yüklenmesi), işlemsel veriler, metin verileri ve multimedya verileri veri setlerini örneklendirmektedir. Veri madenciliği işlevleri genel olarak tanımlama ve tahminleme olmak üzere iki ana bölüme ayrılmaktadır. Tanımlayıcı işlevler, verinin genel niteliklerini karakterize etmektedir. Tahmin edici işlevler ise veriyi kullanarak çıkarsamalarda bulunmakta ve kestirim yapmaktadır.

3.1. Tanımlayıcı İşlevler

Tanımlama, veri hakkında bilgi edinmek ve verileri karakterize etmek amacıyla yapılmaktadır. Tanımlama işlemi sonunda, verilerin gizli bağıntıları, ilişkileri ve birliktelikleri açıkça tanımlanabilmelidir. Veri madenciliğinde tanımlamada kullanılan yöntemler; kümeleme analizi, faktör analizi, uyum analizi, birliktelik kuralları ve aykırı değer analizidir.

3.2. Tahmin Edici İşlevler

Tahmin işlemi, bir tahminleyici üretilebilmesi için bütün örüntüler ve gizli kalmış bağlantılar kullanılarak gözlenmiş veriler üzerinden model oluşturulmasını sağlamaktadır. Daha sonra yeni bir gözlem için, tahminleyiciden faydalanılarak yanıt

değişkeni tahmin etmektir. Yanıt değişkeni kategorik bir değişken olduğu zaman tahminleme işlemi sınıflama işlemi adını almaktadır. Tahmin edici işlevler sınıflandırma ve regresyon olmak üzere genel olarak iki kısımdan oluşur. Bu kısımda sınıflandırmaya değinilecektir.

3.3. Sınıflandırma

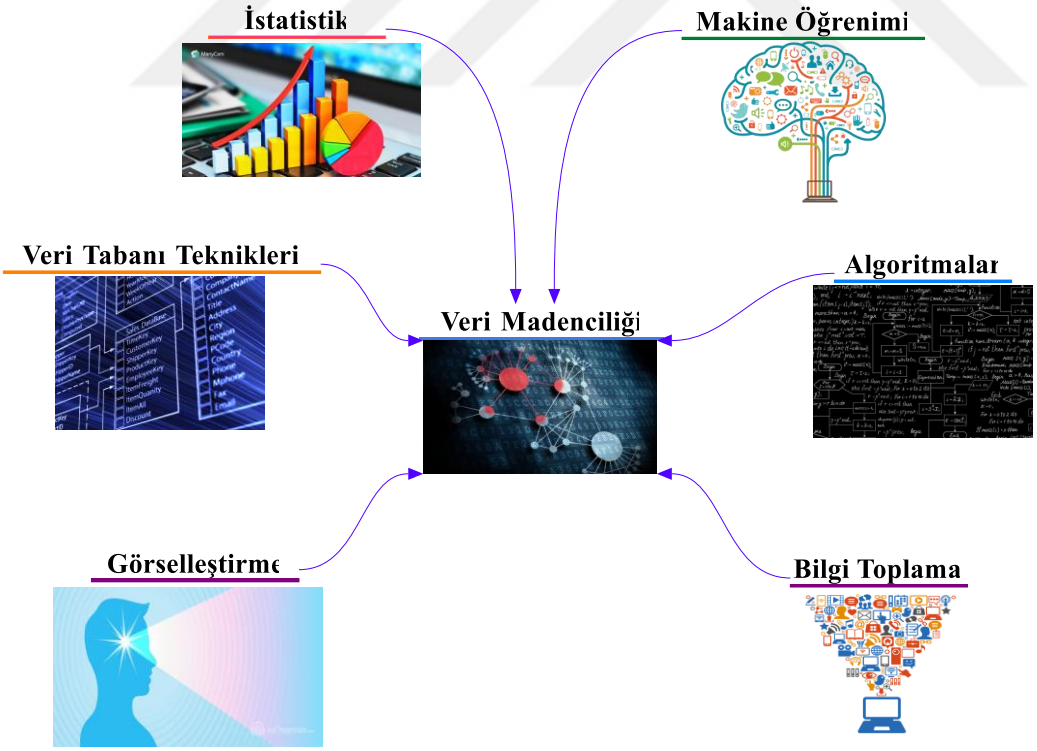
Sınıflandırma, veri sınıflarını tanımlayan ve ayıran bir model (veya fonksiyon) bulma işlemidir. Model, bir takım eğitim verilerinin (yani sınıf etiketleri bilinen verilerin) analizine dayanılarak türetilir. Model, sınıfları bilinmeyen gözlemlerin sınıf etiketlerini tahmin etmek için kullanılır (2). Sınıflama veri madenciliğinde en çok kullanılan tahmin edici tekniklerden biridir. Sınıflama işlemi genel olarak veriyi sadeleştirmek ve sınıfı bilinmeyen gözlemler için kestirim yapmakta kullanılır. Verilerin sınıflandırılması iki adımda gerçekleşir; ilk adımda, etiketlenmiş olan sınıfı tahmin etmek için kullanılacak bir model oluşturulur. Bu adımda gözlemlerden öğrenme yapıldığı için, öğrenme adımı olarak da tanımlanmaktadır. Bir sonraki adımda oluşturulan model, sınıfı belli olmayan yeni gözlemlere uygulanarak tahmin işlemi yapılır (4). Sınıf etiketleri {*0,1*} değerlerini alabileceği gibi {*var-yok*} değerlerini de alabilmektedir. Önemli olan etiketlerin biri sınıfa dahil olma durumunu, diğeri ise sınıfa dahil olmama durumunu gösteren iki farklı kategori olmasıdır. İkili sınıflandırma problemi için özetlenen bu tanımlama, uygun sayıda farklı etiketler belirtilmek koşuluyla çoklu sınıflandırma için genişletilebilir (5).

Sınıflandırma yanlılığının önüne geçebilmek için her bir sınıf kategorisinin birbirinden bağımsız ve eşit önemlilikte olduğu kabul edilir. Ancak bu durum veride bazı bağımlılıklar varsa yeniden değerlendirilebilir. Bir sınıflandırma algoritması, veri setinin yığınları (batchs) ihtiva ettiğini öğrenirse bir yığının bitip diğer yığının başladığı nokta(ları) belirleyerek yüksek doğruluk değerine ulaşır. Örneğin, bir hekim bulaşıcı hastalığa yakalanan bireylerin yığınlar halinde hastaneye başvurduklarını bilir. Dolayısıyla art arda belli bir bulaşıcı hastalığa yakalanan birkaç hastayı gördükten sonra, büyük ihtimalle bir sonraki hastayı semptomları önceki hastalar kadar belirgin olmasa bile söz konusu bulaşıcı hastalığa yakalanmış birey

olduğu yönünde sınıflandırma eğilimi gösterebilir (6). Ayrıca objektif sınıflamadan kaynaklanan yüksek maliyetlerin veya alınmak istenmeyen risklerin olduğu durumlarda sınıfların eşit önemlilikte olması durumu gözardı edilebilir. Sınıflama yöntemlerinden bazıları; karar ağaçları, naive bayes, k-en yakın komşuluk, yapay sinir ağları, destek vektör makineleri, genetik algoritmalar ve rasgele ormandır.

3.4. Veri Madenciliğinde Kullanılan Teknikler

Uygulamaya dayalı bir alan olan veri madenciliği, birçok tekniği bünyesinde barındırmaktadır. İstatistik, makine öğrenimi, veri tabanları ve veri ambarı sistemleri, bilgi toplama, görselleştirme, algoritmalar, yüksek performans hesaplama gibi pek çok uygulama alanına ve (Şekil 1) diğer alanlardan elde edilen verilere dayanmaktadır. Veri madenciliği, araştırma ve geliştirmenin disiplinlerarası yapısına ve yaygın uygulama alanlarına katkıda bulunmaktadır.



Şekil 1: Veri madenciliğinde kullanılan teknikler

Veri madenciliği bankacılık, pazarlama, sigortacılık, telekomünikasyon, borsa, sağlık, endüstri, astronomi, kalite kontrolü, genetik ve mühendislik gibi birçok alanda kullanılırken sağlık alanında da çeşitli amaçlar için kullanılmaktadır.

Bu amaçlar;

- Hastalıklara etki eden faktörlerin ortaya çıkartılması
- Hastalıklara erken teşhis koyabilmenin önünü açacak etkenlerin belirlenmesi
- Sağlık hizmetlerindeki kalitenin arttırabilmesi için eksikliklerin belirlenmesi
- Sağlık harcamalarındaki hileli işlemlerin ortaya çıkartılması ve maliyeti düşürülebilecek faktörlerin belirlenmesi
- İlaç geliştirici firmaların, sağlık veri tabanlarından yararlanarak doğru ilaçları geliştirilebilmesi için alt yapı hazırlanmasıdır.

3.5. Makine Öğrenimi

Günlük hayatın bir parçası olmaya aday makine öğrenimi (ML), son yirmi yılda bilgi teknolojisinin en önemli dayanak noktalarından biri haline gelmiş bulunmaktadır. ML bilgisayarların veriler üzerinden nasıl bir öğrenme (veya performanslarını nasıl geliştireceğini) gerçekleştireceğini araştırır. Ana araştırma alanı, bilgisayar programlarının karmaşık kalıplarını otomatik olarak tanımak ve veriye dayalı akıllı karar vermeyi öğrenmektir. Örneğin tipik bir ML problemi, bir dizi örnekten bağıntıları öğrendikten sonra, postadaki el yazısıyla yazılan posta kodlarını otomatik olarak algılayacak şekilde bir bilgisayarı programlamaktır (2). ML, *"bilgisayarın bir dizi ile ilgili bağıntı ve tecrübeleri öğrenerek, gelecekte oluşacak benzer durumlar için kararlar verebilmesi ve problemlere çözümler üretebilmesi"* olarak da tanımlanabilir (7). ML 'yi başlangıçta çok fazla benimsemeyen istatistik bilimciler, yüksek boyutlu verilerle karşılaştıklarında bu karmaşık veri setlerine ve problemlere çözüm üreten öğrenme algoritmalarını ve Vapnik-Chervonenkis tarafından öne sürülen "istatistiksel öğrenme" yöntemlerini benimsemişlerdir.

ML yöntemleri veri madenciliğinin aksine; veri setlerinin büyük olması gerektirmemektedir. Küçük veri setleri için de başarılı sonuçlar vermesinin yanında yeni gözlemler için kendi kendine öğrenebilen modeller oluşturabilmektedir. ML yöntemleri büyük boyuttaki veri setlerine uygulanmak üzere modeller oluşturmakta ve bu modellerle veri madenciliği örüntülerine alt yapı hazırlamaktadır. ML yöntemiyle oluşturulan modeller veri madenciliğinde kullanıldığında, bu modellerlerden öğrenme beklenmemesi gerekmektedir. Bu kısımda veri madenciliğinde de büyük pay sahibi olan ve aynı zamanda ML 'de de kullanılan öğrenme yöntemlerine yer verilecektir.

Bu yöntemler;

- Eğiticili (supervised) öğrenme
    - Sınıflama
    - Regresyon
- Eğiticisiz (unsupervised) öğrenme
- Yarı-eğiticili (semi-supervised) öğrenme
- Aktif (active) öğrenmedir.

Kullanılan veri türünü belirlemek öğrenme problemlerini karakterize etmede araştırıcıya fayda sağlayacaktır. Zira araştırmalar genellikle benzer veri türlerinden oluşabildiğinden çözümleri de benzer tekniklerden oluşabilmektedir. Değişkenlerin sistemleştirildiği vektörler, çalışmalar için temel olgulardır (2).

3.5.1. Eğiticili (supervised) öğrenme

Eğiticili ya da diğer adıyla danışmanlı öğrenme yöntemlerinde model, girdi olarak ele alınan veri setinin çıktı olarak nitelendirilen gözlem setiyle olan ilişkisini belirler. Bu öğrenme türünde araştırıcı tarafından belirlenmiş bir yanıt değişken bulunmaktadır. Bu yöntemde eğitici, sisteme öğrenilmesi istenen durum ile ilgili yapıları eğitim seti olarak verir. Girilen gözlemlerle modellenen değerler arasındaki fark (hata değeri) belirlenen değerden küçük oluncaya modelleme eğitimine devam

edilir. Bu öğrenme yönteminin öne çıkan bazı algoritmaları karar ağaçları, regresyon yöntemleri, sinir ağları, destek vektör makineleri ve k-en yakın komşuluktur (2).

3.5.2. Eğiticisiz (unsupervised) öğrenme

Bu öğrenme yöntemi kümeleme analizine oldukça benzerdir. Bu yöntemde önceden belirlenmiş sınıflar bulunmamaktadır. Dolayısıyla yöntem benzer giriş bilgileri ve veriler arasındaki uzaklığı/benzerliği kullanarak kendi kendine öğrenme gerçekleştirmektedir. Kümeleme analizi, faktör analizi, korelasyon analizi ve uyum analizi bu öğrenme yöntemine örnek olarak verilebilir (8).

3.5.3. Yarı-eğiticili (semi-supervised) öğrenme

Yarı eğiticili öğrenmede, öğrenme makinesi hem sınıflanmış hem de sınıflanmamış gözlemlerden oluşan eğitim kümesini alır ve eğitim setinde yer almayan yeni gözlemler için atamalar yapar. Sınıflama modellerini kurgulayabilmek için sınıflanmış veriler ve sınıf sınırlarını düzenleyebilmek için sınıflanmamış veriler kullanılır. Sınıflama ve regresyonun bazı türleri bu yöntem ile çalışmaktadır (9).

3.5.4. Aktif (active) öğrenme

Aktif öğrenme, araştırıcıların öğrenme sürecinde aktif rol almalarını sağlayan bir makine öğrenimi yaklaşımıdır. Aktif öğrenme yaklaşımı, araştırıcıdan sınıflanmamış örnekleri ya da öğrenme programıyla sentezlenmiş bir örneği sınıflandırmasını isteyebilir. Bu öğrenme yönteminde amaç, araştırıcıyı öğrenme sürecinde aktif hale getirmektir (2). Genellikle sınıflamanın maliyetli olduğu durumlarda kullanılır.

ML ’de uygulama için pek çok teknik bulunmakta ve bu yöntemlere her yıl yenileri eklenmektedir. Bu uygulamada araştırıcıyı bekleyen temel görev, veri setine uygun algoritmanın seçimidir. Bu seçim işlemi için araştırıcının önemle üzerinde durması gereken üç ana başlık bulunmaktadır. Bunlar; temsil etme (representation), değerlendirme (evaluation) ve optimizasyondur.

Representation; bu aşamada araştırıcı veri setini bilgisayar ortamına uygun olacak şekilde düzenler, gerekli değişkenleri ve hipotezini belirler.

Evaluation; sınıflama performansı iyi olanı kötü olandan ayırt edebilmek için gerekli olan fonksiyonun belirlendiği aşamadır. Bu değerlendirme, fonksiyondan optimizasyona geçişi sağlayan algoritmanın kendi içinde oluşturduğu içsel bir değerlendirmedir.

Optimizasyon; en yüksek skoru alan sınıflandırıcının belirlendiği aşamadır. Optimizasyon tekniğinin seçimi, araştırıcının verimliliğinin göstergesidir. Ayrıca değerlendirme fonksiyonunun birden fazla optimum içeriğe sahip olması durumunda üretilen sınıflandırıcının belirlenmesine yardımcı olur (10)**.**

ML 'nin kurgusu, karmaşıklığın içinden basit prototiplerle anlaşılır ve güvenli sonuçlar verme stratejisine dayanmaktadır. Örneğin, bir hayat sigortası şirketi potansiyel bir müşterinin ömrünü tahmin etmek için tansiyon, kalp hızı, boy, kilo, kolesterol düzeyi, sigara içme durumu, cinsiyet değişkenlerinden oluşan vektörle çalışmasına yön verebilmektedir. Ancak vektörlerle yapılan işlemlerin dezavantajlarından biri vektörü oluşturan değişkenlerin ölçeklerinde meydana gelebilecek farklılıktır. Böyle durumlarda yaygın yaklaşımlarından biri verilerin normalleştirilmesidir.

3.6. İstatistiksel öğrenme teorisi

İstatistiksel öğrenme teorisi ML 'nin arkasındaki temel teoremdir. Tümevarımsal nedenler genellikle tümdengelim nedenlerle kıyaslanır. Tümdengelim metoduyla yapılan analizlerde asıl ilgilenilen sonuçlardır ve sonuçların doğruluğu garanti edilir. Ancak tümevarım metotlarda sonuçlardan çok hangi metotlar için garanti verileceği önemlidir. Tümevarım problemini belirleyen çeşitli paradigmatik yaklaşımlar vardır. Örneğin, Reichenbach uzun vadede herhangi bir şey işe yarayacaksa, belirlenen süreç için tümevarımın daha faydalı olacağını savunmaktadır. Buradaki paradigma, sınırsız bir veri akışı olduğunda her sonlu veri seti için yöneltilen bir soruya cevap öneren bir M metodunu öngörmektir. Örneğin bir analizde veri setinin alfabenin bir dizi harfinden oluştuğunu ve diziye "A"

harfinden sonra “B” harfi gelenleri bulma görevinin yüklendiğini düşünelim. Burada ilgilenilmesi gereken, belirlenen her durum için soruyu cevaplamaya yeterli bir yöntem ve optimum sonuçları verebilecek daha başarılı bir yöntem olup olmadığıdır. İkinci paradigma araştırıcının veri seti için çeşitli olasılık dağılımlarını sağladığını kabul edip yöntemleri ve sonuçları belirlemesidir. Üçüncü paradigma olan istatistiksel öğrenme teorisinde, olasılık dağılımının ne olduğuna bakılmaksızın yeni vakalara uygulanmak üzere çıkarımların neler olacağı belirlenmeye çalışılır (11). Yani istatistiksel öğrenme teorisinin amacı, dağılımdan bağımsız yöntemler vasıtasıyla sınıflama ve tahminleme için hata sınırları belirlemek ve küçük örneklemlerdeki istatistiksel ilişkileri araştırmaktır (12). Bilindiği üzere klasik istatistikte doğru modelin bilindiği kabul edilip model parametreleri bulunur. Ancak istatistiksel öğrenme teorisinde model formu bilinmez, doğru olduğu düşünülen modeller arasından optimum sonucu veren modelin bulunması amaçlanmaktadır (13).

3.7. Verilerin Düzenlenmesi

Veri tabanları, büyük boyutlara sahip olduğundan ve heterojen kaynaklardan elde edildiğinden gürültülü, eksik ve tutarsız verilere karşı oldukça savunmasızdır. Verilerin düzenlenmemesi ya da kalitesiz olması yanlış sonuçların alınmasına yol açacaktır. Veri ön işleme de denilen bu süreç verinin temizlenmesi, birleştirilmesi, ayıklanması (azaltılması) ve dönüştürülmesidir (2).

Verinin temizlenmesi (cleaning); bu işlem gürültülü veriyi temizlemek ve verideki tutarsızlıkları gidermek amacıyla uygulanmaktadır. Gürültülü veri tam manasıyla anlamsız veri demektir. Örneğin sürekli verilerden oluşan bir veri setinde gözlemlerden birinde text verisinin olmasıdır (http://searchbusinessanalytics.techtarget.com/definition/noisy-data, Erişim tarihi: 03.09.2018).

Verinin birleştirilmesi (integration); bu işlem farklı kaynaklardan elde edilen bilgileri tıpkı bir veri ambarı gibi tutarlı bir veri seti haline getirmektedir.

Verinin ayıklanması (reduction); bu işlem toplama, gereksiz özellikleri ortadan kaldırma veya kümeleme yoluyla veri boyutunu azaltmaktadır.

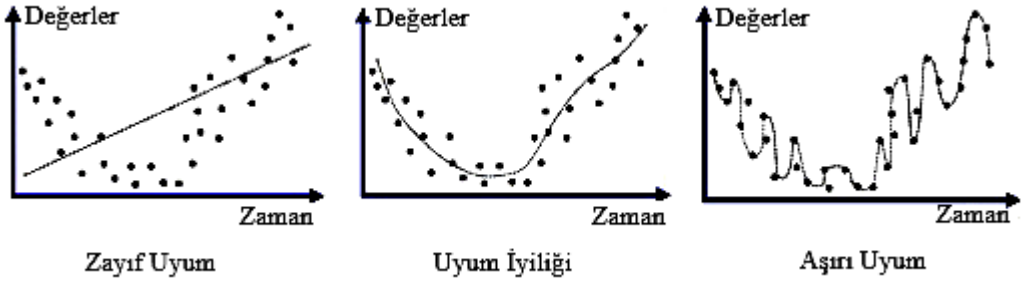
Verinin dönüştürülmesi (transformation); verilerin 0,0 ila 1,0 gibi küçük bir aralığa düşecek şekilde ölçeklendirilmesi için dönüşümler uygulanmaktadır. Bu işlem uzaklık ölçümlerini içeren madencilik algoritmalarının doğruluğunu ve verimliliğini arttırmaktadır. Örneğin bir tarih alanında, tüm girdileri ortak bir formata dönüştürmek gibi yanlış verileri düzeltmek için dönüşümler içerebilir.

3.8. Destek Vektör Makineleri

Sınıflandırma amaçlı kullanılan Destek Vektör Makinelerinin (DVM) temel işlevi bir hiperdüzlem yardımıyla veri setini sınıflara ayırmaktır. DVM algoritmasının doğrusal olarak ayrılabilen veriler için yapması gereken, sonsuz sayıda doğru içerisinden verileri birbirinden ayırabilecek aralığı (marjini) en geniş doğruları seçmektir. Doğrusal olmayan verilerde DVM algoritmasından beklenen, orijinal eğitim setini daha verimli işleyebilmek için verileri daha yüksek boyuta taşımak ve bu işlem için doğrusal olmayan haritalamayı kullanmaktır. DVM bu yeni boyutta verileri optimum ayırabilecek hiperdüzlemi (bir sınıfı diğerinde ayıran "karar sınırını) arar. DVM söz konusu hiperdüzlemi, destek vektörleri ve destek vektörleri tarafından tanımlanan marjinleri kullanarak bulur (2).

Destek vektör makineleri üzerine ilk makale Vladimir Vapnik ve ark. (1992) tarafından yapılmasına rağmen temeli 1960'lı yıllara dayanmaktadır. En hızlı DVM 'lerin bile eğitim süresi çok yavaş olmasına rağmen, karmaşık doğrusal olmayan karar sınırlarını modelleme yetenekleri sayesinde oldukça yüksek doğruluk oranları verebilmektedirler. DVM 'ler diğer yöntemlere göre aşırı uyuma (Şekil 2) daha az eğilimlidirler.

DVM 'ler sayısal tahminlemede kullanılmanın yanı sıra sınıflama için de kullanılabilmektedir. DVM 'ler görüntü ve metin sınıflandırma, nesne tanıma, el yazısı tanıma, ses tanıma ve yüz tanıma gibi çeşitli örüntü tanıma uygulamalarında sıkça kullanılmaktadır. (14). Sağlık bilimlerinde DVM özellikle, kanser morfolojisinde, tedavi başarısının ve ilgili genin belirlenmesinde, bazı hastalıkların teşhisinde kullanılmaktadır (15,16).



Şekil 2. Modellemenin veriye uyum durumu*

3.8.1. Karar Fonksiyonları

Genel olarak verilerin sınıflandırılmasında iki yaklaşımdan söz edilmektedir. Bunlardan ilki verilerin dağılımı hakkında ön bilginin olduğu parametrik yaklaşımlar ikincisi dağılım hakkında bilginin olmadığı parametrik olmayan yaklaşımlardır. Sinir ağları, fuzzy sistemler ve DVM parametrik olmayan sınıflayıcılara örnek olarak verilebilir. Bu algoritmalar, eğitim setindeki girdi (tahminleyiciler) ve çıktıları (yanıtlar) kullanarak yeni bir gözlemi belirlenen sınıflardan birine atamak için karar fonksiyonu oluştururlar (17).

3.8.1.1. İkili sınıflama için karar fonksiyonu

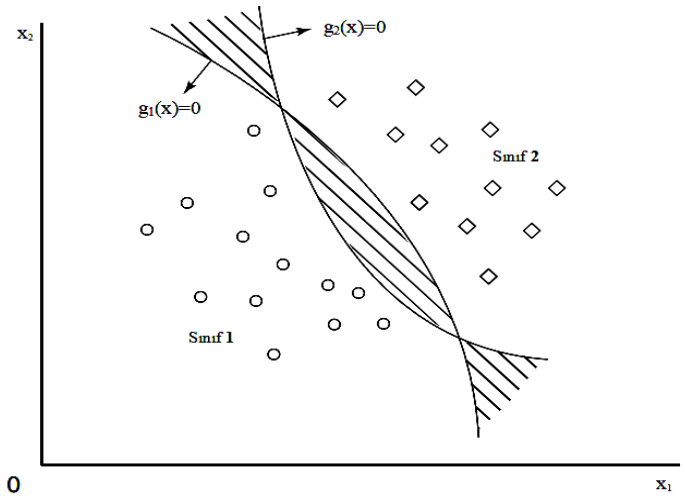
İki sınıflı sınıflayıcı için m boyutlu $\mathrm{x} = (x_1, x_2, \ldots, x_m)^{\mathrm{T}}$ vektörü ele alınsın. Sırasıyla birinci ve ikinci sınıf için $g_1(\mathrm{x})$ ve $g_2(\mathrm{x})$ skaler fonksiyonu olduğu varsayılsın, eğer;

$$g_1(\mathrm{x}) > 0 \quad \text{ve} \quad g_2(\mathrm{x}) < 0 \text{ ise x birinci sınıfa,} \tag{2.1}$$

$$g_1(\mathrm{x}) < 0 \quad \text{ve} \quad g_2(\mathrm{x}) > 0 \text{ ise x ikinci sınıfa atanmaktadır.} \tag{2.2}$$

Bahsi geçen fonksiyonlar karar fonksiyonu adını almaktadır (17).

* Overfitting; öğrenme algoritması esas problemi çözmekten uzaklaşır ve kendisine verilen değerleri tutturmaya odaklanır. Bu işleme o kadar odaklanır ki değerleri mükemmel tutturur ancak bu yüzden temel çözmesi gereken problemi çözmekten uzaklaşır. Bu durumda sınıflayıcının yeni veriyi genelleme yeteneği azalır ve algoritmanın sınıflama performansı düşer.



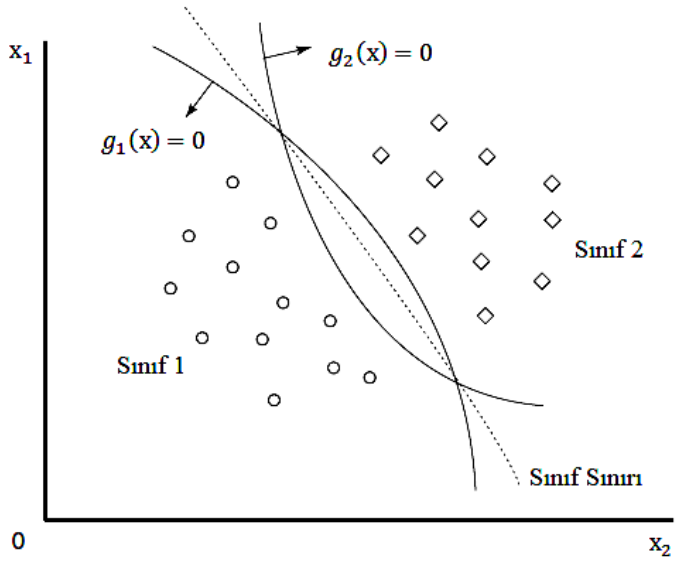
Şekil 3a. İki boyutlu uzayda karar fonksiyonu

Eğer karar fonksiyonları $g_1(x)g_2(x) > 0$ durumunu alırsa x sınıflanamaz (taralı bölgedeki değerler). Çünkü bu iki fonksiyonun çarpımının sıfırdan büyük olduğu yerlerde noktalar ya iki sınıfa birden atanmakta ya da iki sınıfa da atanamamaktadır. Dolayısıyla x 'i sınıflayabilmek için (2.1) ve (2.2) deki eşitlikleri aşağıdaki gibi düzenleyerek sınıflama işlemini gerçekleştirebiliriz.

$$g_1(x) > g_2(x) \quad \text{ise x birinci sınıfa} \tag{2.3}$$

$$g_1(x) < g_2(x) \quad \text{ise x ikinci sınıfa atanır.} \tag{2.4}$$

Bu durumda $g_1(x) = g_2(x)$ denklemi x 'e bağlı olarak çözüldüğünde sınıf sınırları dolaylı olarak elde edilir. Bu tip karar fonksiyonlarına dolaylı karar fonksiyonları adı verilir (17).



Şekil 3b. Dolaylı sınıf sınırı

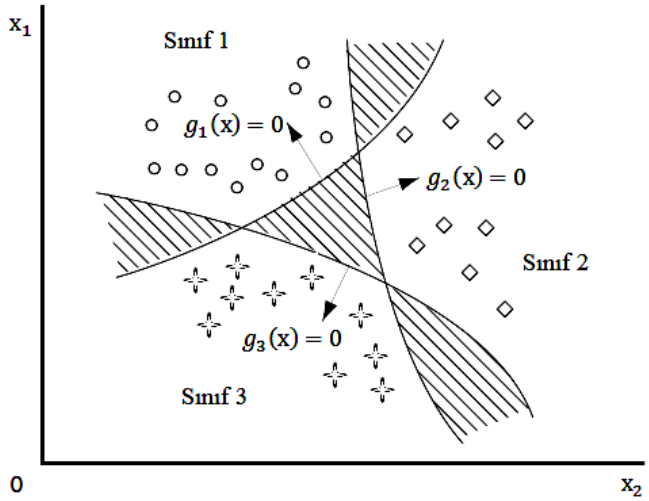
Eğer karar fonksiyonu doğrusal ise $g_1(x)$ fonksiyonu;

$$g_1(x) = w^{\mathrm{T}} + b \tag{2.5}$$

şeklinde ifade edilir. Burada w: m boyutlı vektör ve b yanlılık değeridir. Eğer bir sınıf, hiperdüzlemin pozitif kısmında yani $g_1(x) > 0$ ve diğer sınıf da negatif kısmında ise bu durum doğrusal olarak ayrılabilen problemler diye adlandırılır.

3.8.1.2. Çoklu sınıflama için karar fonksiyonu

Bu kısımda çoklu sınıflama için kullanılan birçok karar fonksiyonu belirleme yöntemi içerisinden bütüne-karşı-bir formülasyonu ele alınacaktır. x vektörünün *i.* sınıfa ait olduğu bilindiğinde *i.* karar fonksiyonu $g_i(x) > 0$'dır. $(i = 1,2, \dots, n)$ x 'in ait olduğu sınıf hariç diğer tüm sınıflarda $g_i(x) < 0$'dır. Ancak bazı durumlarda birden fazla pozitif değer alabilen karar fonksiyonu olabilmektedir (14). Böyle durumlar için geliştirilmiş yöntemler var olmakla birlikte tezde ikili sınıflama kullanılacağından bu kısma değinilmeyecektir.



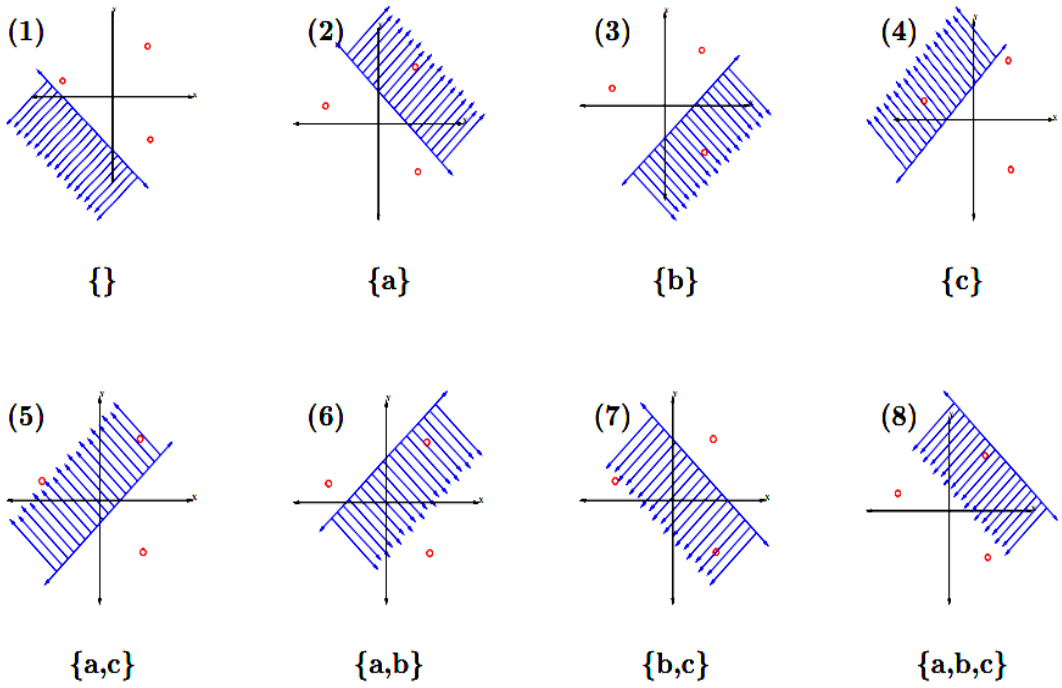
Şekil 3c. Bütüne-karşı-bir formülasyonuyla sınıf sınırları

3.8.3. Vapnik-Chervonenkis (vc) boyutu

VC boyutu bir dizi fonksiyonun kapasitesini ölçen skaler bir değerdir. İki sınıflı öğrenme probleminde {f (α)} fonksiyon seti için (α bir parametre seti olmak üzere), *l* adet nokta dizisi bütün olası 2^l yolla bu iki sınıfa ayrılabildiğinden, noktalar bahsi geçen fonksiyon seti tarafından parçalanabilmektedir (18). Daha açık bir ifadeyle, fonksiyonlar kümesi için bir VC boyutu, doğrusal bir şekilde ayrılabilen eğitim verisi noktalarının maksimum sayısı olarak tanımlanmaktadır. VC boyutu bir model kümesinin karmaşıklığını ölçtüğü için, "eğer iki model veriyi aynı oranda açıklıyorsa, bu durumda daha basit olan model tercih edilmelidir" ilkesini benimsemektedir. Şekil 4a 'da iki boyutlu uzayda (R^2) VC boyutunun a, b ve c gözlemleri için nasıl çalıştığı gösterilmiştir (2^3=8).

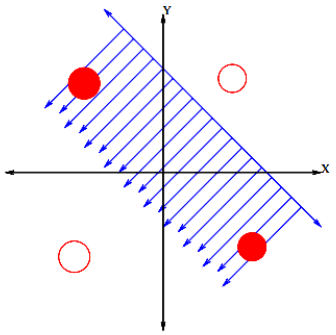
Tablo 1: R^2 'de gözlem dağılımı

Parçalanma Sıra	**Sınıflar 0**	**1**
1	{}	abc
2	a	bc
3	b	ac
4	c	ab
5	ac	b
6	ab	c
7	bc	a
8	abc	{}



Şekil 4a. R^2 de VC boyutu

Şekil 4a 'da, R^2 uzayında a, b ve c noktalarını ayıran 8 olası durum belirtilmektedir. Taralı kısımlar 0 sınıfına ait örnekleri, diğer kısımlar ise 1 sınıfına ait örnekleri göstermektedir. Belirlenen üç adet gözlemin yanına bir gözlem daha eklendiğinde, yani dört gözlem için, doğrusal ayrılma mümkün olamamaktadır. Dolayısıyla R^2 'de olası maksimum boyut sayısı üçtür.



Şekil 4b. R^2 de VC boyutunun dört nokta için doğrusal olarak mümkün olamayacağı

3.8.4. Yapısal risk minimizasyonu

Klasik yöntemler deneysel risk minimizasyonuna dayanırken DVM yapısal risk minimizasyonuna (YRM) dayanmaktadır ve yöntemin farklılığı da buradan kaynaklanmaktadır (19).

Deneysel risk minimizasyonu (DRM) sadece eğitim verisindeki hatayı minimize etmeye çalışmaktadır. Bu durumda eğitim verisinin, örneklemin çekildiği dağılımı temsil etmediği düşünülürse, öğrenme algoritmasının sınıflama yeteneği azalır ve performansı düşük olur. Ancak sadece eğitim setindeki hatayı değerlendiren DRM 'nin yerine YRM 'nin kullanılması durumunda, verideki aykırılıklar değerlendirilebilmekte ve oluşabilecek riskin çerçevesi belirlenebilmektedir (20).

R(f) gerçek risk ve R_{den}(f) de deneysel risk olarak kabul edilsin ve öncelikle deneysel risk incelenecek olsun. Deneysel risk için; *g(x)* fonksiyonu ikili sınıflama problemi Ω=(w_1, w_2) için ayırıcı ve $f: X \rightarrow \{1, -1\}$ karar fonksiyonu olsun. Bu durumda fonksiyon,

$$f = sign(g(x))$$

olarak tanımlanır. $x_1, x_2, \dots, x_n$ dizisi $z_1, z_2, \dots, z_n$ etiketleriyle ilişkili n boyutlu örneklemin eğitim noktaları olsun. Bu durumda x_i değeri w_1 ile etiketlenirse $z_i = 1$ ve x_i değeri w_2 ile etiketlenirse $z_i = -1$ olur. Dolayısıyla (sıfır-bir değeri alan) kayıp fonksiyonu;

$$\frac{1}{2}|f(x) - z| \qquad (2.6)$$

herhangi bir *x* eğitim örneğinin sınıflama doğruluğunu tanımlamaktadır. Yani eğer *x* doğru sınıflandıysa kayıp sıfır, yanlış sınıflandıysa kayıp bir olmaktadır. Buradan da ortalama eğitim hatası veya deneysel risk;

$$R_{den}[f] = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{2}|f(x_i) - z_i| \qquad (2.7)$$

olur. Deneysel risk ile birlikte aşağıdaki eşitlikten,

$$R\ (f) \leq R_{den}(f) + \phi(h, m, \delta) \qquad (2.8)$$

gerçek risk için üst sınır bulunur. Burada $\phi(h, m, \delta)$ terimi, güvenirlik ya da kapasite terimi olarak adlandırılmaktadır. Eşitlik daha açık bir ifadeyle;

$$\phi(h, m, \delta) = \sqrt{\frac{1}{m}\left(h\left(ln\frac{2m}{h} + 1\right) + ln\frac{4}{\delta}\right)} \tag{2.9}$$

$$R\,(f) = R_{den}(f) + \sqrt{\frac{1}{m}\left(h\left(ln\frac{2m}{h} + 1\right) + ln\frac{4}{\delta}\right)} \tag{2.10}$$

şeklinde ifade edilir. Burada *h,* VC boyutunu gösteren pozitif bir çarpan olarak ifade edilmektedir. Böylece birçok ML için, yeterli küçüklükte ve sabit bir δ belirlenerek eşitliğin sağ tarafını en küçük yapan öğrenme algoritması seçilir ve gerçek risk üzerindeki en düşük sınırlar bulunmuş olur (19). Eğitim verilerinin miktarına uygun fonksiyon sınıfını seçerek, denklem (2.10) 'daki hem $R_{den}(f)$ hem de $\phi(h, m, \delta)$ ifadeleri minimize edilerek yapısal risk minimizasyonunu gerçekleştirilmiş olur.

3.8.5. İki sınıflı destek vektör makineleri

İkili sınıflama problemlerinde, destek vektör makinesi eğitilmekte ve karar fonksiyonunun genelleme becerisi en üst düzeye çıkarılmaktadır. Yani *m* boyutlu giriş vektörü olan x, *l* boyutlu nitelik uzayı olan z ye haritalanır (17).

Bu bölümde iki sınıflı DVM için eğitim verilerinin doğrusal olarak ayrılabildiği sert marjin (hard-margin) DVM ve verilerin doğrusal olarak ayrılamadığı durum için geliştirilmiş yumuşak marjin (soft-margin) DVM ele alınacaktır. Daha sonra doğrusal olmayan DVM 'ler için çekirdek düzenlemesinden bahsedilecektir.

3.8.5.1 Sert marjin (hard margin) dvm

Hangi sınıftan olduğu bilinmeyen (sınıf 1 veya sınıf 2 'ye ait olabilir) m boyutlu girdi vektörü x_i i=(1, 2, 3, …, M), $y_i = +1$ ve $y_i = -1$ gibi iki sınıftan birine atanmak istensin, buradaki doğrusal ayrım (2.11) 'deki karar fonksiyonları yardımıyla belirlenebilir.

$$D(x) = w^{\mathrm{T}}x + b \tag{2.11}$$

burada w; m boyutlu vektör ve b yanlılık terimidir (*i*=1, 2, …, *M*)

$$D(x) = w^{\mathrm{T}}x + b \begin{cases} > 0 & \mathrm{y}_i = +1 \\ < 0 & \mathrm{y}_i = -1 \end{cases} \qquad (2.12)$$

Eğitim seti doğrusal olarak ayrılabildiğinden, eğitim setinin hiçbir gözlemi $w^{\mathrm{T}}x + b = 0$ denklemi ile ifade edilememektedir. Dolayısıyla ayrılabilirliği kontrol etmek için aşağıdaki eşitsizlik yardımıyla söz konusu problem çözüme ulaştırılabilmektedir.

$$w^{\mathrm{T}}x + b \begin{cases} \geq 1 & \mathrm{y}_i = +1 \\ \leq -1 & \mathrm{y}_i = -1 \end{cases} \qquad (2.13)$$

Buradaki +1 ve -1'i, eşitsizliğin sağ tarafı *a* ve *–a* gibi iki sabit olarak düşünebilir ve buradan (2.13) denklemini eşitsizliğin her iki tarafını *a* ya bölerek elde ederiz. (2.13) eşitliğini kullanarak aşağıdaki sonuç elde edilir.

$$\mathrm{y}_i(w^{\mathrm{T}}x_i + b) \geq 1 \qquad i=1,2, \ldots, M \qquad (2.14)$$

ve

$$D(x) = w^{\mathrm{T}}x + b = c \qquad -1 < c < +1 \qquad (2.15)$$

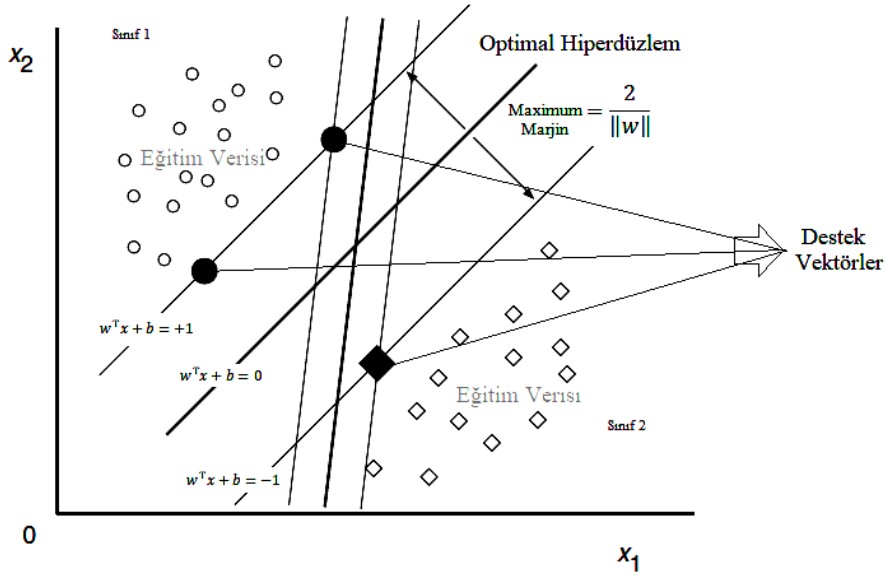
böylece eşitlik (2.15) ile x_i (*i*=1,2, …, *M*) için ayırıcı formda bir hiperdüzlem oluşmuş olur. Ayrıca $c = 0$ durumu, $c = \pm 1$ ile belirtilen iki hiperdüzlemin orta noktasını ifade etmektedir. Burada $D(x)$ 'in +1 ve -1 olduğu durumlarda hiperdüzlemlerden en az bir tanesi eğitim verisinin bir gözlemini ihtiva etmektedir.

Ayırıcı hiperdüzlem ile hiperdüzleme en yakın eğitim verisindeki nokta arasındaki uzaklık marjin olarak adlandırılmakta ve ayırıcı hiperdüzleme uzaklığı $\frac{2}{\|w\|}$ kadar olmaktadır. Şekil 5 'te görüldüğü üzere, (2.14) denklemi yardımıyla verileri birbirinden ayıran çok sayıda hiperdüzlem çizilebilmektedir. Ancak bu düzlemler içerisinden marjini en geniş olan hiperdüzlem kabul edilmekte ve bu en geniş marjinli hiperdüzleme *optimum ayırıcı hiperdüzlem* denilmektedir. (17).

Optimum ayırıcı hiperdüzlem belirlenirken öklid uzaklığı kullanılmaktadır. Bu uzaklık ($|D(x)|/\|w\|$) eğitim verisinin örneklerinden olan x 'in ayırıcı hiperdüzleme olan uzaklığı olarak ifade edilmektedir. Optimum ayırıcı hiperdüzlem, w ve b 'yi en küçük yapacak değerleri bulma problemini çözerek elde edilmektedir.

$$minimize\ Q(w,b) = \frac{1}{2}\|w\|^2 \qquad (2.16)$$

$$kısıt\ \ y_i(w^{\mathrm{T}}x_i + b) \geq 1 \quad i=1,2,\ldots,M$$



Şekil 5. İki boyutlu uzayda maksimum ayırıcı hiperdüzlem

kısıtı kaldırıp denklemi çözebilmek için Lagrange çarpanları ($\alpha = (\alpha_1, \alpha_2, \ldots, \alpha_M)^T$) eklenmektedir.

$$Q(w,b,\alpha) = \frac{1}{2}w^T w - \sum_{i=1}^{M} \alpha_i\{y_i(w^T x_i + b) - 1\} \qquad (2.17)$$

Eşitlik (2.17) 'nin w 'ya göre minimize ve α'ya göre maksimize edilmesi gerektiğinden çözüm karmaşıklaşmaktadır. Dolayısıyla denklemin çözümü için KKT koşulları yardımıyla çözüme gitmek gerekmektedir.

$$\frac{\partial Q(w,b,\alpha)}{\partial w} = 0 \qquad (2.18)$$

$$\frac{\partial Q(w,b,\alpha)}{\partial b} = 0 \tag{2.19}$$

$$\alpha_i\{y_i(w^T x_i + b) - 1 = 0 \tag{2.20}$$

$$\alpha_i \geq 0$$

Eşitlik (2.17) kullanılarak (2.18) ve (2.19) eşitlikleri aşağıdaki gibi sırasıyla azaltılmış olur.

$$w = \sum_{i=1}^{M} \alpha_i y_i x_i \tag{2.21}$$

$$\sum_{i=1}^{M} \alpha_i y_i = 0 \tag{2.22}$$

Eşitlik (2.21) ve (2.22) ifadeler (2.17)'de yerine konursa;

$$Q(\alpha) = \sum_{i=1}^{M} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{M} \alpha_i \alpha_j y_i y_j x_i^T x_j$$

denklemi elde edilir. Formüle edilen destek vektör makinesi, sert marjin destek vektör makinesi olarak adlandırılmaktadır. Buradan karar fonksiyonu;

$$D(x) = \sum_{i \in S} \alpha_i y_i x_i^T x + b$$

olarak bulunur.

3.8.5.2. Yumuşak marjin (soft margin) dvm

Sert marjinli destek vektör makinelerinde, eğitim verilerinin doğrusal olarak ayrılabilir olduğu varsayılmaktadır. Verilerin doğrusal olarak ayrılamadığı durumlarda, sert marjin destek vektör makinesi uygun çözüm sunamamaktadır. Soft marjin yaklaşımı pozitif değerli gevşek (slack) ξ_i (≥ 0) değişkenini kullanmaktadır. Sert marjinle elde edilen eşitliklere (2.14) slack değişkenler eklenerek, uygun çözüm elde edilebilmektedir.

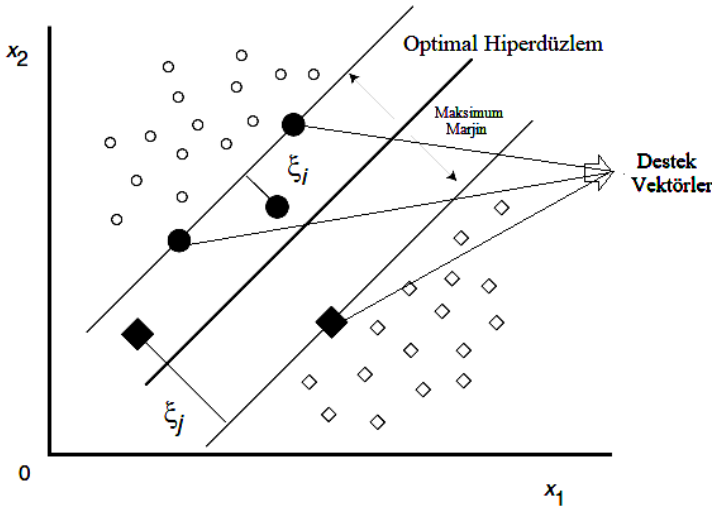
$$Y_i(w^T x + b) \geq 1 - \xi_i \qquad i=1, 2, \ldots, M \qquad (2.22)$$

Slack değişkenler ξ_i yardımıyla uygun çözümler her zaman elde edilebilir. Eğitim verisi x_i için eğer $0 < \xi_i < 1$ ise veriler maksimum marjine sahip olmadan da doğru sınıflanabilmektedir. Eğer $\xi_i \geq 1$ ise, veriler optimum hiperdüzlem ile yanlış sınıflandırılmaktadır. En geniş marja sahip olamayan optimum hiperdüzlemi elde etmek için aşağıdaki ifadeyi minimize etmek gerekir ki;

$$Q(w) = \sum_{i=1}^{M} \theta(\xi i) \quad \text{burada} \quad \theta(\xi i) = \begin{cases} 1 & \xi i > 0 \\ 0 & \xi i = 0 \end{cases}$$

bu çözümü zor kombinasyonal eşitliğin yerine aşağıdaki ifadenin minimize edilmesi daha kolaydır.

$$Q(w, b, \xi) = \frac{1}{2}\|w\|^2 + \frac{c}{p}\sum_{i=1}^{M} \xi_i^p \qquad (2.23)$$



Şekil 6. Slack değişkenler ile doğrusal ayırma

(2.23) eşitliği

$$y_i(w^T x + b) \geq 1 - \xi_i \qquad (2.24)$$

kısıtına maruz kalarak bulunmuş olur. Burada $\xi = (\xi_1, \xi_2, \dots \xi_M)^{\mathrm{T}}$ ve c minimum sınıflama hatası ve maksimum marjin arasındaki ödünleşimi (trade-off) sağlayan parametredir (21). Buradaki p değeri genellikle 1 ya da 2 seçilerek yumuşak marjinli hiperdüzlem elde edilebilmektedir. Eğer p=1 seçilirse bu destek vektör makinesine L1 yumuşak marjin DVM, p=2 seçildiği durumda ise destek vektör makinesine L2 yumuşak marjin DVM denilmektedir. Doğrusal olarak ayrılabilen durumda olduğu gibi pozitif Lagrange çarpanlarıyla;

$$Q(w, b, \xi, \alpha, \beta) = \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{M}\xi_i - \sum_{i}^{M}\alpha_i(y_i(w^T x_i + b) - 1 + \xi_i) - \sum_{i}^{M}\beta_i\xi_i \quad (2.25)$$

denklemi elde edilir. Burada $\alpha = (\alpha_1, \dots, \alpha_M)^T$ ve $\beta = (\beta_1, \dots, \beta_M)^T$'dir. Optimum çözüm için KKT koşullarıyla düzenlenen denklem;

$$\frac{\partial Q(w,b,\xi,\alpha,\beta)}{\partial w} = 0\ ,\ \frac{\partial Q(w,b,\xi,\alpha,\beta)}{\partial b} = 0 \text{ ve } \frac{\partial Q(w,b,\xi,\alpha,\beta)}{\partial \xi} = 0$$

$$\alpha_i(y_i(w^T x_i + b) - 1 + \xi_i) = 0 \text{ ve } \beta_i\xi_i = 0 \ (i=1, 2, \dots, M)$$

eşitliğine dönüşür. Buradan (2.25) denklemi kullanılarak w ve ξ 'ye bağlı türev denklemleri sırasıyla;

$$w = \sum_{i}^{M}\alpha_i y_i x_i \qquad \text{ve} \qquad \sum_{i}^{M}\alpha_i y_i = 0$$

eşitliklerine dönüşür ($\alpha_i + \beta_i = C$). Bu eşitlikler (2.25) 'de yerine konulursa;

$$maximize\ Q(\alpha) = \sum_{i=1}^{M}\alpha_i - \frac{1}{2}\sum_{i,j=1}^{M}\alpha_i\alpha_j y_i y_j x_i^T x_j$$

L1 yumuşak marjin destek vektör makinesi elde edilmiş olur. L1 yumuşak marjin destek vektör makineleri ile sabit marjin destek vektör makineleri arasındaki tek fark, α_i 'nin c 'yi geçememesidir (17).

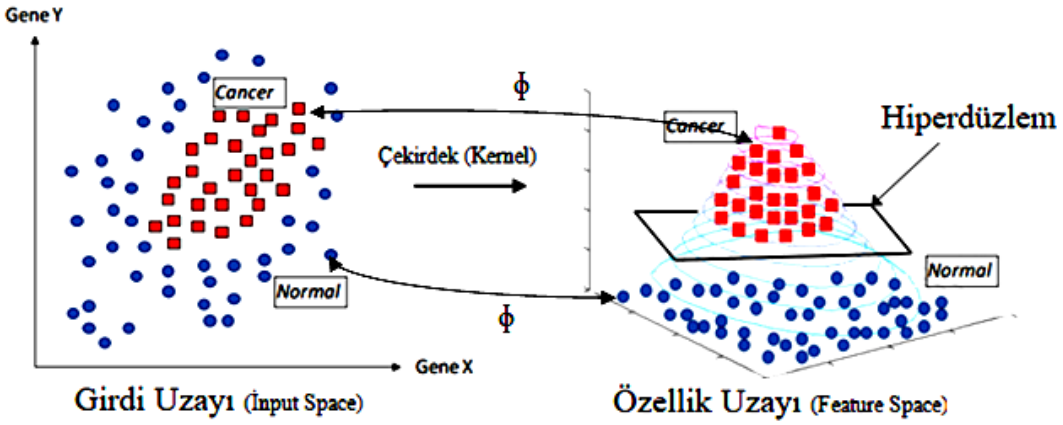
3.8.6. Doğrusal olmayan DVM, yüksek boyutlu uzaya haritama ve çekirdek düzenlemesi

Destek vektör makinelerinde üzerinde durulan en önemli konulardan biri genelleme kabiliyeti oluşturmak için hiperdüzlem belirlemektir. Ancak eğitim verisi doğrusal ayrılamıyorsa elde edilen hiperdüzlem optimal olsa bile sınıflandırıcının performansı düşebilmektedir. Bu sorunun önüne geçebilmek için orijinal girdi uzayı, özellik uzayı da (feature space) denilen bir nokta çarpım uzayına haritalanır ($x \in R^n \rightarrow \phi(x) \in R^f$).

Doğrusal olmayan ϕ(x) fonksiyonu $\phi(x) = (\phi_1(x), \ldots, \phi_l(x))^{\mathrm{T}}$ şeklinde tanımlansın. Bu fonksiyon *m*-boyutlu girdi vektörü x 'i *l* boyutlu özellik uzayına taşıyan doğrusal karar fonksiyondur ve aşağıdaki gibi tanımlanır.

$$D(x) = w^{\mathrm{T}}\phi(x) + b$$

burada *w* değeri *l* boyutlu vektör ve *b* yanlılık terimidir. Hilbert–Schmidt teoremine göre simetrik bir *K(x,x')* fonksiyonuyla;



Şekil 7. Doğrusal ayrılamayan verilerin yüksek boyutlu uzaya haritalanması

$$\sum_{i,j=1}^{M} h_i h_j K(x_i, x_j) \geq 0 \qquad (2.28)$$

eşitliğinden (burada M doğal sayı ve $h_{i,j}$ reel sayı olmak üzere) nokta çarpımla üretilen ϕ(x) fonksiyonu;

$$K(x, x') = \phi^{\mathrm{T}}(x)\phi(x')$$

eşitliğine dönüşür. Bu denklemle birlikte (2.28) eşitliği açılırsa;

$$\sum_{i,j=1}^{M} h_i h_j K(x_i, x_j) = \left(\sum_{i=1}^{M} h_i \phi^{\mathrm{T}}(x_i)\right)\left(\sum_{i=1}^{M} h_i \phi(x_i)\right) \geq 0 \qquad (2.29)$$

(2.28) ve (2.29) eşitliğinde bahsi geçen kısıtlar Mercer Koşulu olarak adlandırılır. Doğrusal olmayan DVM 'de çekirdek fonksiyonları söz konusu Mercer koşullarına uymak zorundadır (17). Bahsi geçen ϕ(x) haritalama fonksiyonunun kısıtlarının ve yapısının belli olmaması bununla birlikte yüksek boyutlu uzaylarda hesaplama karmaşası yaşanmasından dolayı çekirdek düzenlemesi yapılmaktadır.

3.8.6.1. Çekirdek fonksiyonlarla düzenleme

Doğrusal olmayan veriler doğrusal olarak ayrılabilen yüksek boyutlu bir uzaya dönüştürülürken, bu dönüşümler çekirdek fonksiyonlar aracılığıyla yapılmaktadır. Doğrusal DVM 'den doğrusal olmayan DVM 'ye geçişi sağlayan bu işleme çekirdek düzenlemesi (veya kernel trick) adı verilmektedir. Burada amaç düşük boyutlu girdi uzayında çözülmesi güç olan karmaşık sınıflama probleminin kolay çözümlenebileceği daha büyük boyutlu bir özellik uzayına taşınmasıdır. Çekirdek fonksiyonlardan bazıları;

Doğrusal çekirdek; sınıflama problemi girdi uzayında doğrusal olarak ayrılabiliyorsa girdi uzayını yüksek boyutlu bir uzaya haritalamaya gerek duyulmaz ve bu durumda kullanılacak olan fonksiyon;

$$K(x_i, x_j) = x_i^{\mathrm{T}} x_j \text{ 'dir.} \qquad (2.30)$$

Polinomiyal çekirdek; polinomiyal çekirdek d derece için verilecek olup d doğal sayı olmak üzere, bu fonksiyon aşağıdaki gibi ifade edilir.

$$K(x_i, x_j) = (x_i^{\mathrm{T}} x_j + 1)^d \tag{2.31}$$

Radyal tabanlı çekirdek; kullanımı en yaygın çekirdek fonksiyon olmak üzere, eşitlik aşağıda belirtildiği gibidir,

$$K(x_i, x_j) = \exp(-\gamma \|x_i - x_j\|^2) \tag{2.33}$$

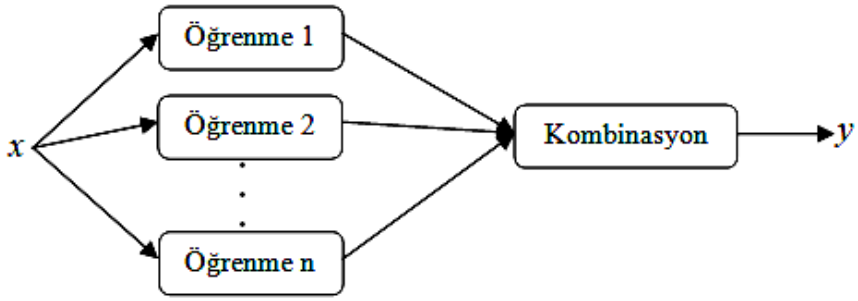
burada γ yarıçap kontrolünü sağlayan pozitif parametredir.

Sigmoid çekirdek; sadece belli δ değerleri için tanımlanan Sigmoid fonksiyon k ve δ gibi iki parametre içermektedir ve fonksiyon aşağıda olduğu gibi ifade edilir.

$$K(x_i, x_j) = \tanh(k x_i x_j - \delta) \tag{2.34}$$

3.9. Topluluk (Ensemble) Yöntemler

ML 'de topluluk metodu, aynı problemin daha yüksek performanslı çözümü için birden fazla modeli birleştirme olarak tanımlanmaktadır. Topluluk yönteminde amaç, bir ML yöntemini değişik şekillerde uygulayarak elde edilen modellerin doğrusal bir kombinasyonuyla yüksek doğruluk yakalamaktır. Modellerin bir araya gelerek oluşturdukları topluluğun genelleme yeteneği, yalnızca bir modelden elde edilen sonuçtan çok daha güçlüdür.



Şekil 8. Toplu öğrenmenin mimarı yapısı

Modeller bir araya getirilirken, modeller arasında farklılıkların olması topluluk metodunda istenen bir durumdur. Çünkü farklı sonuçların olmaması topluluk metodunun farkını ortaya çıkaramayacaktır (22). Bu kısımda temel topluluk yöntemleri olan Boosting ve RF algoritmalarından ve bu algoritmaların alt yapısını oluşturan Bagging yönteminden bahsedilecektir.

3.9.1. Bagging (Bootstrap Aggregration)

Ağaç tabanlı sınıflama yöntemlerinde kullanılan bu teknik, sınıflama kararlılığını arttırmak ve daha yüksek doğrulukla sonuç elde etmek için geliştirilmiştir. Varyans azaltıcı etkisi sayesinde sınıflamadaki hataları minimuma indirmektedir (23). Bagging yöntemini genel olarak örneklendirecek olursak; sizin bir hasta olduğunuzu ve semptomlara bakarak kendinize teşhis koymak istediğinizi varsayalım. Hastalığınızın ne olduğunu öğrenmek için kendinize yönelttiğiniz sorulardan size en yakın ve en çok cevap aldığınız durumlar, kendiniz için nihai ve en iyi teşhis olacaktır. Yani konulan teşhis eğer doktorlara sorulsaydı, her doktorun koyacağı teşhisin eşit değerde olduğu kabulüyle en çok oyu alan karar, final teşhis olarak değerlendirilecekti (2).

Bagging yönteminin algoritması şu şekilde çalışır: Örneklem hacmi N olan eğitim setinden bootstrap* örnekleme tekniğiyle m adet n örneklem hacimli küçük eğitim setleri oluşturulur ($n \leq N$). Oluşturulan m adet bootstrap örneklemi (B_1, B_2, …, B_m) ve her bir bootstrap örneklemi için m (ağaç sayısı) adet sınıflayıcı oluşturulur. Final sınıfı ise m adet sınıflayıcı içerisinden en çok oyu alan sınıf olarak belirlenir. Her bir örnekleme yerine koyma metodu ile örneklendiğinden bazı gözlemler, örneklemde birden fazla bulunabilmektedir (23).

3.9.2. Boosting

Boosting terimi zayıf öğrenmeyi güçlü öğrenmeye döndüren algoritmalar ailesi olarak ifade edilir. Zayıf (performansı düşük) öğrenmeler rasgele tahminden daha iyidir, güçlü (performansı yüksek) öğrenmelerden elde edilen bulgular ise hatasız sonuçlara oldukça yakın olmaktadır. Boosting prosedüründe, eğitim seti

* Bir bootstrap örneklemi, eğitim setinden yerine koyarak örnekleme yöntemiyle x adet örneğin seçilmesinden oluşur.

üzerinden tekrar tekrar bir dizi öğrenme gerçekleştirilir ve bu öğrenmelere ağırlıklar verilir. Elde edilen öğrenmelerin kombinasyonuyla doğru tahmin için minimum hatalı öğrenmeyi gerçekleştirir (22). Bagging 'de birden çok doktora sorup teşhis koymaya çalıştığımız örneği ele alırsak bu örnek Boosting için, her bir doktora yaptığı tahmin için ağırlıklar verilmesi ve ağırlıklandırılmış teşhislerin kombinasyonuyla en iyi sonucun alınması şeklini almaktadır (2).

3.9.3. Rasgele Orman

Forest (Orman) ifadesi, topluluk yöntemlerinde kullanılan sınıflayıcıların her birinin karar ağaçları olduğu durum için geliştirilmiş bir terimdir. Karar ağaçlarının her biri bölünmeyi belirleyen düğümlerdeki (nod) nitelikleri (attribute) *rasgele* seçerek oluşmaktadır. Başka bir ifadeyle karar ağaçlarının her biri ormandaki tüm ağaçlardan bağımsız olarak ve aynı dağılımla örneklenmiş bir rasgele vektörün değerlerine bağlıdır (2). RF yönteminde amaç, diğer topluluk yöntemlerinde olduğu gibi, birden fazla karar vericiyle daha verimli sonuçlar almaktır. RF yöntemi Bagging yöntemi üzerine inşa edilmiştir. Bagging yönteminden farkı, karar ağaçları dallara ayrılırken RF yönteminde bu ayrımı belirleyen (ileride anlatılacak olan m_{try} sayıdaki) değişkenler rasgele seçilir (24).

Karar ağaçları, tümevarımsal mantığın bilgisayar ortamına taşındığı gürültülü veriye karşı performanslı bir algoritmadır. Ağaçlar kök düğümünden, giriş verilerinin test edildiği iç düğümlerden, test sonuçlarını gösteren dallardan ve sınıf etiketlerinin bulunduğu yapraklardan oluşmaktadır. Örneğin bir kişinin bilgisayar alıp almamasına karar verme durumunu karar ağacıyla betimleyelim. Kullanılacak değişkenler kişinin yaşı, öğrenci olup olmaması ve kredi notu olsun. Bu niteliklerle PC alıp almama sınıflarına atama yapılırken Şekil 9'daki algoritmaya benzer bir yol izlenmektedir. Örneğimizde kök düğümü olarak belirlediğimiz yaş değişkeni üç farklı kategoriye ayrıldıktan sonra orta yaş grubu için direkt olarak PC alır sınıfına atanmıştır.

p boyutlu rasgele bir giriş vektörü $X=(X_1, \ldots, X_p)^T$ ve *Y* gerçek değerli bir yanıt değişkeni olsun. Amaç *Y* 'yi tahmin edebileceğimiz bir *f(X)* fonksiyonu bulmaktır. Bu tahmin fonksiyonu *L(Y, f(X))* şeklinde tanımlı kayıp fonksiyonu ile belirlenir ve kaybın beklenen değerini en aza indirmek için E_{XY} fonksiyonu, $E_{XY}(L(Y,$

f(X))) olarak tanımlanır. Sınıflamada eğer Y 'nin olası değerlerini $\hat{y}$ ile tanımlarsak, en aza indirilmiş $E_{XY}(L(Y, f(X)))$ ifadesi 0-1 kayıp fonksiyonu* için;

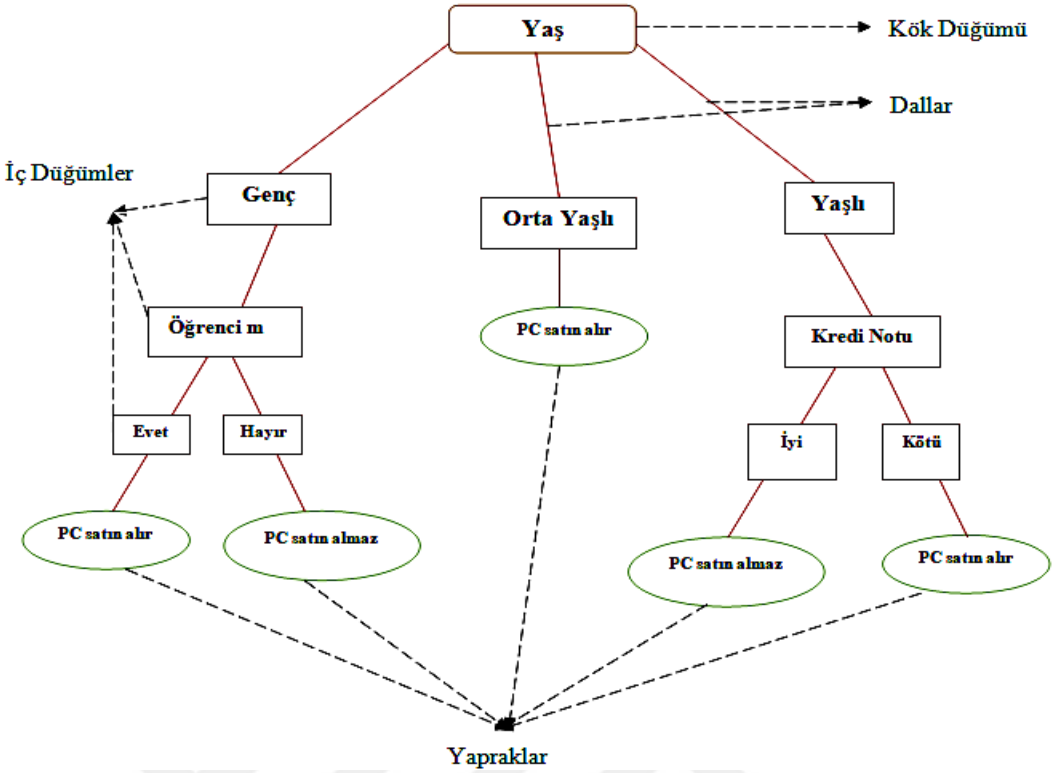
$$f(x) = \arg\underbrace{max}_{y\in\hat{y}} P(Y = y|X = x) \tag{3.1}$$

eşitliği elde edilir. Her bir karar ağacıyla yapılan öğrenmeyle, bunlar temel öğreniciler olarak ifade edilen $h_1(x), \ldots, h_J(x)$ fonksiyonlar olmak üzere, $f(x)$ tahminleyici fonksiyonu elde edilir. Sınıflama için bu fonksiyon (oylama ile) en çok tahmin edilen sınıf olarak aşağıdaki eşitlik ile elde edilir.

$$f(x) = \arg\underbrace{max}_{y\in\hat{y}} \sum_{j=1}^{J} I(y = h_j(x)) \tag{3.2}$$

RF metodu "sınıflandırma" olarak adlandırılan kategorik bir yanıt değişkeni ya da "regresyon" olarak adlandırılan sürekli bir yanıt için kullanılabilir. Tahminleyici (bağımsız) değişkenler ise kategorik veya sürekli olabilmektedir.

* *Kayıp Fonksiyon;* İstatistikte parametre tahmininde kullanılan kayıp ya da maliyet fonksiyonu, bir gözlem için gerçek değerlerle tahmin edilen değerler arasındaki farktır. Sınıflamada ise yanlış sınıflandırılmış bir gözlemin ceza skorunu temsil eder. 0-1 kayıp fonksiyonu, I indikatör fonksiyon olmak üzere; $L(\hat{y}, y) \neq I(\hat{y}, y)$'dir.



Şekil 9. Karar Ağacı Yapısı

RF 'nin istatistik alanındaki uygulamalarına örnek olarak; değişken önem ölçümleri, kayıp veri atama, aykırı değer tespiti, danışmansız öğrenme ve kademeli sınıf ağırlıklandırma verilebilir.

RF algoritması şu şekilde çalışır;

1) Bootstrap (yinelemeli örnekleme) ile n adet örneklem belirlenir. Her bir örneklemin % 66'sı (2/3) eğitim verisi (inbag) ve % 33'ü (1/3) test verisi (Out-of Bag, OOB) olarak ayrılır.

2) Eğitim seti içerisinden, m adet değişken (m<p, p:veri setindeki bütün değişkenler) rasgele seçilir. m adet değişken içerisinden maksimum bilgi kazancını sağlayan değişken belirlenir. Bu değişken belirlenirken; ID3 veya C4.5 algoritmasında kullanılan entropiye dayalı bilgi kazancına bakılır. Entropi, 0-1 aralığında değer alır. Veri setindeki tüm verilerin sadece bir sınıfa atanabildiği durumda entropi değeri

sıfır olur. Tüm veriler için sınıflara atanabilme olasılığı eşit olduğunda entropi maksimum değere ulaşır. Entropi belirsizliğin miktarını ölçtüğü için karar ağaçlarında, dallara ayırıcı değişkeni belirlemede kullanılmaktadır (25).

D: Bootstrap örneklemini ifade etmek üzere;

$$Info(D) = \sum_{i=1}^{n} -p_i log_2 p_i \qquad (3.3)$$

P=(p_1, p_2, …, p_n) olasılıklarını ifade etmek üzere; $p_i \geq 0$ (i=1,2,…,n) $\sum_{i=1}^{n} p_i = 1$'dir. Her bir değişken için bilgi kazancı hesaplanırken; örneğin A değişkeninden *D* örneklemi için elde edilen bilgi aşağıdaki gibi hesaplanır;

$$Info_A(D) = \sum_{i=1}^{n} \frac{|D_j|}{|D|} Info(D_j) \qquad (3.4)$$

A değişkeninden elde edilen kazanç ise

$Gain\ (A) = Info(D) - Info_A(D)$ ile hesaplanır.

Dallanma işlemi için, yani dallanma işleminin hangi değere göre yapılacağını belirlemede CART algoritmasında da kullanılan GINI indeksinden yararlanılır. Gini indeksi bir düğümdeki bootstrap örnekleminin karışıklık ya da safsızlık seviyesini ölçmektedir. İki sınıflı bir araştırma probleminde p; t. düğümde sınıflardan birindeki gözlemlerin oranını ve 1-p de diğer sınıftaki gözlemlerin oranını göstersin. Bu durumda t düğümü için Gini indeksi (https://www.researchgate.net/publication /226246437_Tree-Based_Methods, Erişim Tarihi 15 Ekim 2018);

$$G_t = 2p(1-p)\text{'dir} \qquad (3.5)$$

3) Kök düğümü için değişken ve dallara ayırma işlemleri için değişkenler belirlendikten sonra her yaprağa bir sınıf atanır. Bu işlemlere, yeni bir dal oluşturulmasına gerek kalmayana devam edilir.

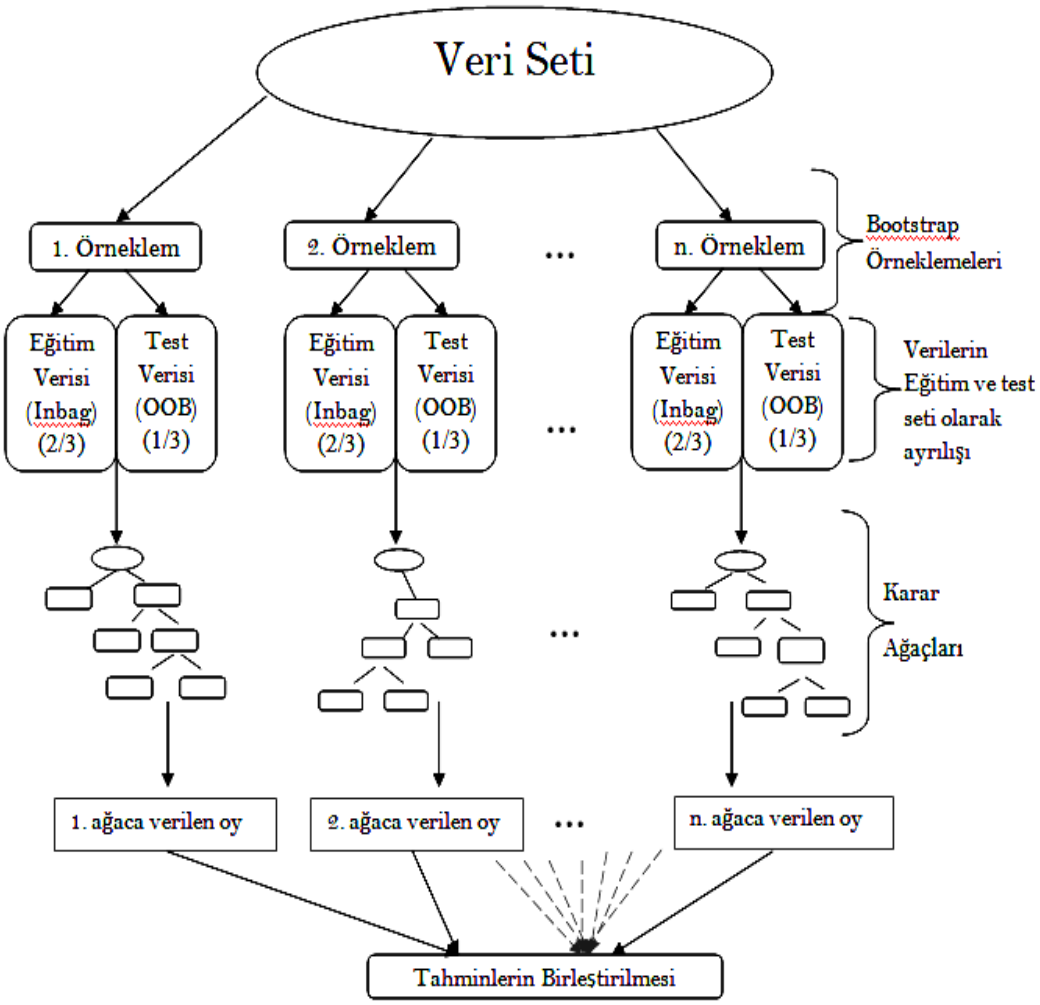
4) Eğitim veri setiyle oluşturulan karar ağaçlarına test veri seti uygulanarak OOB hata oranı hesaplanır. Böylece her bir karar ağacı OOB ile belirlenen ağırlığıyla ters orantılı olacak şekilde oylanır.

5) Bütün karar ağaçlarının yaptığı tahminlere göre en çok oyu alan sınıfın nihai sınıf olduğu kabul edilir (26).

OOB (Out of Bag) hesabı için test verisindeki (yani veri setinin % 33'ünü temsil eden küçük veri seti) her bir gözlem karar ağaçlarıyla elde edilen sınıflamaya tabi tutulur. Karar ağacının test verisindeki her bir gözlem için yaptığı sınıflamayla veri setindeki gerçek sınıflar karşılaştırılır. Elde edilen iki sonucun birbirine eşit olmama durumunun oranı OOB' yi verir (https://stat.ethz.ch/education/semesters /ss2012/ams/slides/v10.2.pdf, Erişim Tarihi: 15 Ekim 2018). Tüm karar ağaçlarından bulunan OOB hata oranlarının ortalaması, RF karar ormanının OOB hata miktarını vermektedir.

3.9.3.1. RF yönteminde hiperparametreler

RF yönteminde değişkenleri dallara ayıracak olan "m" adet değişken, veri setindeki bütün değişkenler (p) içerisinden rastgele seçilmektedir. Uygun "m_{try}" sayısının belirlenmesi öğrenme yönteminin performansı açısından oldukça önemlidir. "m_{try}" sayısının olması gerekenden küçük alınması durumunda karar ağaçlarında sınıflama için yeterli argüman toplanamayacak ve sınıflama performansı düşecektir. Diğer taraftan "m_{try}" sayısının gereğinden yüksek seçilmesi durumunda, ağaçların yaptıkları sınıflama benzerlik göstereceğinden sınıflar birbirinden tam anlamıyla ayrılamayacak ve sınıflama doğruluğu düşük çıkacaktır. Sınıflamada kullanılacak değişken sayısı Brieman tarafından $\sqrt{p} = m_{try}$ olarak önerilmiştir. Bu konuda başka değerlere (p/2, 0.1p vs.) imkan tanıyan çalışmalar olmakla birlikte uygun "m_{try}" sayısı belirlemede, farklı değerlerle sınıflama gerçekleştirilerek optimum performans gözetilmelidir (27). Bir diğer parametre kullanılacak ağaç sayısı "ntree" parametresidir. Ağaç sayısı azaltıldıkça sonuçlar önemli ölçüde bozulmaktadır.



Şekil 10. RF algoritması akış şeması

Hatta ntree=1 seçildiğinde RF yöntemi tamamen sadece bir karar ağacıyla sınıflama gerçekleştiren öğrenme yöntemi şeklini alır. Bununla birlikte "ntree" parametresinin gereğinden fazla olması durumu ekstra işlemlerin yapılmasına ve aşırı uyuma (over-fitting) neden olabilmektedir. Parametre olarak "ntree" için de Brieman tarafından önerilen 500 adet ağacın kullanılması gerektiğidir. Ancak "m_{try}" değerinde olduğu gibi burada da optimum sonuçlar için başka değerler ile sınıflama performansı ve OOB' nin artık sabit bir değer aldığı nokta incelenmelidir (27).

3.9.3.2. RF yönteminde yakınlık (proximity)

RF de yakınlık (proximity) benzerlik gösteren gözlemlerin daha çok aynı yaprak (terminal) düğümde, benzer olmayan gözlemlerin ise daha az aynı yaprak düğümde yer almasıdır (28). Yakınlık (proximity) gözlemlerin örüntüleri, aykırı ya da gürültülü veriler, gözlemlerin ilişkileri gibi konularda araştırıcıya bilgi vermektedir. Bu bilgiyi, iki gözlem arasındaki uzaklığı (proximity measure) hesaplayarak elde etmektedir. Uzaklık matrisi oluşturulurken veri setinin tamamı sınıflama ağacından geçirilir, herhangi iki gözlem eğer aynı yaprakta yer alırsa aralarındaki uzaklık bir arttırılır. Bu işlem ormandaki bütün ağaçlar için tekrarlandıktan sonra ortaya çıkan matrisin her bir hücresi, ormandaki ağaç sayısına bölünür. Böylece uzaklık oranları elde edilmiş olur. Uzaklık oranları arttıkça gözlemler birbirlerine benzer bir yapı gösterirken, uzaklık oranı azaldıkça gözlemlerin benzerliği azalmakta ve düşük oranlı gözlem aykırı değer (outlier) şüphesi taşımaktadır (24-26).

3.9.3.3. RF yönteminin avantajları ve dezavantajları

Avantajları;

- Büyük veri tabanlarında oldukça başarılı sonuçlar verir.
- Değişken silme işlemine gerek kalmadan binlerce giriş değişkeniyle işlem yapabilir.
- Sınıflamada değişkenlerin göreli önemini değerlendirme imkanı sağlar.
- Eksik veri atanmasında ve eksik verinin çok fazla olduğu veri setlerinde sınıflamada oldukça başarılıdır.
- Örneklem hacmi eşit olmayan sınıflamada performansı yüksektir.
- Aşırı uyum problemi yoktur ve ağaçların budanmasına gerek duymaz (https://www.coursehero.com/file/27020301/Random-Forestpptx/, Erişim Tarihi: 26.04.2018).

Dezavantajları;

- Veri setinde farklı seviyelere sahip kategorik değişkenler olduğunda, daha fazla kategorisi olan değişken için yanlı sonuçlar verebilmektedir.

- Gürültülü veriler* içeren veri setlerinde aşırı uyum problemi sergileyebilmektedir (29).
- RF 'yi oluşturan birçok ağaç görülemediğinden sonuçların güvenirliliği tartışılabilmektedir (30).

3.10. Yapay Sinir Ağları

Yapay sinir ağları (YSA) beynin nöral yapısına göre dizayn edilmiş, beyne göre daha az işlevsel sayılabilecek elektronik modellerdir. Beynin yapısından esinlenerek elde edilmiş bu yöntem, bilgisayar endüstrisinin geleceği için oldukça önemli bir gelişme olarak karşımıza çıkmaktadır.

Nöral ağlar, orijinal olarak nöronların hesaplama benzetimlerini geliştirmeye ve test etmeye çalışan psikologlar ve nörobiyologlar tarafından hazırlanmıştır. Genel ifade ile bir sinir ağı, her bir bağlantının onunla ilişkili bir ağırlığa sahip olduğu bir dizi bağlantılı giriş/çıkış birimidir. Kökeni 1900'lü yıllara dayanan ve beyni oluşturan sinir hücrelerinin belli bir örüntü oluşturduğu algısıyla başlayan YSA, günümüzde çok çeşitli amaçlarla kullanılmaya devam etmektedir (Kalite kontrol, sistematik modelleme, finansal tahminleme, güvenlik teçhizatları vb).

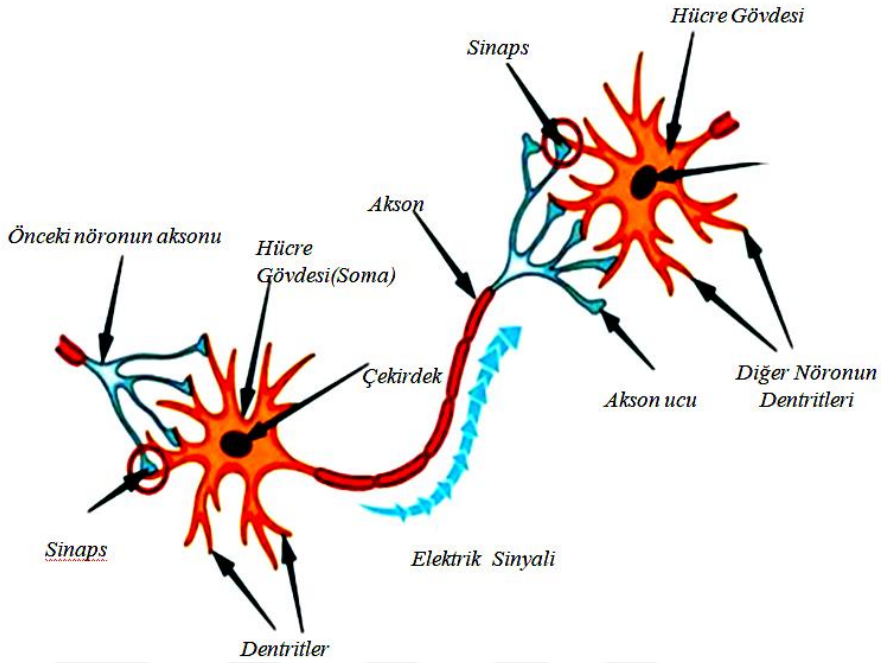
Genel bir tanım olarak YSA; öğrenilen bilginin saklanması ve analize uygun hale getirilmesi için, bilgiyi işleme yeteneğine sahip basit işlem birimlerinden oluşan büyük bir kısmı paralel dağılan işlemcidir (31). Paralel dağılma ile anlatılmak istenen, insan hafızasında olduğu gibi depolanan bilginin hemen her türünün, işlenmeye ihtiyaç olduğunda veya çağrıldığında edinilebilmesidir.

3.10.1. Biyolojik sinir hücreleri (nöronlar)

En basit ifade ile nöronlar bilgi giriş ve çıkışı olan anahtardan başka bir şey değildir. Bilgi girişine izin veren diğer nöronlar yeterli düzeyde uyarı gönderdiğinde, anahtar aktif hâle gelmektedir. Bu işlem devam ederek bilgi çıkışında diğer nöronlara uyarılar gönderilmektedir (32). İnsan beyninin korteks kısmında yer alan nöronlar on binlerce hücre ile etkileşim halindedir (33). Koloniler inşa eden ve aynı zamanda navigasyon harikası olan bal arılarında yaklaşık $0{,}8x10^6$ nörona ihtiyaç vardır. Bu

* *Gürültülü veri;* gürültü denilen büyük miktarda ek anlamsız bilgi içeren verilerdir. Terim genellikle bozuk veriler için eşanlamlı olarak kullanılmıştır.

rakam insan sinir sistemi için yaklaşık $2x10^{11}$ ve daha fazlası olabilmektedir. Nöronlar aynı zamanda biyolojik sinir ağlarının temelini oluşturmaktadır. (32).



Şekil 11. Nöron yapısı

3.10.1.1 Nöronların temel bileşenleri

Sinir hücresinin temel bileşenleri dentrit, soma ve aksondan oluşmaktadır. Nöronun dentritleri, bilgiyi özel bağlantılar (sinapslar) ile alır. Diğer nöronlardan veya hücrelerden gelen sinyaller, sinapslar ile bir sonraki nörona transfer edilmektedir. Söz konusu sinapslar dentritlerde veya doğrudan somada bulunabilir.

Dentrit; dentritler ağaçlar gibi dalları ile nöronun hücre çekirdeğinden veya birçok farklı kaynaktan gelen elektrik sinyallerini aldıktan sonra bunları hücrenin çekirdeğine aktarmaktadır.

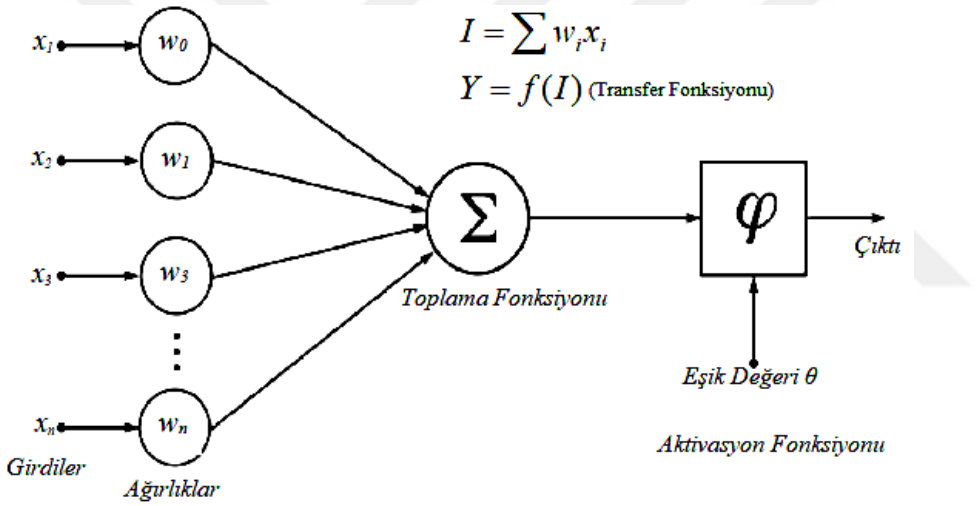
Soma; somalar, sinapslar ve dentritler ile aktive edici ve aktivasyon gerektirmeyen bütün sinyalleri aldıktan sonra bunları bünyesinde tutar. Birikmiş olan bu sinyaller belli bir değeri (eşik değeri veya treshold değeri) aşar aşmaz nöronun hücre çekirdeğini ve daha sonra mevcut olana bağlı olan nöronlara iletilecek olan uyarıyı (elektrik sinyalini) harekete geçirir.

Akson; Nöronlar uyarıyı (elektrik sinyalini) akson vasıtasıyla diğer nöronlara aktarılır. Akson, somanın daha uzun ve ince hali olarak düşünülebilir. Olağan dışı bir

durum geliştiğinde aksonlar bir metreye kadar uzayabilir (örneğin omurilikte). Aksonlar, elektrik sinyalini daha iyi iletebilmek için yalıtkan bir madde olan miyelin kılıfla sarılmıştır (32).

3.10.2. Yapay sinir hücreleri

Yapay sinir hücreleri, biyolojik sinir hücrelerine 4 hususta benzemektedir. Şekil 12'de görüldüğü üzere yapay sinir ağına gelen çeşitli girdiler *x(n)* ile ifade edilmektedir. Girdilerin her biri bir bağlantı ağırlığı ile çarpılır *w(n).* Daha sonra elde edilen bu ağırlıklandırılmış değerler belirlenen transfer (aktivasyon) fonksiyonuna gönderilir. Aktivasyon fonksiyonunda işlenen değerler çıktı kısmına iletilir. Bu işlem farklı toplama fonksiyonu ve farklı aktivasyon fonksiyon yapılarıyla da gerçekleştirilebilir (34).



Şekil 12: Temel Yapay Sinir Hücresi

Bir yapay nöron yedi ana bileşenden oluşmaktadır. Bu kısımda bu bileşenlerden bahsedilecektir.

3.10.2.1. Ağırlık faktörleri

Bir nörona eşzamanlı birçok bilgi girişi olmaktadır. Her girdinin kendi nispi ağırlığı vardır ve bu da her bir girdinin toplama fonksiyonunda ne kadar değer aldığını belirlemektedir. Toplama fonksiyonunda biriken girdiler ağırlıklarına göre önem sırasına konmaktadırlar (34).

3.10.2.2. Toplama fonksiyonu

Toplama fonksiyonu nörona gelen net girdiyi hesaplar (35). Girdiler *(x_1, x_2, x_3, ... x_n)* ve belirlenen ağırlıklarına *(w_1, w_2, w_3,... w_n)* göre toplama fonksiyonundaki değerleri bu iki vektörün nokta çarpımı cinsinden ifade edilmektedir *(x_1* w_1, x_3* w_2, x_3* w_3, ..., x_n* w_n)*. Toplama fonksiyonu girdiler ve onların ağırlıklarıyla oluşan nokta çarpım değerinden daha farklı değerlerle de ifade edilebilmektedir. Örneğin çarpım fonksiyonuyla $\prod_{i=1}^{n} x_i w_i$ veya girdilerin ve ağırlıkların çarpımından elde edilebilecek maksimum değerle (max (x_i,w_i)) hesaplanabilmektedir (34).

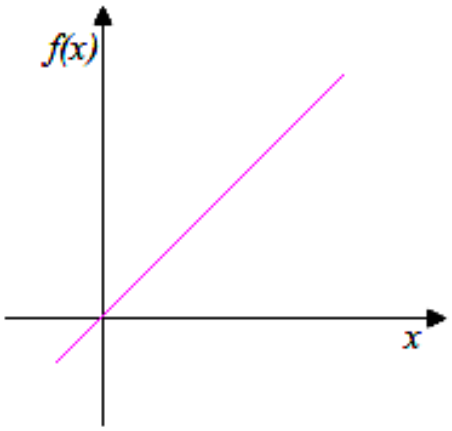
3.10.2.3. Transfer (aktivasyon) fonksiyonu

Transfer fonksiyonu olarak belirlenen bir algoritmik işlem yoluyla toplama fonksiyonundan elde edilen sonuçları analiz çıktısına dönüştüren birimdir. Transfer fonksiyonu belirlenen algoritma ile elde edilen sonuçları eşik (treshold) değeri ile kıyasladıktan sonra eğer bulgu eşik değerinden büyükse işleme elemanı bir sinyal üretir. Eğer bulgu eşik değerinden küçükse herhangi bir sinyal üretilmez.

Eşik değerinin üretildiği fonksiyon ve aktivasyon fonksiyonu genellikle doğrusal olmayan fonksiyondur (34). Yapay sinir ağlarında kullanılan birçok aktivasyon fonksiyonu olmakla birlikte bu fonksiyonlar türevlenebilir ve sürekli olmalıdır. Bu kısımda bir takım aktivasyon fonksiyonundan bahsedilecektir.

3.10.2.3.1. Doğrusal aktivasyon fonksiyonu (f(x)=x)

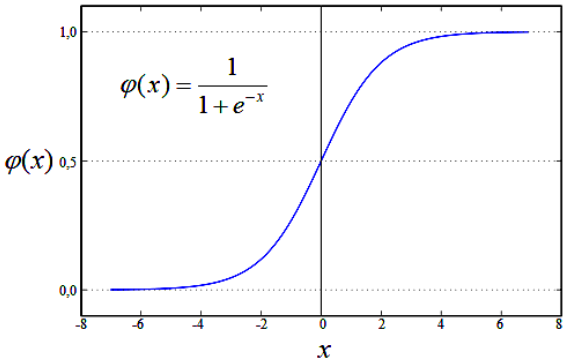
Doğrusal olmayan fonksiyonların nöronlar için daha elverişli olmasından dolayı daha az kullanılan doğrusal aktivasyon fonksiyonu, çıkış birimi için herhangi bir aralık sunamamaktadır. Ancak nöronun diğer aktivasyon fonksiyonlarıyla uyumunu ortaya koymak adına tercih edilebilmektedir.



Şekil 13. Doğrusal fonksiyon

3.10.2.3.2. Sigmoid (lojistik) fonksiyon

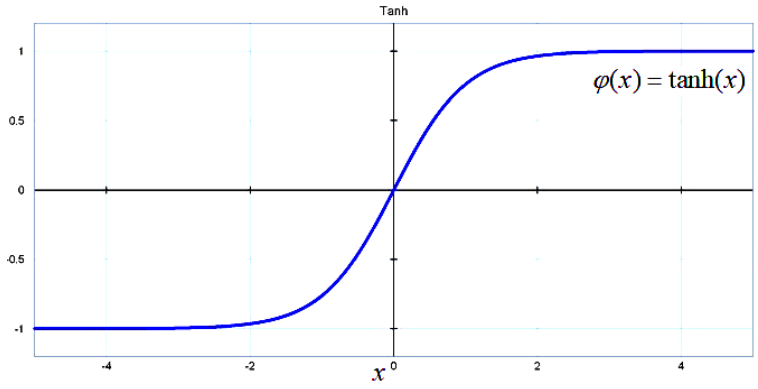
Bu fonksiyonun kullanılmasının en önemli gerekçesi, sigmoid fonksiyonun 0 ile 1 arasında değer almasıdır. Bu değer aralığı bize, işlem çıktısını bir olasılık tahmini gibi yorumlayabilme imkanı sağlamaktadır (32). Yapay sinir ağlarında en çok tercih edilen aktivasyon fonksiyonlarından biridir.



Şekil 14. Sigmoid fonksiyon

3.10.2.3.3.Hiperbolik tanjant fonksiyonu

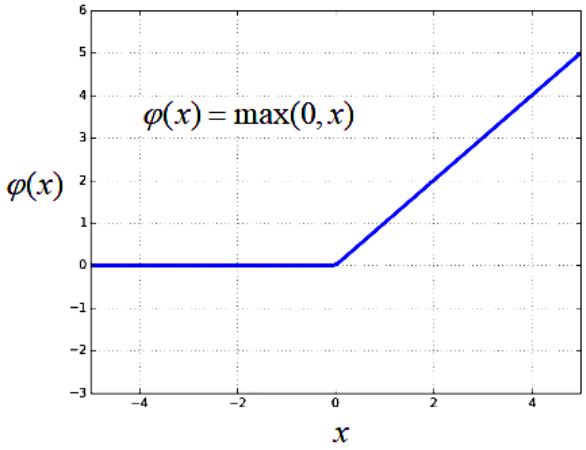
Hiperbolik tanjant fonksiyonu -1 ile +1 aralığında değer aldığından negatif girdiler, negatif değer olarak işlenmekte ve değeri olmayan girdiler sıfıra yakın hesaplanmaktadır. Bu fonksiyon genelde ikili sınıflamada tercih edilmektedir (https://towardsdatascience.com/activation-functions-neural-networks1cbd9f8d91d6, Erişim Tarihi: 08.02.2018).



Şekil 15. Hiperbolik tanjant fonksiyonu

3.10.2.3.4. ReLU (düzeltilmiş doğrusal birim) aktivasyon fonksiyonu

Popüler aktivasyon fonksiyonlarından biridir. 0-1 aralığında değer almaktadır. Dezavantajı, 0-1 aralığı için bütün negatif değerleri sıfır olarak işlemesi ve dolayısıyla modelin eğitilme yeteneğinin az oluşudur.



Şekil 16. ReLU fonksiyon

3.10.2.4. Ölçekleme ve sınırlandırma

Ölçekleme işleminde transfer fonksiyonundan elde edilen değer bir skaler ile çarpılır ve bir dengeleme değeri eklenir. Sınırlandırma işlemi ise ölçeklendirilmiş sonucun bir üst veya alt sınırı aşmadığını garanti eden mekanizmadır (34).

3.10.2.5. Çıktı fonksiyonu

Sonuçların alındığı birimdir. Her işleme elemanına yüzlerce başka nörona verebileceği bir çıkış sinyali verilir. Nöron kendi çıktısını yine kendisine girdi olarak da gönderebilmektedir (35).

3.10.2.6. Hata fonksiyonu ve geriye yayılım değeri

Öğrenmelerin çoğunda mevcut çıkış ile istenen çıkış arasındaki fark, hata fonksiyonu tarafından belirlenen bir ağ mimarisine uyacak şekilde dönüştürülen hata olarak hesaplanmaktadır. En temel sinir ağları bu hatayı olduğu gibi kullanırken ileri yöntemlerde çeşitli dönüştürme işlemleri yapılmaktadır. Elde edilen hata geriye (bir önceki) katmana yayılır. Bu geri yayılan değer ya hata olarak kabul görür ya da istenen ağ tipine bağlı olarak (genellikle aktarım işlevinin türevi alınarak yapılan) ölçeklendirilir (34).

3.10.2.7. Öğrenme fonksiyonu

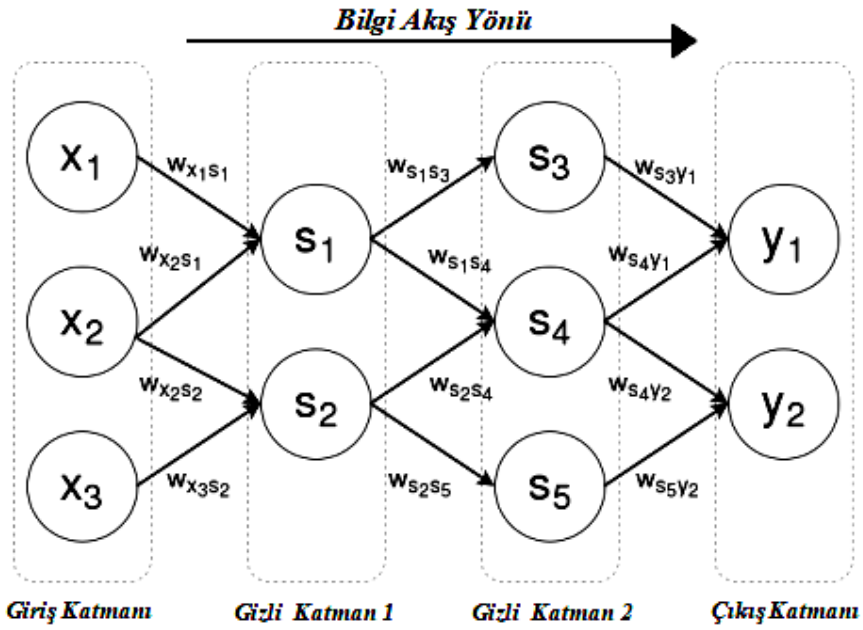
Bu fonksiyonun amacı bazı nöral tabanlı algortimalara göre her işleme elemanının ağırlığını değiştirmektir.

3.10.3. Yapay sinir ağlarının oluşumuna göre sınıflandırılması

YSA genel olarak üç başlıkta ele alınmaktadır. Birincisi nöronların bağlanış dizaynına göre ileri beslemeli ve geri beslemeli YSA mimarilerinden oluşmaktadır. İkincisi öğrenme algoritmasına göre yapılan sınıflandırmadır. Öğrenme algoritmaları, tezin ilk bölümünde bahsedildiği gibi danışmanlı öğrenme (YSA ’da kullanımı Hebb Kuralı, Hopfield Kuralı, Delta Kuralı, vb.), danışmansız öğrenme (YSA ’da kullanımı Kohonen kuralı ve Adaptif Rezonans Teorisi) ve destekleyici öğrenmeden (YSA ’da kullanımı Boltzman makineleri) oluşmaktadır. Üçüncü oluşum olan öğrenme zamanına göre YSA, Statik ve Dinamik öğrenmeden oluşmaktadır.

3.10.3.1. İleri beslemeli yapay sinir ağları

Bu mimaride veriler girdi (input) katmanından çıktı (output) katmanına doğru tek yönlü bağlantı ile iletilmektedir. Diğer bir ifade ile bir katmandaki her bir nöron, sadece bir sonraki tabakanın (çıkış katmanına doğru) nöronlarına yönelik bağlantılara yönlendirilmiştir (32). Veriler aracılığı ile bilgiler öğrenilirken herhangi bir katmanın (giriş, gizli veya çıkış) kendisinde bağlantılar kurulmamaktadır. İleri beslemeli ağ çok katmanlı YSA 'da kullanılabildiği gibi tek katmanlı (sadece girdi ve çıktı katmanından oluşan ağlar) YSA 'da da kullanılabilmektedir (Şekil 18). İleri beslemeli ağları, danışmanlı öğrenme algoritması ile birlikte kullanarak çok katmanlı yapay sinir ağı modelleri ve destekleyici öğrenme algoritması ile birlikte kullanarak da vektör kuantizasyon modelleri (LVQ) oluşturulmaktadır (35, 36).

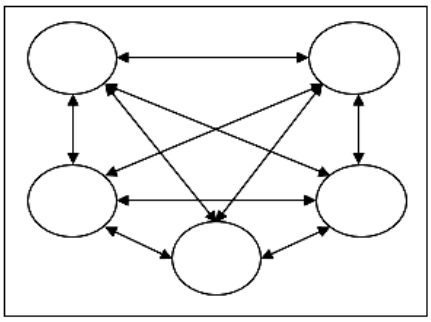


Şekil 17. İleri beslemeli ağ mimarisi

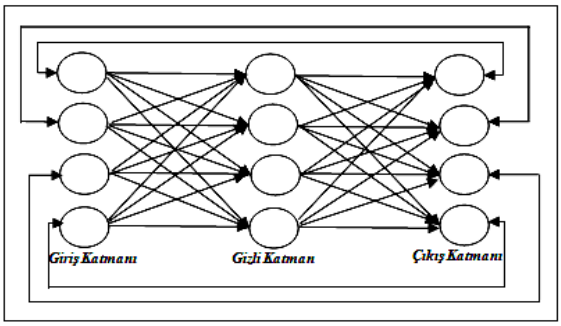
3.10.3.2. Geri beslemeli yapay sinir ağları

Geri beslemeli yapay sinir ağlarında bir hücreden alınan çıktı, kendinden önceki katmana veya kendi katmanında bulunan herhangi bir hücreye girdi olarak kabul edilebilmektedir. Geri beslemeli yapay sinir ağları katmanlar ve işlemciler arası iletişime olanak sağlamaktadır. Dolayısıyla geri beslemeli yapay sinir ağları doğrusal olmayan dinamik bir davranış göstermektedir (36). Geri beslemeli yapay

sinir ağları, kapalı döngülere sahip olan yinelemeli ağları da barındırmaktadır. Bu yinelemeli ağlar genel olarak iki kısımdan oluşmaktadır. Birincisi, bütün nodların birbiriyle iletişimine imkan veren ve bütün nodların giriş ve çıkış ünitesi gibi çalışabileceği "tamamen yinelemeli ağlar" adını almaktadır. İkincisi çıkışın (output) geri besleme olarak tekrar girişe (input) gideceği "Jordan" kapalı döngü ağlarıdır (https://www.tutorialspoint.com/artificial_neural_network/artificial_neural_network_building_blocks.htm, Erişim Tarihi: 29.06.2018).



Şekil 17.a. Tamamen yinelemeli ağlar



Şekil 17.b. Jordan ağları

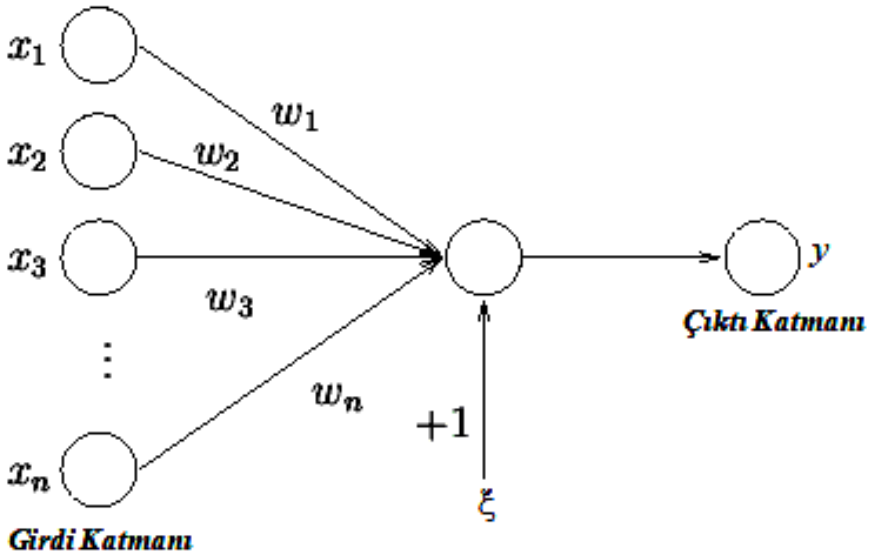
3.10.4.Yapay sinir ağı modelleri

Yapay sinir ağları oluşumlarını ele aldığımız bir önceki bölümde bahsi geçen ileri beslemeli yapay sinir ağlarıyla kurulan modellere değinilecektir. Bu modellerden ileri beslemeli tek katmanlı ve tezin uygulama bölümünde kullanılacak olan çok katmalı YSA anlatılacaktır.

3.10.4.1. Tek katmanlı YSA

Rosenblatt tarafından 1958 geliştirilen model, eğitilebilme niteliğini taşıyan ilk yapay sinir ağıdır. Sadece girdi ve çıktı katmanından oluşan tek katmanlı ağlarda çıktı birimleri girdi birimlerinin her birine bağlanmaktadır. Bu modellerde daima ağın çıktısından sıfır sonucunun alınmasına engel olan eşik değeri (ξ) vardır (32). Tek katmanlı YSA, hem giriş hem de çıkış katmanı olmasına rağmen tek katmanlı diye anılmasının sebebi giriş ünitesinde hesap yapılmadığı için bu katmanın dikkate alınmamasıdır. Şekil 18 'de gösterildiği üzere x_1, x_2, ..., x_n girdi katmanının değişkenlerini, w_1, w_2, ..., w_n her bir bağlantının ağırlığını, ξ eşik değerini (+1), y

çıktı birimini ifade etmektedir. Tek katmanlı YSA 'da ağdan istenilen, bir doğrusal fonksiyon yardımıyla gelen verileri öğrendiği forma göre iki sınıfa ayrılacak şekilde sınıflamaktadır. Tek katmanlı YSA genel olarak (4.2) formunda ifade edilir.



Şekil 18. Tek katmanlı YSA

y çıktıyı ifade eden doğrusal fonksiyon ve 0 ile 1 sınıflar olmak üzere;

$$y = f(\sum_{k=1}^{n} x_k w_k + \xi) \qquad (4.1)$$

$$\sum_{k=1}^{n} x_k w_k + \xi = Ķ \text{ ise } f(Ķ) = \begin{cases} 1 & y > 0 \\ 0 & y \leq 0 \end{cases} \qquad (4.2)$$

Tek katmalı algılayıcılardan; Basit Algılayıcı Modeli (Perceptron), Adaptif Doğrusal Eleman (ADALINE) ve Çoklu Adaptif Doğrusal Eleman (MADALINE) en çok bilinen modellerdir. Ancak söz konusu modeller burada incelenmeyecektir.

3.10.4.2. Çok katmanlı algılayıcılar

Çok katmanlı algılayıcılar (MLP); girdi katmanı, çıktı katmanı ve bu iki katman arasındaki bir ya da daha fazla gizli katmandan oluşan yapay sinir ağlarıdır. MLP 'yi tek katmanlı algılayıcılardan ayıran özelliği olan gizli katman sayesinde

doğrusal olmayan problemler için çözüm geliştirilmektedir. İleri beslemeli-geriye yayılımlı MLP birçok yapay sinir ağı uygulamasında kullanılmaktadır (37). Giriş katmanı *d* boyutlu *M* doğrusal kombinasyonu içerdiği ve (1) ilk katmanın ağırlığını ifade ettiği kabulüyle;

$$b_j = \sum_{i=0}^{d} w_{ij}^{(1)} x_i \qquad j = 1,2,\dots,M$$

YSA 'da eşitliğiyle verilen her bir işlem doğrusal olmayan aktivasyon fonksiyonları kullanılarak *g* fonksiyonuna dönüştürülür. Örneğin sigmoid aktivasyon fonksiyonu kullanılarak;

$$z_j = h(b_j) = \frac{1}{1+\exp(-b_j)} \qquad (4.3)$$

eşitliğine dönüşür. z_j değerleri gizli katmanın çıktısı olarak yorumlanmaktadır. İkinci katmanda gizli ünitelerin çıktıları, *K* çıkış aktivasyon toplamını elde edebilmek için doğrusal olarak birleştirilir. İkinci katmanın ağırlığı ve yanlılık değeri olarak $z_0 = 1$ kabul edildiğinde;

$$a_k = \sum_{j=0}^{M} w_{kj}^{(2)} z_j \qquad k = 1,2,\dots,K \qquad (4.4)$$

eşitliği elde edilir. Bu çıktı birimleri bir aktivasyon fonksiyonu (örneğin sigmoid fonksiyon) kullanıldıktan sonra aşağıdaki eşitliğe dönüştürülerek,

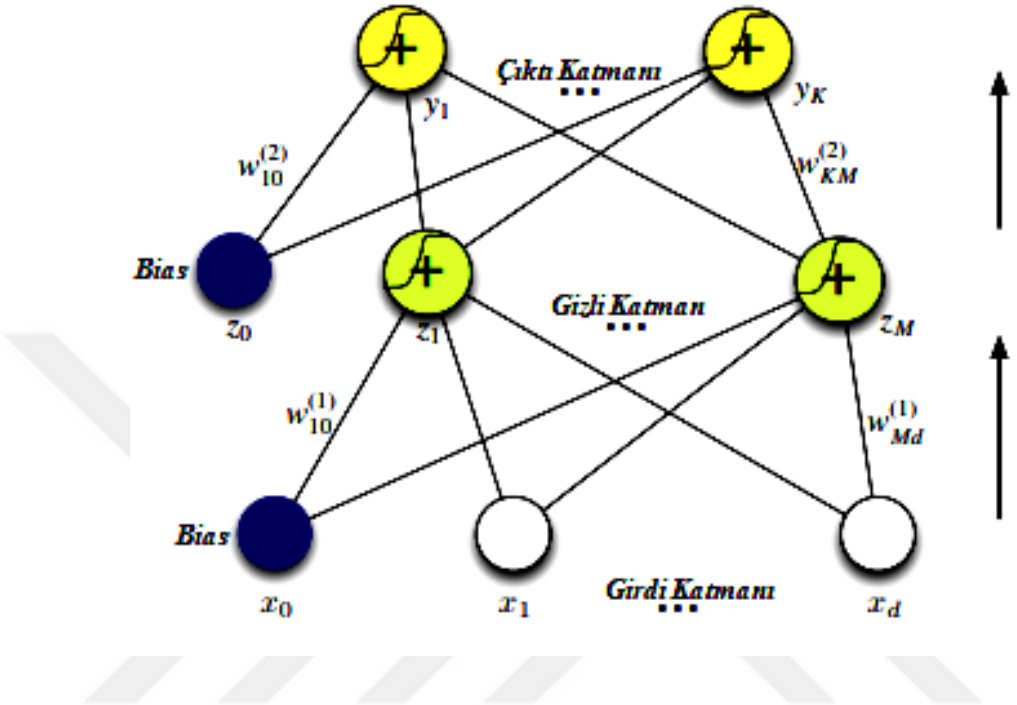
$$y_k = g(a_k) = \frac{1}{1+\exp(-a_k)}$$

(4.5)

ikili sınıflamada kullanılan eşitlik elde edilir. Çoklu sınıflama için softmax aktivasyon fonksiyonu kullanılarak;

$$g(a_k) = \frac{\exp(a_k)}{\sum_{l=1}^{K} \exp(a_l)} \qquad (4.6)$$

çoklu sınıflama denklemi elde edilir (https://www.inf.ed.ac.uk/teaching/courses/inf2b/- learnnotes/inf2b-learn12-notes-nup.pdf, Erişim tarihi: 16.12.2017).



Şekil 19. Çok katmanlı algılayıcı

MLP 'de danışmanlı öğrenme yöntemi kullanılmaktadır. Yani MLP 'ye girdiler ve çıktılar verilerek ağın modeller geliştirmesi ve yeni veri setleri için çözümleme yapması beklenir. MLP ağının öğrenme kuralı en küçük kareler yöntemine dayalı genelleştirilmiş delta kuralıdır. Bu öğrenme kuralı, nöronun gerçek çıkışı ile istenilen çıkış değerleri arasındaki farkı azaltan, giriş bağlantılarını güçlendiren ve sürekli olarak değiştiren bir düşünceye dayanmaktadır.

3.10.5. İleri beslemeli ağlarda geriye yayılım (BP)

Geriye yayılım (BP) algoritması, ileri beslemeli ağlar için en yaygın kullanılan eğitim yöntemidir. "*k*" adet gizli katmanı olan bir MLP' yi ele aldığımızda giriş (0) ve çıkış (*k*+1) katmanlarıyla birlikte toplam *k*+2 adet katmanı olan YSA elde edilir. Giriş katmanı birimlerini *K*, çıkış katmanı birimlerini *L* ve *m* gizli katmanının

birimlerini N^m olarak kabul edelim. *m* gizli katmanında *j*. birimin ve *m+1* gizli katmanındaki *i*. birimin ağırlıkları w_{ij}^m ile gösterilmektedir. *m* katmanındaki *i*. birimin aktivasyonu x_i^m ile gösterilsin. İleri beslemeli ağların eğitilmesi için kullanılan eğitim seti *T*, giriş-çıkış veri çiftlerinden oluşur, *n* eğitim örneğini belirtmek üzere; $u(n) = (x_1^0(n), \dots, x_K^0(n))^t$ ve $d(n) = (d_1^{k+1}(n), \dots, d_L^{k+1}(n))^t$ eşitlikleriyle giriş birimi dışındaki birimlerin aktivasyonu;

$$x_i^{m+1} = f(\sum_{j=1,\dots,N^m} w_{ij}^m x_j(n)) \tag{4.7}$$

eşitliği ile elde edilir. *u(t)* eğitim girdisi ile belirlenen denklem, çıktı katmanında $y(n) = (x_1^{k+1}(n), \dots, x_L^{K+1}(n))^t$ formunda bir ağ yanıtı elde edilene kadar art arda gelen gizli katmanlardaki birimlerin aktivasyonlarını hesaplamak için kullanılır. Eğitimin amacı hata kareler toplamı olan;

$$E = \sum_{n=1,\dots,T} \|d(n) - y(n)\|^2 = \sum_{n=1,\dots,T} E(n) \tag{4.8}$$

(4.8) 'in minimum sonuçlanacağı şekilde bir dizi ağ ağırlığı bulmaktır. Bu ağırlık bulma işlemi ağırlıkları, küçük bir öğrenme oranı olan γ 'yı kullanarak;

$$\frac{\partial E}{\partial w_{ij}^m} = \sum_{t=1,\dots,T} \frac{\partial E(n)}{\partial w_{ij}^m} \tag{4.9}$$

(4.9) 'un ağırlıklarına göre hata eğimi yönü doğrultusunda aşamalı olarak değiştirme işlemi ile yapılır. (γ 'nın kullanımı: yeni $w_{ij}^m = w_{ij}^m - \gamma \frac{\partial E}{\partial w_{ij}^m}$)

Bu formül tüm eğitim örneklerini sunduktan sonra yeni ağırlıkların hesaplandığı toplu (batch) öğrenme modunda kullanılan formüldür.*

Veri setindeki bütün örnekler böyle bir devir (döngü) denen işlemden geçmektedir. Döngü başlatılmadan önce ağırlıklar küçük rasgele sayılar ile

* ML 'de YSA gibi öğrenen sistemlerde öğrenme bazı kurallara göre gerçekleştirilmektedir. Bunlar çevrimiçi (on-line), kullanıma alınmadan önce örnekler üzerinde eğitilen ve çalışmamızda kullanacağımız çevrimdışı (off-line-batch) yöntemidir.)

oluşturulur. Bir varyant (versiyon/senaryo) eğitim örneklerinin sunumundan sonra ağırlıkların değiştiği artan (öğrenilen bilgilerle optimize edilen) öğrenmedir. Geriye yayılım algoritmasının hesaplama adımları şunlardır:

- Her n boyutlu örneklem için katman içindeki ve çıkış birimlerindeki aktivasyonlar hesaplanır (ileri besleme yapılır)
- Geriye doğru hesaplamalar (*m=k+1, k, ..., 1* şeklinde) yapılır. x_{ij}^{m} ifadesinin her birimi için hata yayılım terimi $\delta_i^m(n)$ olmak üzere çıkış katmanı;

$$\delta_i^{k+1}(n) = (d_i(n) - y_i(n))\frac{\partial f(u)}{\partial u}|_{u=z_i^{k+1}}$$

ile hesaplanır. Gizli katman;

$$\delta_j^m(n) = \sum_{i=1}^{N^{m+1}} \delta_i^{m+1} w_{ij}^m \frac{\partial f(u)}{\partial u}|_{u=z_j^m}$$

şeklinde düzenlenir. Burada $z_i^m(n) = \sum_{j=1}^{N^{m-1}} x_j^{m-1}(n) w_{ij}^{m-1}$ denklemi, x_i^m biriminin içindeki ifadeyi belirtmektedir.

- Son adımda ağırlıklar aşağıdaki forma göre ayarlanır:

$$new\ w_{ij}^{m-1} = w_{ij}^{m-1} + \gamma \sum_{t=1}^{T} \delta_i^m(n) x_j^{m-1}(n)$$

Her döngüden sonra hata hesaplanır. Döngünün durması için; hata miktarının veya hatadaki değişimin önceden belirlenen bir eşiğin altına düşmesi veya döngü sayısının maksimuma ulaşması gerekmektedir. Yeterince küçük hata için döngüler işlemeye devam eder (34).

3.11. Model Performans Ölçütleri

Makine öğrenimi yaklaşımlarında ikili (binary) sınıflandırıcılar oldukça yaygın bir şekilde kullanılmaktadır. Hem sınıflandırıcıların performansını değerlendirmek hem de modelin geçerliliğini inceleyebilmek için performans ölçütlerine ihtiyaç duyulmaktadır. Bu kısımda binary sınıflandırıcılar için oluşturulan sınıflandırma tablosu (Tablo 2) yardımıyla hesaplanan performans ölçütlerine yer verilecektir. Literatürde birçok ölçüt bulunmasına rağmen tez çalışmasında doğruluk (accuracy), Matthews korelasyon kastsayısı (MCC), F ölçütü (F-measure), ve Ayırsama Gücü (AG) yöntemine yer verilecektir.

Tablo 2. Hesaplamaların yapıldığı sınıflama tablosu

		Gerçek Sınıflama		
		Hasta (+)	**Sağlam (-)**	
Tahmin Edilen Sınıflama	**Hasta (+)**	Doğru Pozitif DP	Yanlış Pozitif YP	Pozitif Kestirim Değeri $\frac{DP}{DP+YP}$
	Sağlam (-)	Yanlış Negatif YN	Doğru Negatif DN	Negatif Kestirim Değeri $\frac{DN}{YN+DN}$
		Duyarlılık $\frac{DP}{DP+YN}$	Seçicilik $\frac{DN}{YP+DN}$	Doğruluk $\frac{DP+DN}{DP+YP+YN+DN}$

3.11.1. Doğruluk (Accuracy)

Testin hasta-sağlam olarak doğru bir şekilde tahmin edilmiş toplam doğru tanı oranına veya geçerlilik katsayısına doğruluk denir. Diğer bir ifade ile doğruluk, bir belirleyicinin tüm örnekleri doğru bir şekilde tanımlama yeteneğini ölçmektedir (38).

$$Doğruluk = \frac{DP + DN}{DP + YP + YN + DN}$$

3.11.2. Matthews korelasyon katsayısı (MCC)

Gözlenen ve tahmin edilen sınıflamalar arasındaki korelasyon katsayısıdır. Bağımlı değişkene ait prevalans değerinin dengeli olmadığı durumlarda oldukça başarılı sonuçlar vermektedir. MCC, -1 ile +1 arasında değer almaktadır. MCC değeri; 0'a yaklaştıkça rasgele bir tahmin yapıldığını, -1 değerine yaklaştıkça yanlış bir tahminleme yapıldığını, +1 değerine yaklaştıkça doğru tahminleme yapıldığını belirtmektedir (39).

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

3.11.3. F ölçütü (F-measure)

Bu ölçüt kesinlik ile duyarlılık arsındaki dengeyi ifade eder. Kesinlik veya duyarlılıktan herhangi biri sıfır olduğunda bu ölçüt 0 olarak hesaplanır. Bağımlı değişkene ait prevalans değeri dengeli olduğunda iyi sonuçlar alınır. PKD pozitif kestirim değerini ifade etmek üzere;

$$F - measure = 2 * \frac{PKD * Duyarlılık}{PKD + Duyarlılık}$$

eşitliği ile hesaplanır (39).

3.11.4. ROC eğrisi ve eğri altında kalan alan (AUC)

ROC eğrisinin ML 'deki algoritması şu şekildedir: Öncelikle bilgisayar programı üzerinden aralıklar ve her bir aralıkta cut-off değeri belirlenir. Belirlenen aralıklar için olasılık tablosu hesaplanır. Her bir aralık için hesaplanan duyarlılık ve 1- seçiciliğe göre ROC grafiği çizilir. ROC grafiği (eğrisi) altında kalan alan (AUC) hesaplanır. AUC değeri 1'e yaklaştıkça modelin performansı artmaktadır.

3.11.5. Ayırsama Gücü (Discriminant Power - AG): Bu ölçüt duyarlılık ve seçiciliği özetleyerek sınıflandırıcının pozitif ve negatif durumlar arasında ne kadar iyi ayrım yaptığını değerlendirir. Aşağıdaki formülle hesaplanır;

$$AG = \frac{\sqrt{3}}{\pi}(\log\left(\frac{duyarlılık}{1-duyarlılık}\right) + \log(\frac{seçicilik}{1-seçicilik})$$

AG değeri 0-1 aralığında olduğunda sınıflandırıcının zayıf olduğu, 1-2 aralığında orta iyilikte, 2-3 aralığında ise başarılı olduğu anlaşılır (39).

4. GEREÇ ve YÖNTEM

Tez çalışmasında iki farklı yaklaşım ile performans karşılaştırmaları yapıldı. Birinci yaklaşımda gerçek veri seti, İnönü Üniversitesi Turgut Özal Tıp Merkezi Tıbbi Patoloji Anabilim Dalı ‘ndan edinildi. Çalışma için gerekli izinler alındıktan sonra, veri seti olarak kullanılan 302 adet invaziv duktal karsinomlu (IDC) hastanın değerleri, 01.10.2008-30.06.2016 tarihleri arasındaki 75.000 biyopsi raporunun her biri incelenerek 1470 meme kanseri vakasından elde edildi. Çalışmada meme kanseri için prognostik faktör olduğu bilinen değişkenlerle birlikte toplam 15 bağımsız değişken kullanılarak iki sınıflı bağımlı değişkenin sınıflandırılmasına çalışıldı (Tablo 3). Bu sınıflandırmadaki amaçlar; bahsi geçen bağımsız değişkenlerle sağkalımın (sağ/ölü) ne kadar doğrulukla tahmin edildiğini belirlemek, farklı modellerin performanslarını başarı ölçütlerine göre değerlendirmek ve hangi değişkenin bu sınıflamaya daha çok katkısı olduğunu öğrenmektir. Tablo 3 ’te cinsiyet değişkeninin olduğu görülmektedir. Genellikle meme kanseri vakaları kadınlarda görülmesine rağmen veri setinin orjinalliğinin bozulmaması açısından bu değişkene yer verildi.

İkinci yaklaşımda ise benzetim (simülasyon) çalışmasıyla, bağımlı değişkende riskli durumun görülme oranına (prevalans), gözlem sayısına ve bağımsız değişken sayısına göre farklı veri setleri ile analizler gerçekleştirildi. Bu veri setlerinde bağımlı değişken için 0,2-0,4-0,6 ve 0,8 gözlem oranlarına göre binom dağılımdan elde edilen veri grupları analize dahil edildi. Bağımsız değişkenler, sıfır ortalamalı bir standart sapmalı çok değişkenli normal dağılımdan türetilen veri setlerinden oluşturuldu. Çalışma için 15, 25 ve 35 bağımsız değişken kullanılarak oluşturulan veri setleri ile binom dağılımdan elde edilen veri grupları birleştirilerek 100 tekrarlı 24 farklı veri seti elde edildi. Her bir tekrar için elde edilen performans ölçütleri ayrı ayrı hesaplanarak ölçütler arasındaki ilişki ve bağlantılara göre makine öğrenimi yöntemlerinin performansı değerlendirildi.

Veri seti için sınıflama yaparken R 3.3.3 (R programlama dilleri/Project) ve RSudio ara yüzü kullanıldı. DVM için gerekli komutların bulunduğu "e1071" paketi yüklendi. Bu paket doğrusal, polinomiyal, sigmoid ve radyal tabanlı çekirdek fonksiyonları içermektedir. RF için “randomForest” paketi ve YSA için “neuralnet”,

"nnet" ve "mxnet" paketleri yüklendi. Hiperparametreleri belirlemek için "caret" paketi kullanıldı. Çok değişkenli normal dağılımdan veri türetmek için "MASS" paketi ve diğer uygulamalar için SPSS 21 (Armonk, NY: IBM Corp.) paket programı kullanıldı. Bazı bağımsız değişkenlerdeki ortalama % 3-4'lük kayıp verinin yanı sıra özellikle ki-67 proliferasyon indeksi bağımsız değişkenindeki yaklaşık %65'lik kayıp veriye "rfimpute" komutuyla atama yapıldı. Ki-67 proliferasyon indeksindeki % 65 'lik kayıp verinin sebebi, patoloji biriminin sadece son yıllarda "ki-67" değerini değerlendirip raporlara yazmasından kaynaklanmaktadır.

Meme karsinomlu veri setinin analizinde DVM, RF ve YSA modelleri için en uygun hiperparametreler, 10 katlı 5 tekrarlı çapraz geçerlilik metodu ve farklı hiperparametreler denenerek, yanlış sınıflandırma oranının küçüklüğüne göre belirlendi. Bütün makine öğrenimi yöntemlerinde veri setinin %70'i modeli eğitmek (eğitim seti) için %30'u ise öğrenilen modeli test etmek için (test seti) olarak kullanıldı. Modeller veri setine uygulanırken DVM için doğrusal, polinomiyal, sigmoid ve radyal tabanlı çekirdeklerden faydalanıldı. RF yöntemi için dallara ayırıcı karar değişkeni sayısı (m_{try}) 3 ve 4 olarak belirlendikten sonra her değişken sayısı için 100 ve 500 karar ağacı ile analiz gerçekleştirildi. YSA 'da ise sigmoid, hiperbolik tanjant, relu ve doğrusal aktivasyon fonksiyonlarına göre sonuçlar elde edildi.

Tez çalışması birinci yaklaşım için kullanılan değişkenlere ilişkin açıklamalar:

- *Tümör çapı;* tümör boyutu, ilgili aksiller lenf nodlarının varlığı ve sayısı ile ilişkilidir. Küçük tümör çapının prognozunun büyük çaplı tümörlere göre daha iyi olduğu bilinmektedir. Ayrıca tümör çapı, hastalık nüksüyle korelasyona sahip bağımsız bir prognostik faktördür (40).
- *Tümör nekrozu faktörü;* birçok hücre tipi tarafından salgılanan ve kanserli hücrelerin yıkımını sağlayan bir sitokindir. Tümör nekrozunun kötü prognostik faktör olduğu ayrıca aksiler LD (+) olgularda ölüm oranının daha fazla olduğu bildirilmektedir (41).

Tablo 3: Birinci yaklaşım için bağımlı ve bağımsız değişkenler

Bağımlı Değişken	Bağımsız Değişkenler
Sağkalım (Sağ/Ölü)	Yaş
	Cinsiyet
	Beden Kitle İndeksi (BKI)
	Tümör Çapı
	Tümör Nekrozu
	Er (Östrojen Reseptörü)
	Pr (Progesteron Reseptörü)
	C-erb B2 (Her2)
	Lenf Nodu Sayısı
	Lenf Nodu Metastazı
	LVI (Lenfovasküler Invazyon)
	Ki-67 (Proliferasyon Indeksi)
	Kemoterapi
	Radyoterapi
	Hormonoterapi

IDC veri setinde bağımlı değişken iki sınıflı (binary) değişken olup hastaların hayatta olup olmamasına göre değer almaktadır. Bağımsız değişkenlerden "yaş" ve "lenf nodu sayısı" değişkenleri sürekli diğer bütün değişkenler kategorik değişkenlerdir.

- *Hormon reseptörleri Er (Östrojen) & Pr (Progesteron);* bu reseptörler, konsantrasyondaki değişime bağlı olarak, dolaşım vasıtasıyla hücre içine alınıp hormon molekülüne seçici olarak bağlanarak hormon reseptör kompleksini oluştururlar. Er ve Pr pozitif tümörler hormonal tedaviye yanıt vermekte ve iyi prognoz gösterme eğilimindedirler (42).

- *Her-2 (CerbB-2);* hücre büyümesine neden olan özel bir büyüme faktörünün reseptör proteinini kodlamaktadır. Her-2/neu pozitif olduğu belirlenen meme kanseri hastaları genellikle daha kötü bir prognoza sahiptir (https://labtestsonline.org.tr/tests/her-2neu, Erişim Tarihi: 25.09.2018).

- *Lenf nodu metastazı;* metastatik aksiller lenf nodu oranı meme kanserinde önemli bir prognostik faktör olarak bilinir. Genellikle yüksek metastatik lenf

nodu sayısı kötü prognoza sahiptir. Kanser hücrelerinin bölgesel lenf düğümlerine yayılması meme kanserinde en önemli prognostik faktördür (43).

- *LVI (Lenfovasküler Invazyon);* meme tümörlerinin üçte birinde lenfovasküler invazyon (LVI) mevcuttur. LVI, adjuvan kemoterapinin tek bir göstergesi olarak artmış lenf nodu metastazı ve sistemik hastalığa doğru ilerleme/artma riski ile birliktelik gösterir (44).

- *Ki-67 skoru;* tümör hücrelerinin G1, S, G2 ve M fazlarında exprese edilen core proteini ve solid tümör proliferasyon markerıdır (45).

5. BULGULAR

Birinci yaklaşım için kullanılan veri setinde 302 meme karsinomlu hastanın yaş ortalaması (yıl) 50.54 ± 12.38 ve ortanca sağkalım süresi 39 ay olarak hesaplandı. Meme kanseri hastalarının, 222'sinin (%73,5) hayatta olduğu 80 hastanın (%26,5) meme kanseri nedeniyle hayatını kaybettiği anlaşıldı.

Ki-67 skoru raporlarda yüzde değerler üzerinden nümerik olarak ifade edilmektedir. Ancak cerrahi yaklaşıma göre %14 değerinin kritik bir değer olabileceği öngörüldüğünden kategorilendirme, bu değer esas alınarak gerçekleştirildi. Kategorilendirilmiş ki-67 skorlarının sağkalım sınıflamasına katkı sağlayacağı düşüncesi ön plana çıkmaktadır (p=0,045).

Lenf nodu sayısı ham haliyle Tablo 4 'te belirtildiği gibi değer almaktadır (p=0,549). Literatürde bu değerin sağkalıma ciddi bir etkisinin olamayacağı öngörülmekle birlikte makine öğrenimi yöntemlerinin performansına katkı sağlayacağı düşünüldüğünden standardize edilerek sınıflamaya dahil edilmiştir.

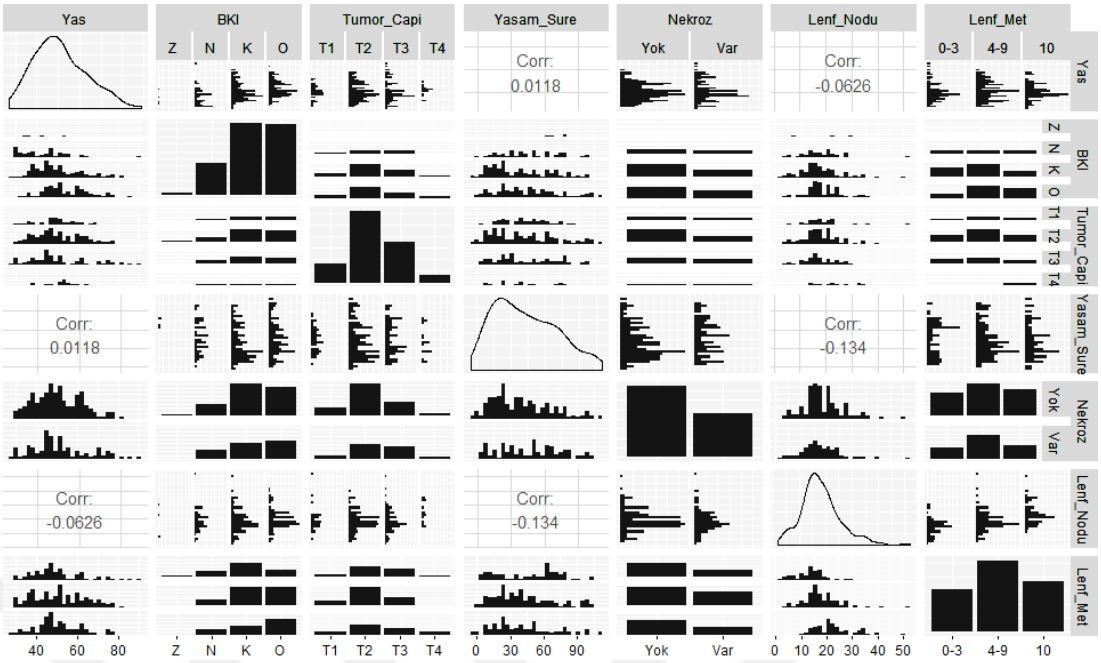
Bir diğer önemli faktör olan BKI kategorilerinin sağkalıma göre önemlilik ifade etmemesi (p=0,066), zayıf kategorisinde yeterli veriye ulaşılamamasından kaynaklanmaktadır. Bahsi geçen kategori analizden çıkarıldığında BKI 'nin sağkalım için önemli olduğu görülmektedir (p=0,035).

Alınan rutin tedaviler bakımından kemoterapinin (p=0,631) ve radyoterapinin (p=0,488) sağkalıma katkı sağlayamama durumu ortaya çıkmaktadır. Ancak bu tedaviler prognostik sürecin oldukça önemli parametreleri olduğundan modellemeye dahil edilmiştir. Yukarıda anlamlılığı değerlendirilen değişkenler dışındaki bütün değişkenler sağkalım ile ilişki göstermektedir (Tablo 4). Tablo 3 'te bağımlı-bağımsız değişken kurgusu anlatılan veri seti için danışmanlı makine öğrenimi algoritmalarıyla analiz gerçekleştirilmiştir.

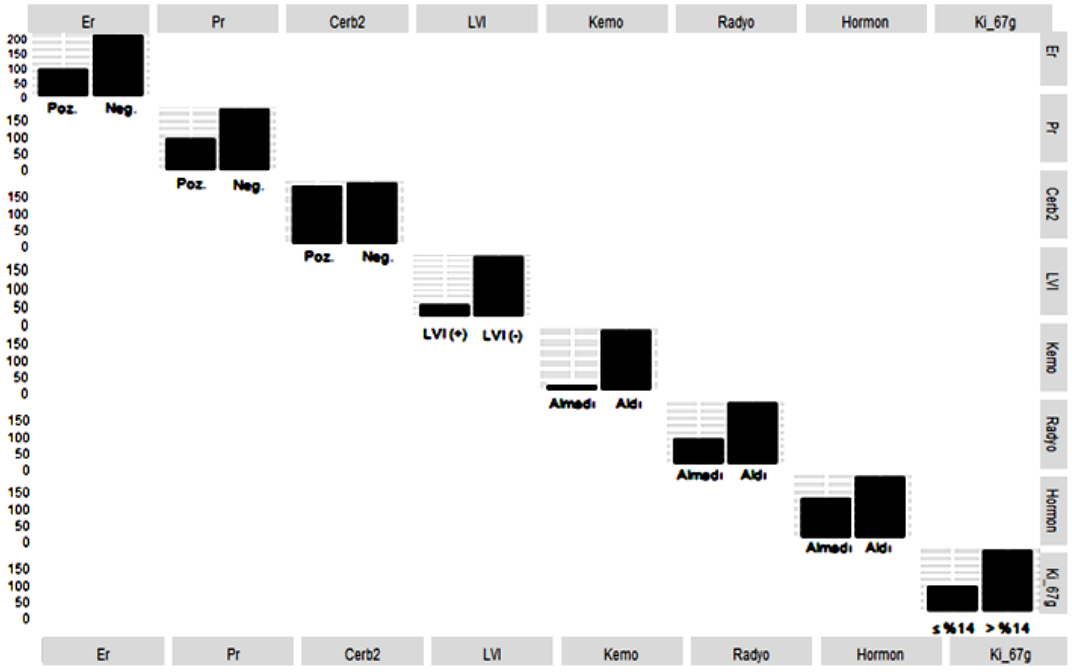
Tablo 4 : IDC veri setinde sağkalıma göre dağılım

		Sağkalım		
		Sağ	**Ölü**	**p**
Sağkalım		222 (73,5)	80 (26,5)	0,001*
Yaş		49,54 ± 12,16	53,32 ± 12,64	0,019**
Cinsiyet	Kadın	216 (75,3)	71 (24,7)	0,005****
	Erkek	6 (40)	9 (60)	
BKI	Zayıf	5 (83,3)	1 (16,7)	0,066*
	Normal	35 (67,3)	17 (32,7)	
	Aşırı Kilolu	100 (81,3)	23 (18,7)	
	Obez	82 (67,8)	39 (32,2)	
Tümör Çapı	T1	35 (85,4)	6 (14,6)	0,001*
	T2	123 (80,4)	30 (19,6)	
	T3	56 (63,6)	32 (36,4)	
	T4	8 (40)	12 (60)	
Tümör Nekrozu	Yok	154 (81,9)	34 (18,1)	0,001*
	Var	68 (59,6)	46 (40,4)	
Er	Negatif	54 (57,4)	40 (42,6)	0,001*
	Pozitif	168 (80,8)	40 (19,2)	
Pr	Negatif	68 (64,8)	37 (35,2)	0,012*
	Pozitif	154 (78,2)	43 (21,8)	
C-erb B2 (Her2)	Negatif	111 (79,3)	29 (20,7)	0,034*
	Pozitif	111 (68,5)	51 (31,5)	
Lenf Nodu Sayısı		18,37 ± 7,61	17,77 ± 7,78	0,549**
Lenf Nodu Metastazı	0-3 Lenf. Met.	67 (84,8)	12 (15,2)	0,009*
	4-9 Lenf. Met.	94 (73,4)	34 (26,6)	
	10+ Lenf. Met.	61 (64,2)	34 (35,8)	
LVI	LVI (-)	51 (87,9)	7 (12,1)	0,009***
	LVI (+)	171 (70,1)	73 (29,9)	
Ki-67	Ki-67 ≤ % 14	80 (80,8)	19 (19,2)	0,045*
	Ki-67 > % 14	142 (70)	61 (30)	
Kemoterapi	Almadı	14 (66,7)	7 (33,3)	0,631***
	Aldı	208 (74)	73 (26)	
Radyoterapi	Almadı	63 (70,8)	26 (29,2)	0,488*
	Aldı	159 (74,6)	54 (25,4)	
Hormonoterapi	Almadı	68 (57,1)	51 (42,9)	0,001*
	Aldı	154 (84,2)	29 (15,8)	

:** Pearson Ki-kare Test, *:** Bağımsız Student-t Test, *****:** Yates Düzeltmeli Ki-kare Test,
******:** Fisher Exact Ki-kare Test,



Şekil 20.a. Meme karsinoma verilerinin dağılımı



Şekil 20.b. Meme karsinoma verilerinin dağılımı

ML yöntemleri IDC veri setine uyguladığında eğitim verisi için en yüksek sınıflama doğruluğu (%98) doğrusal aktivasyon fonksiyonu ile YSA 'dan elde edildi. Test verisi için en yüksek sınıflama doğruluğu ise (%80) radyal tabanlı çekirdek ile DVM 'den elde edildi. RF yöntemi, eğitim verisi için diğer modellere göre genel olarak performanslı sonuçlar vermesine rağmen test verisinde benzer performansı gösteremediği anlaşılmaktadır. Test verisi için bütün modeller genel olarak değerlendirildiğinde sonuçların çok farklı olmadığı görülmektedir (%71-%79). Ancak en uygun ve performanslı model arayışında küçük farklılıklar bile tıbbi veriler için hayati önem arz edebilmektedir.

Model performansları, hem eğitim hem de test verileri için elde edilen sonuçlar dikkate alınarak değerlendirildi. Genellikle sınıflama doğruluğu dışındaki diğer performans ölçütleri (duyarlılık, seçicilik, F-ölçütü, MCC, AUC, AG) doğruluk ile paralel büyüme veya küçülme göstermektedir. Ancak bazı modellerde (polinomiyal çekirdekli DVM, relu aktivasyon fonksiyonlu YSA) doğruluk için yüksek değerler alınamamasına rağmen duyarlılıklarının %100 olduğu görülmektedir. Bu modeller incelendiğinde (Tablo 5), yüksek duyarlılığa rağmen seçicilikleri %0 olarak bulunmuş olup diğer performans ölçütlerinin (MCC, AG) hesaplanmasına da engel oldukları görülmektedir. Seçiciliğin sıfır olması modelin sınıf ayrımını yapamadığını ve ayırsama gücünün olmadığını göstermektedir. Dolayısıyla model değerlendirilirken sınıflama doğruluğu değerinden modelin performansı hakkında önemli bir ön bilgi edinilmektedir. Daha sonra diğer ölçütler de incelendiğinde nihai bir karara varılabilmektedir.

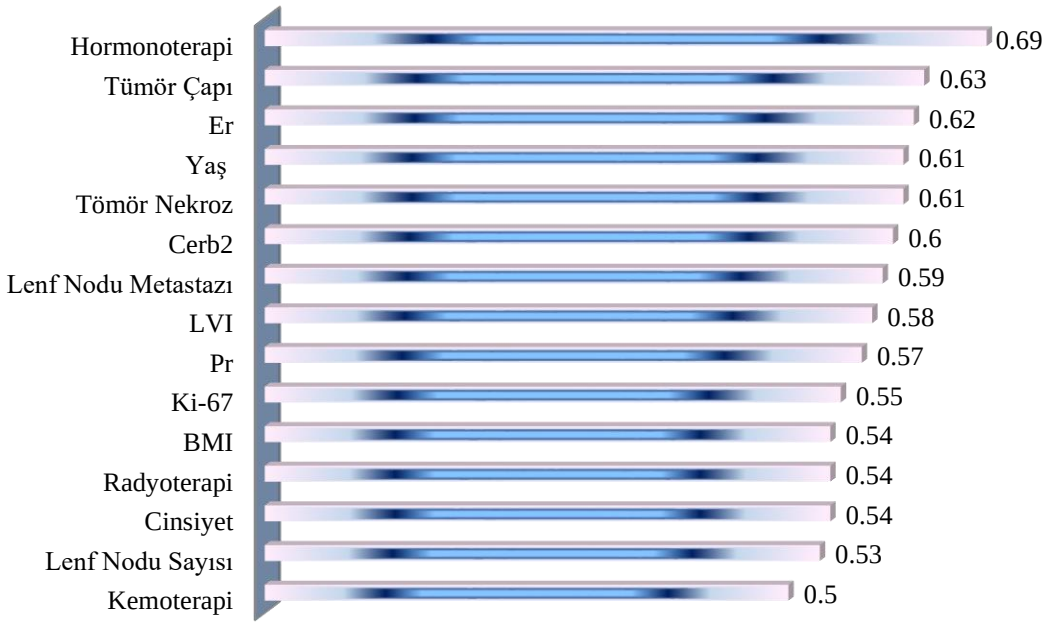
Tablo 5 : Gerçek veri seti için model performansları

		D^1	D^2	S^1	S^2	SP^1	SP^2	F^1	F^2	MCC^1	MCC^2	AUC^1	AUC^2	AG^1	AG^2
DVM	Doğrusal	0,82	0,76	0,93	0,91	0,50	0,33	0,88	0,85	0,49	0,16	0,86	0,84	0,62	0,39
	Polinomiyal	0,74	0,73	1,00	1,00	0,00	0,00	0,85	0,85	!	!	0,78	0,77	!	!
	Sigmoid	0,75	0,73	1,00	1,00	0,00	0,00	0,85	0,85	!	!	0,78	0,77	!	!
	Radyal	0,93	0,80	0,99	0,92	0,77	0,42	0,96	0,87	0,83	0,22	0,98	0,89	1,49	0,52
RF	m_{try}=3, ntree=100	0,82	0,77	0,83	0,79	0,76	0,64	0,88	0,86	0,48	0,14	0,83	0,7	0,65	0,45
	m_{try} =3, ntree=500	0,82	0,75	0,83	0,78	0,74	0,55	0,88	0,84	0,48	0,11	0,83	0,7	0,63	0,34
	m_{try} =4, ntree=100	0,82	0,73	0,83	0,78	0,73	0,47	0,88	0,82	0,48	0,10	0,82	0,66	0,62	0,28
	m_{try} =4, ntree=500	0,81	0,73	0,84	0,77	0,64	0,46	0,88	0,83	0,40	0,09	0,84	0,69	0,53	0,25
YSA	Sigmoid	0,72	0,77	1,00	1,00	0,00	0,00	0,84	0,87	!	!	0,8	0,83	!	!
	Hiperbolik Tanjant	0,75	0,77	0,96	0,97	0,17	0,14	0,85	0,86	0,23	0,18	0,82	0,84	0,39	0,40
	Relu	0,73	0,76	1,00	1,00	0,00	0,00	0,84	0,86	!	!	0,78	0,82	!	!
	Doğrusal	0,98	0,71	0,99	0,74	0,93	0,63	0,98	0,79	0,94	0,12	0,98	0,81	1,82	0,38

[1]: Eğitim verisi, **[2]:** Test verisi; **D:** Doğruluk, **S:** Duyarlılık, **SP:** Seçicilik, **F:** F-ölçütü, **MCC:** Matthews korelasyon kastsayısı, **AG:** Ayırsama gücü, **!:** Hesaplanamayan değer.

Kullanılan makine öğrenimi yöntemleriyle hangi değişkenin sınıflamaya ne kadar katkısının olduğu da anlaşılabilmektedir. Bu analiz için performansı en yüksek modellemelerle (DVM 'de radyal tabanlı çekirdek, RF 'de m_{try}=3 ve ntree=100 ve YSA 'da doğrusal aktivasyon fonksiyon) değişken önemliliği analizi yapılmıştır. Şekil 21.a.'da DVM ile yapılan sıralamaya göre hastalara hormonoterapi uygulamak, sağkalımı belirlemede en önemli değişken (0,69) olmakla birlikte daha sonra gelen tümör çapı (0,63), hormon reseptörlerinden Er (0,62), hastanın yaşı (0,61), tümörde

nekroz varlığı (0,61) ve Her2 (0,60) değişkenleri birbirine yakın önemlilik göstermektedirler. Ancak cinsiyet (0,54) ve kemoterapinin (0,54) sınıflamaya çok az katkı sağladığı anlaşılmaktadır.



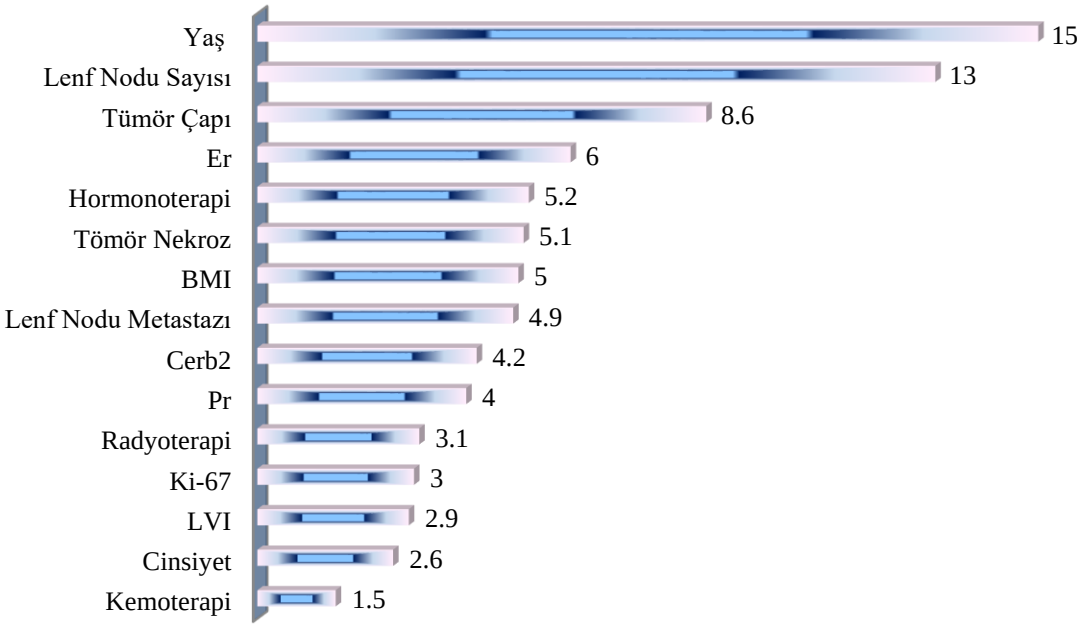
Şekil 21.a. DVM radyal tabanlı çekirdek ile önemli değişken sıralaması

RF yönteminde ise hastanın yaşı (15), lenf nodu sayısı (13) ve tümör çapı (8,6) en önemli üç değişken olarak karşımıza çıkmaktadır. RF yönteminde de cinsiyet (2,6) ve kemoterapinin (1,5) sınıflamaya çok fazla katkısının olmadığı görülmektedir (Şekil 22.b).

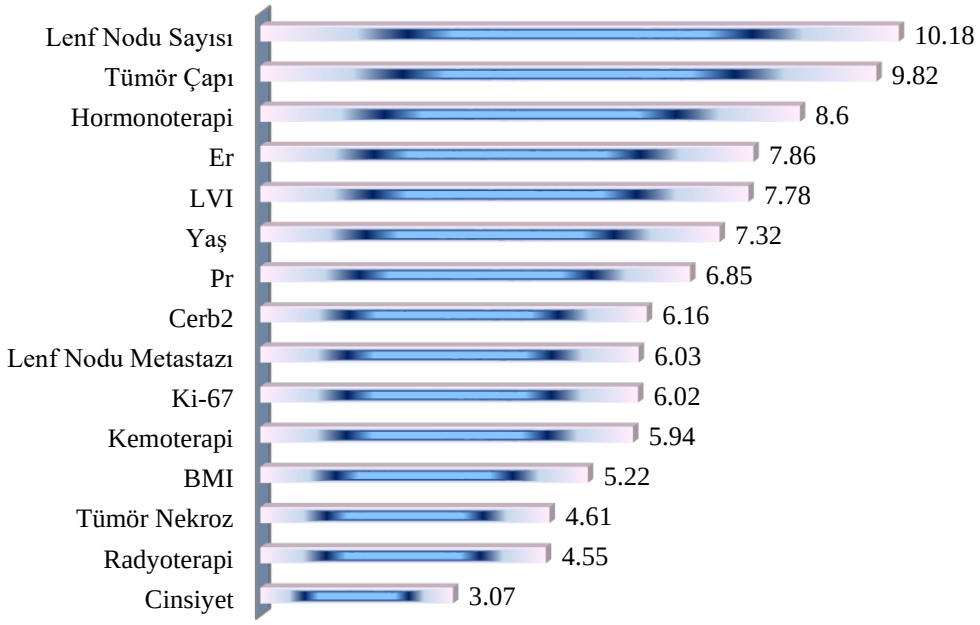
YSA 'da sağkalım sınıflaması için en önemli değişkenin lenf nodu sayısı (13) olduğu daha sonra tümör çapı (9,82) ve hormonoterapinin (8,6) geldiği anlaşılmaktadır. Sınıflamaya en az katkısı olan değişkenler ise radyoterapi (4,55) ve cinsiyet (3,07) olduğu görülmektedir (Şekil 22.c).

Bütün yöntemler birlikte değerlendirildiğinde sağkalımı belirlemede en önemli değişkenlerin; yaş, lenf nodu sayısı, tümör çapı, hormonoterapi ve Er hormon reseptörü olduğu anlaşılmaktadır. Orta önemlilikte olduğu anlaşılan değişkenler; tümör nekrozu, lenf nodu metastazı, cerb2, LVI ve Pr hormon reseptörüdür. Diğer

değişkenlere göre katkısı daha az olan değişkenler ise kemoterapi, BKI, radyoterapi, ki-67 ve cinsiyet olduğu görülmektedir.



Şekil 21.b. RF'de m_{try}=3 ve ntree=100 ile önemli değişken sıralaması



Şekil 21.c. YSA 'da sigmoid aktivasyon fonksiyon ile önemli değişken sıralaması

Simülasyon ile elde edilen 24 farklı veri seti için; eğitim-test seti doğrulukları, doğruluk farkları, doğruluk farklarının önemlilik testi (H_0: Doğruluk farkı sıfırdan farksızdır hipoteziyle, tek örneklem t testi) eğitim-test verisi için pozitif tahminleyici ve F-ölçütü değerleri analiz edildi. Veri setlerinden 100 tekrar için ayrı ayrı hesaplanan yukarıdaki ölçütlerin ortalamaları tablolara (Tablo 6-9) yerleştirildi. Ancak önemlilik testinde, doğruluk farklarının her bir model ve her bir prevalanstaki değişken sayısından elde edilen değerlerine göre istatistiksel anlamlılık hesaplandı.

Risk görülme oranı yaklaşık % 20 (0,2 prevalans) olduğunda 500 gözlem için bütün değişken sayılarında (15, 25 ve 35) test seti için en başarılı sonuç (0,8000-0,7998-0,7954) DVM 'den elde edildi. Eğitim ve test setinden elde edilen doğrulukların farkına bakıldığında en düşük fark ortalamalarının genel olarak DVM modellerinde olduğu görülmektedir. Aynı risk oranında 1000 gözlem içeren veri setinde, 15 ve 25 bağımsız değişken olduğunda en yüksek sınıflama doğruluğu (0,8003-0,8040) ve en düşük doğruluk farkı DVM 'den elde edilirken 35 bağımsız değişken olduğunda en yüksek sınıflama doğruluğu (0,8041) ve en düşük doğruluk farkı RF 'den elde edildi (Tablo 6).

Risk görülme oranı yaklaşık % 40 (0,4 prevalans) olduğunda 500 ve 1000 gözlem için bütün değişken sayılarında (15, 25 ve 35) test seti için en başarılı sonuç (0,5917-0,5923-0,5977 ve 0,5971-0,6009-0,5949) DVM' den elde edildi. Ancak eğitim ve test setinden elde edilen doğrulukların farkına bakıldığında en düşük fark ortalamalarının genel olarak RF yönteminden elde edildiği görülmektedir (Tablo 7).

Risk görülme oranı yaklaşık % 60 (0,6 prevalans) olduğunda 500 ve 1000 gözlem için bütün değişken sayılarında (15, 25 ve 35) test seti için en başarılı sonuç (0,5904-0,6034-0,6007 ve 0,5892-0,5975-0,5994) DVM' den elde edildi. Ancak eğitim ve test setinden elde edilen doğrulukların farkına bakıldığında en düşük fark ortalamalarının genel olarak RF yönteminden elde edildiği görülmektedir (Tablo 8).

Risk görülme oranı yaklaşık % 80 (0,8 prevalans) olduğunda 500 gözlem için bütün değişken sayılarında (15, 25 ve 35) test seti için en başarılı sonuç (0,8002-0,8028-0,8063) DVM 'den elde edildi. Aynı risk oranında 1000 gözlem içeren veri setinde, 15 ve 25 bağımsız değişken olduğunda en yüksek sınıflama doğruluğu (0,8029-0,8019) ve en düşük doğruluk farkı DVM 'den elde edilirken 35 bağımsız

değişken olduğunda en yüksek sınıflama doğruluğu (0,7994) ve en düşük doğruluk farkı RF ’den elde edildi (Tablo 9).

Tablo 6 : Bağımlı değişkende riskli durumun yaklaşık % 20 olduğunda performanslar

			D^1	D^2	ΔD	p	PKD^1	PKD^2	F^1	F^2
N=500	*p*=15	DVM	0,8004	0,8000	0,0040	0,861	0,8000	0,8000	0,8886	0,8882
		RF	0,7933	0,7899	0,0034	0,116	0,9914	0,9842	0,8844	0,8819
		YSA	0,9064	0,7319	0,1745	0,001	0,9061	0,7982	0,9460	0,8399
	p=25	DVM	0,7981	0,7998	-0,0019	0,001	0,7981	0,7998	0,8876	0,8887
		RF	0,7931	0,7906	0,0025	0,014	0,9974	0,9937	0,8845	0,8828
		YSA	0,9221	0,7410	0,1811	0,001	0,9200	0,8024	0,9558	0,8456
	p=35	DVM	0,7943	0,7954	-0,0010	0,809	0,7943	0,7954	0,8853	0,8859
		RF	0,7913	0,7930	-0,0160	0,517	0,9979	0,9950	0,8833	0,8843
		YSA	0,9304	0,7477	0,1826	0,001	0,9330	0,8085	0,9602	0,8507
N=1000	*p*=15	DVM	0,7999	0,8003	-0,0011	0,006	0,7997	0,8003	0,8886	0,8890
		RF	0,7931	0,7880	0,0051	0,001	0,9942	0,9880	0,8845	0,8813
		YSA	0,8542	0,7489	0,1053	0,001	0,8580	0,8045	0,9159	0,8513
	p=25	DVM	0,8026	0,8040	-0,0013	0,012	0,8026	0,8040	0,8904	0,8912
		RF	0,8052	0,8012	0,0039	0,239	0,9992	0,9980	0,8920	0,8894
		YSA	0,9061	0,7237	0,1823	0,001	0,9079	0,8015	0,9446	0,8335
	p=35	DVM	0,7966	0,7936	0,0030	0,346	0,7992	0,7995	0,8883	0,8884
		RF	0,8007	0,8041	0,0002	0,428	0,9999	0,9866	0,8893	0,4811
		YSA	0,9309	0,7265	0,2043	0,001	0,9417	0,8025	0,9589	0,8359

[1]: Eğitim verisi, **[2]:** Test verisi; **D:** Doğruluk, ***p:*** Bağımsız değişken sayısı, Δ**D:** Doğruluk farkı ortalamaları, **p:** Tek örneklem-t test anlamlılığı, **PKD** pozitif kestirim değeri, **F:** F-ölçütü.

Tablo 7: Bağımlı değişkende riskli durumun yaklaşık % 40 olduğunda performanslar

			D^1	D^2	ΔD	p	PKD^1	PKD^2	F^1	F^2
N=500	p=15	DVM	0,6265	0,5917	0,0347	0,008	0,6223	0,5988	0,7642	0,7407
		RF	0,5459	0,5476	-0,0016	0,867	0,7767	0,7599	0,6671	0,6630
		YSA	0,7778	0,5372	0,2405	0,001	0,7843	0,5982	0,8322	0,6317
	p=25	DVM	0,6237	0,5923	0,0314	0,018	0,6166	0,5969	0,7605	0,7391
		RF	0,5610	0,5585	0,0025	0,745	0,8213	0,7946	0,6865	0,6769
		YSA	0,8731	0,5215	0,3516	0,001	0,8796	0,5945	0,9003	0,6155
	p=35	DVM	0,6456	0,5977	0,0479	0,014	0,6434	0,5999	0,7794	0,7434
		RF	0,5654	0,5547	0,1060	0,155	0,8449	0,8232	0,6973	0,6875
		YSA	0,9312	0,5188	0,4124	0,001	0,9384	0,5979	0,9435	0,5980
N=1000	p=15	DVM	0,6062	0,5971	0,0090	0,049	0,6024	0,5987	0,7508	0,7453
		RF	0,5616	0,5461	0,0154	0,001	0,8335	0,7846	0,6945	0,6746
		YSA	0,7276	0,5419	0,1857	0,001	0,8335	0,7846	0,7936	0,6472
	p=25	DVM	0,6158	0,6009	0,0148	0,007	0,6109	0,6016	0,7580	0,7495
		RF	0,5692	0,5625	0,0066	0,049	0,8751	0,8400	0,7074	0,6950
		YSA	0,8378	0,5173	0,3205	0,001	0,8397	0,5934	0,8720	0,6081
	p=35	DVM	0,6056	0,5949	0,0107	0,045	0,6019	0,5957	0,7512	0,7452
		RF	0,5857	0,5831	0,0025	0,380	0,9111	0,8881	0,7257	0,7193
		YSA	0,8670	0,5324	0,3345	0,001	0,8785	0,6135	0,8916	0,6223

[1]: Eğitim verisi, **[2]:** Test verisi; **D:** Doğruluk, ***p:*** Bağımsız değişken sayısı, Δ**D:** Doğruluk farkı ortalamaları, **p:** Tek örneklem-t test anlamlılığı, **PKD** pozitif kestirim değeri, **F:** F-ölçütü.

Tablo 8: Bağımlı değişkende riskli durumun yaklaşık % 60 olduğunda performanslar

			D[1]	D[2]	ΔD	p	PKD[1]	PKD[2]	F[1]	F[2]
N=500	p=15	DVM	,6616	,5904	,0712	,001	,9420	,3105	,5002	,2221
		RF	,5624	,5638	-,0014	,926	,2230	,2396	,2930	,2970
		YSA	,7939	,5249	,2689	,001	,7742	,3885	,7071	,3570
	p=25	DVM	,6080	,6034	,0047	,394	,9792	,4167	,1998	,2326
		RF	,5693	,5643	,0050	,480	,1660	,1726	,2261	,2311
		YSA	,9162	,5042	,4120	,001	,9167	,3851	,8993	,3696
	p=35	DVM	,6285	,6007	,0279	,070	,9804	,3900	,5390	,2980
		RF	,5632	,5620	,0012	,882	,1391	,1651	,1975	,2235
		YSA	,9394	,5249	,4145	,001	,9413	,4014	,9188	,3914
N=1000	p=15	DVM	,6004	,5892	,0112	,036	,8956	,2071	,2037	,0426
		RF	,5631	,5587	,0044	,560	,1486	,2033	,2101	,2654
		YSA	,7276	,5405	,1872	,001	,6972	,3978	,5869	,3295
	p=25	DVM	,6040	,5975	,0065	,227	,9691	,3704	,2211	,2260
		RF	,5735	,5772	-,0037	,518	,1166	,1483	,1766	,2110
		YSA	,8838	,5192	,3646	,001	,8884	,3815	,8557	,3653
	p=35	DVM	,6218	,5994	,0224	,168	,9828	,2238	,4539	,3947
		RF	,5752	,5667	,0084	,022	,1051	,1373	,1638	,2006
		YSA	,8881	,5432	,3449	,001	,8837	,4217	,8546	,4086

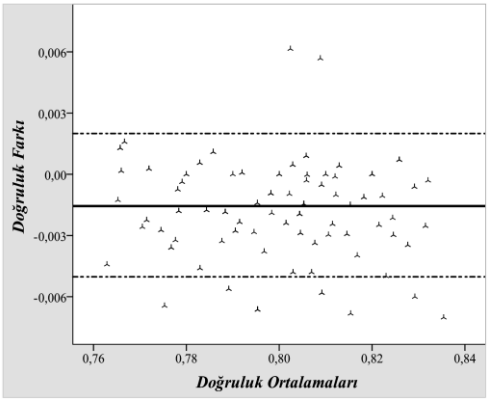
[1]: Eğitim verisi, **[2]:** Test verisi; **D:** Doğruluk, ***p:*** Bağımsız değişken sayısı, Δ**D:** Doğruluk farkı ortalamaları, **p:** Tek örneklem-t test anlamlılığı, **PKD** pozitif kestirim değeri, **F:** F-ölçütü.

Tablo 9: Bağımlı değişkende riskli durumun yaklaşık % 80 olduğunda performanslar

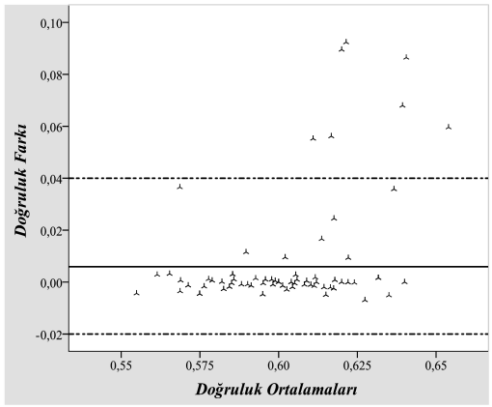
			D[1]	D[2]	ΔD	p	PKD[1]	PKD[2]	F[1]	F[2]
N=500	p=15	DVM	,8002	,8002	,0000	,985	1,0000	,0000	,3664	!
		RF	,7961	,7945	,0016	,693	,0121	,0053	,0499	,0785
		YSA	,8863	,7423	,1439	,001	,8809	,1946	,6544	,1347
	p=25	DVM	,8022	,8028	-,0006	,777	1,0000	!	,4444	!
		RF	,8004	,8012	-,0007	,431	,0009	!	,0286	!
		YSA	,9388	,7416	,1972	,001	,9830	,1894	,9082	,1368
	p=35	DVM	,8004	,8063	-,0058	,063	1,0000	!	,3902	!
		RF	,8024	,8023	,0001	,910	,0019	!	,0273	!
		YSA	,9448	,7472	,1976	,001	,9983	,2027	,9633	,1697
N=1000	p=15	DVM	,8018	,8029	-,0011	,001	1,0000	!	,0676	!
		RF	,7940	,7894	,0047	,002	,0049	,0083	,0242	,0438
		YSA	,8497	,7426	,1071	,001	,8513	,1456	,5137	,1156
	p=25	DVM	,8003	,8019	-,0016	,001	!	!	!	!
		RF	,8004	,8002	,0003	,592	,0010	,0012	,0142	,0351
		YSA	,9047	,7198	,1849	,001	,8949	,1906	,7497	,1663
	p=35	DVM	,7978	,7991	-,0012	,001	!	!	!	!
		RF	,8027	,7994	,0033	,415	!	!	,0137	!
		YSA	,9259	,7272	,1988	,001	,9659	,1982	,8857	,1697

[1]: Eğitim verisi, **[2]:** Test verisi; **D:** Doğruluk, ***p:*** Bağımsız değişken sayısı, Δ**D:** Doğruluk farkı ortalamaları, **p:** Tek örneklem-t test anlamlılığı, **PKD** pozitif kestirim değeri, **F:** F-ölçütü, **!:** Hesaplanamayan değer.

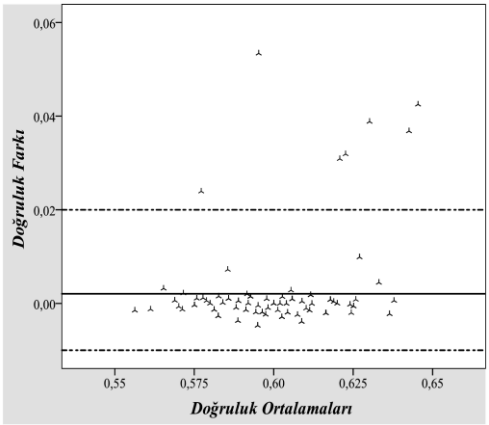
Simüle edilmiş verilerde makine öğrenimi yöntemlerinin her bir risk oranında değişken ve gözlem sayısı önemsenmeden eğitim ve test doğrulukları için Bland-Altman grafikleri çizdirilerek doğruluklar arasındaki uyum değerlendirildi. Bu grafiklerde de DVM yöntemi eğitim ve test doğrulukları özellikle ilgilenilen riskin bağımlı değişkende görülme oranı arttıkça grafiklerdeki uyum göze çarpmaktadır. RF yönteminde ise bağımlı değişkendeki gözlemlerin bir sınıfta toplandığı durumlarda eğitim ve test seti doğruluklarının uyumu ön plana çıkmaktadır. Ancak YSA için öne çıkan bir durum saptanamamıştır.



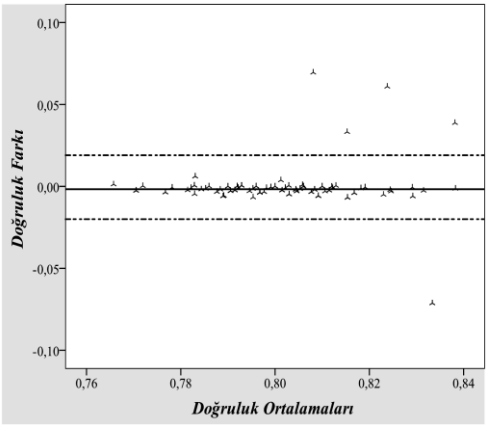
Şekil 22.a. 0,2 Prevalans için DVM uyum



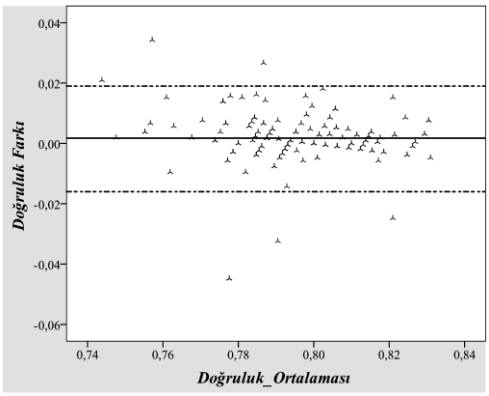
Şekil 22.b. 0,4 Prevalans için DVM uyum



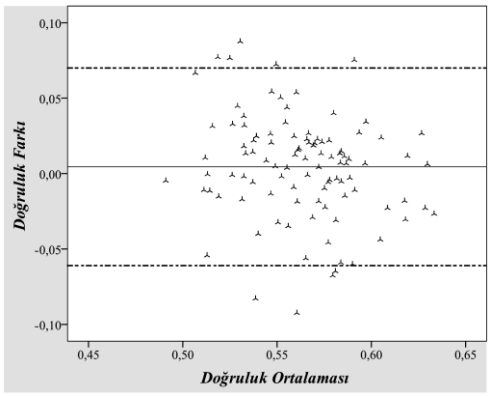
Şekil 22.c. 0,6 Prevalans için DVM uyum



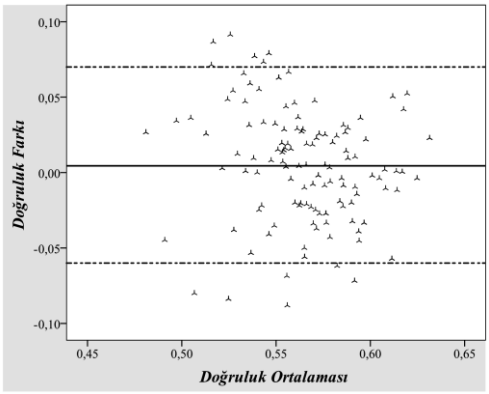
Şekil 22.d. 0,8 Prevalans için DVM uyum



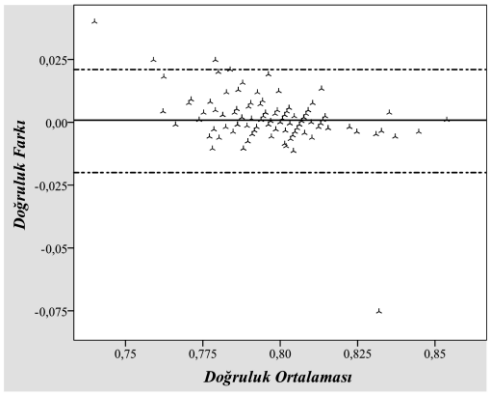
Şekil 23.a. 0,2 Prevalans için RF uyum



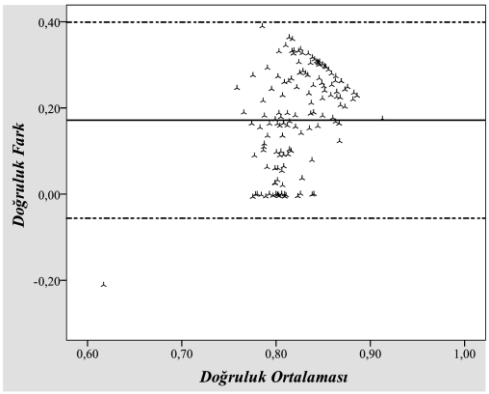
Şekil 23.b. 0,4 Prevalans için RF uyum



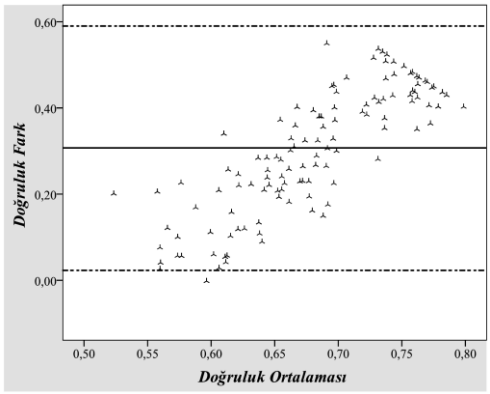
Şekil 23.c. 0,6 Prevalans için RF uyum



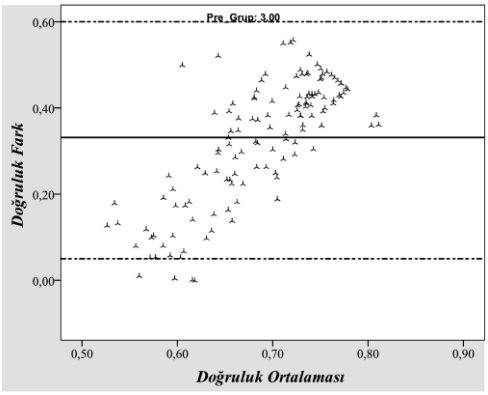
Şekil 23.d. 0,8 Prevalans için RF uyum



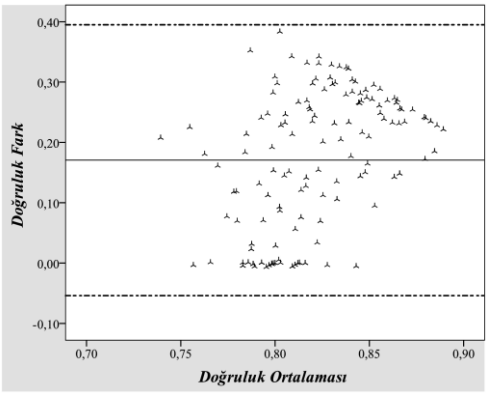
Şekil 24.a. 0,2 Prevalans için YSA uyum



Şekil 24.b. 0,4 Prevalans için YSA uyum



Şekil 24.c. 0,6 Prevalans için YSA uyum



Şekil 24.d. 0,8 Prevalans için YSA uyum

6. TARTIŞMA

Tez çalışmasındaki veri setleri 12 alt birime ayrılmış üç adet makine öğrenimi yöntemi, gerçek veri seti için yedi performans değerlendirme kriteri ve simülasyon verileri için beş performans değerlendirme kriteri ile karşılaştırıldı.

Gerçek veri setinde test verileri için en yüksek sınıflama doğruluğu radyal tabanlı DVM 'den (% 80) elde edildi. Simülasyon verilerinde genel olarak en yüksek sınıflama doğruluğu DVM 'den elde edilirken RF yönteminin de başarılı sonuçlar verdiği görüldü. Bağımlı değişkende riskli durumun dağılımının dengesiz olduğu değişken sayısının 15-25 olduğu durumlarda DVM yönteminin sınıflamada oldukça başarılı olduğu değişken sayısı ile birlikte gözlem sayısı da arttığında RF yönteminin başarılı sonuçlar verdiği görülmektedir. Bağımlı değişkendeki riskli durum yüzdesinin çok farklı olmadığı durumlarda da DVM yönteminin performansı tez çalışmasındaki diğer yöntemlere göre yüksek bulundu.

Çalışmadaki simülasyon verilerinden eğitim ve test verileri için doğruluklar hesaplanırken alınan sonuçlara göre bir hususun ön plana çıktığı görüldü. Özellikle bağımlı değişkende riskli durum oranının çok düşük veya çok yüksek olduğu durumlarda eğitim ve test doğrulukları birbirine yakın olan makine öğrenimi yönteminin genel olarak daha performanslı sonuçlar verdiği anlaşıldı. Eğitim ve test doğruluk değeri birbirine uzak olan yöntemin performansının da düşük olabileceği sonucuna varıldı. Bu sonuca, taranan çalışmalarda rastlanılmadığı görüldü. Ancak gerçek veri setinde bu durumun paralellik göstermemesi, veri seti için değişken seçim (feature selection) işleminin yapılmamasından kaynaklandığı düşünülmektedir. Gerçek veri setinde değişken seçim işleminin yapılmamasının nedeni, ML 'de daha çok değişken ile performanslı sonuç alabilme öngörüsüdür. Meme kanserli hastalardan oluşan veriler için literatürdeki çalışmalara bakıldığında IDC Evre III için sınıflama veya sağkalım çalışması bulunamamıştır. Meme kanserine genel yaklaşım olarak;

Kate R. J. ve Nadig R.'nin 2017 yılında yapmış oldukları "*Stage-specific predictive models for breast cancer survivability*" isimli çalışmalarında sağkalım sınıflamasında 16 bağımsız değişken kullanmışlardır (46). Söz konusu çalışma için

makine öğrenimi yöntemlerinden Naive Bayes, Lojistik Regresyon ve Karar Ağaçları metotlarını kullanıp sonuçları AUC kriterine göre değerlendirmişlerdir. En başarılı performansın Naive Bayes yönteminden alındığını belirtmişlerdir.

Chao C-M. ve arkadaşlarının 2014 yılında yapmış oldukları "*Construction the Model on the Breast Cancer Survival Analysis Use Support Vector Machine, Logistic Regression and Decision Tree*" isimli çalışmalarında meme kanseri için sağkalım sınıflamasında 7 bağımsız değişken kullanmışlardır (47). Bu çalışmada makine öğrenimi yöntemlerinden DVM, Lojistik Regresyon ve Karar Ağaçları metotlarını kullanıp sonuçları eğitim ve test seti doğruluğu şeklinde sunmuşlardır. Çalışma sonucuna göre DVM yüksek sınıflama doğruluğu göstermiştir.

Tapak L. ve arkadaşlarının 2018 yılında yapmış oldukları "*Prediction of survival and metastasis in breast cancer patients using machine learning classifiers*" isimli çalışmalarında meme kanserinde sağkalım ve metastaz için yapılan sınıflamada 9 bağımsız değişken kullanmışlardır (48). Bu çalışmada makine öğrenimi yöntemlerinden DVM, RF, YSA, Naive Bayes, Liner Diskirminant Analizi (LDA), Adaboost ve Adabag metotlarını kullanıp sonuçları doğruluk, likelihood ratio, duyarlılık ve seçicilik kriterlerine göre değerlendirmişlerdir. Belirlenen kriterlere en yüksek performansı DVM ve LDA metodunun gösterdiğini bildirmişlerdir.

Meme kanseri türlerinden sadece IDC 'li hastaların veri seti olarak kullanıldığı "*Predictive value of CD24 and CD44 for neoadjuvant chemotherapy response and prognosis in primary breast cancer patients*" sağkalım çalışmasında makine öğrenimi tekniklerinden Lojistik Regresyon ve Alternatif Karar Ağaçları (ADTree) kullanılmıştır (49). Yöntemler doğruluk ve AUC kriterlerine göre kıyaslanmıştır.

Simülasyon ile veri üretilip makine öğrenimi yöntemlerinin sınıflama performansının kıyaslandığı literatür çalışmalarında;

Khondoker M. ve arkadaşlarının 2018 yılında yapmış oldukları "*A comparison of machine learning methods for classification using simulation with multiple real data examples from mental health studies*" isimli çalışmalarında hem gerçek veri seti üzerinden simülasyonlarla hem de normal olmayan (poisson dağılımından) tamamen simüle edilmiş veri setiyle RF, DVM, LDA ve k-En Yakın

Komşuluk (kNN) yöntemlerini; sınıflama hatası, duyarlılık ve seçicilik kriterlerine göre kıyaslamışlarıdır (50). Genel sonuç olarak *p* değişken sayısını ve *n* örneklem hacmini göstermek üzere yüksek korelasyonlu veri setinde $p/n < 0,5$ *olduğunda* LDA 'yı $p/n \geq 0,5$ olduğunda korelasyona değinmeden radyal tabanlı DVM 'yi önermişlerdir. Söz konusu dört yöntemden RF hariç diğer yöntemlerin yüksek korelasyonda daha iyi performans gösterdiğini bildirmişlerdir.

Engelhardt A. arkadaşlarının 2014 yılında yapmış oldukları "*Comparing classification methods for diffuse reflectance spectra to improve tissue specific laser surgery*" isimli çalışmalarında küçük bir veri kümesini simüle ederek çok değişkenli normal dağılım gösteren daha büyük veri seti elde etmişlerdir (51). Bu veri setlerini LDA, Quadratic Liner Diskirminant Analizi (QDA), Penalized Diskirminant Analizi (PDA), RF, CART, YSA ve kNN gibi ML yöntemleriyle yanlış sınıflama oranları ve Student-t testi ile kıyaslamışlardır. En düşük yanlış sınıflama oranının PDA 'da olduğu sonucuna ulaşılmıştır.

7. SONUÇ

Sınıflamanın yapıldığı birçok veri setini içeren tez çalışmasında literatürle uyumlu olarak DVM 'nin, çalışmada kullanılan diğer yöntemlere göre (RF ve YSA) daha yüksek sınıflama doğruluğu ve performans kriteri değerine ulaştığı sonucuna varılmıştır. Simülasyon çalışmalarında da yine DVM test doğruluğunun en yüksek değerlere ulaştığı gözlenmiştir.

RF yöntemi işlem süresinin kısalığı ve sınıflama doğruluğu bakımından oldukça başarılıdır. Özellikle bağımlı değişkende risk görülme oranının homojen dağılmadığı durumlarda değişken ve gözlem sayısı arttıkça sınıflama doğruluğu yükselmektedir.

YSA 'da hiperparametrelerin belirlenme ve modelleme süresi diğer yöntemlere göre oldukça zaman alıcıdır. Aynı zamanda sınıflama doğruluğu bakımından kötü performans göstermektedir. Dolayısıyla makine öğrenimi çalışmaları için araştırıcılara, hem farklı değişken sayısı ve örneklem hacmi için performanslı sonuçlar verdiğinden hem de hiperparametreleri belirleme işlemi süresinin kısalığından dolayı DVM yöntemi önerilmektedir.

Tez çalışmasında simülasyon verilerinden eğitim ve test setleri için elde edilen sınıflama doğrulukları incelendiğinde, sınıflama doğruluğu birbirine yakın olan yöntemlerden daha performanslı sonuç alındığı görülmektedir. Ancak bu durumun simülasyon verisinin bütün kombinasyonları için geçerli olmadığı anlaşılmaktadır. Dolayısıyla bu alanda daha fazla çalışma yapılmasına, farklı yapıdaki veri setleri üzerinde ML teknikleri uygulanarak eğitim ve test doğruluğu farkının analiz çalışmalarına ihtiyaç olduğu düşünülmektedir.

8. KAYNAKLAR

1. Aydoğan Ü. Destek Vektör Makinalarında Kullanılan Çekirdek Fonksiyonların Sınıflama Performanslarının Karşılaştırılması. Hacettepe Üniversitesi, Sağlık Bilimleri Enstitüsü, Yüksek Lisans Tezi, 2010, Ankara (Prof. Dr. Osman Saraçbaşı Doç. Dr. Hafize Sezer).

2. Han J, Kamber M and Pei J. Foreword. In: Data Mining Concepts and Techniques. 3st ed. Waltham: Morgan Kaufmann (Elseiver); 2012, p:xix, 327, 24, 25, 408.

3. Smola A, Vishwanathan SVN. Preface. In: Introduction to Machine Learning. 2st ed. Cambridge: Cambridge University Press; 2010, p:5, 6.

4. Larose DT. Introduction to Data Mining. In: Discovering Knowledge in Data. 1st ed. New Jersey: Wiley-Interscience; 2005, p:14.

5. Hamel LH. What is Knowledge Dıscovery. In: Knowledge Discovery with Support Vector Machines. 1st ed. New Jersey: Wiley-Interscience; 2009, p:6.

6. Dietterich TG. An experimental comparison of three methods for constructing ensembles of decision trees: Bagging, boosting, and randomization. Machine Learning. 2000; 40(2):139–157.

7. Tolun S. Destek Vektör Makineleri: Banka Başarısızlığının Tahmini Üzerine Bir Uygulama. İstanbul Üniversitesi, Sosyal Bilimler Enstitüsü, Doktora Tezi, 2008, İstanbul (Prof. Dr. Öner ESEN).

8. Kaygulu MS. Supervised and Unsupervised Learning Techniques in Data Mining. Dokuz Eylül Üniversitesi, Fen bilimleri Enstitüsü, Yüksek Lisans Tezi, 1999, İzmir (Doç. Dr. Alp KUT).

9. Mohri M, Rostamizadeh A, Talwalkar A. Introduction. In: Foundations of Machine Learning. 2st ed. London: The MIT Press; 2012, p:7,8.

10. Domingos P. A Few Useful Things to Know about Machine Learning. Communications of the ACM. 2012; 55(10):78-87.

11. Harman G, Kulkarni S. Statistical Learning Theory. In: Statistical Learning Theory as a Framework for the Philosophy of Induction. 3st ed. New Jersey: Princeton University The MIT Press; 2008, p:1-4.

12. Hurtado JE. An Examination of Methods for Approximating Implicit from the Viewpoin Statistical Learning Theory. Structural Safety. 2004; 26(3):271-293.

13. Vapnik VN. An Overview of Statistical Learning Theory. IEEE Transactions on Neural Networks. 1999; 10(5):988-999.

14. Kıyak E. CRISP-DM Yöntembilimi Kullanılarak Deniz Kuvvetleri Verisi Üzerinde Veri Madenciliği Sınıflandırma Tekniklerinin Karşılaştırılması. Kocaeli Üniversitesi, Fen Bilimleri Enstitüsü, Yüksek Lisans Tezi, 2006, Kocaeli (Yrd. Doç. Dr. Nevcihan DURU).

15. Avşar E. Tek-Sınıf Destek Vektör Makineleri Kullanılarak Epilektik EEG İşaretlerinin Sınıflandırılması. İstanbul Üniversitesi, Fen Bilimleri Enstitüsü, Yüksek Lisans Tezi, 2009, İstanbul (Yrd. Doç. Dr. Mustafa E. KAMASAK).

16. Guyon I, Weston J, Barnhıll S, Vapnik V. Gene Selection for Cancer Classification using Support Vector Machine. Machine Learning. 2002; 46:389–422.

17. Abe S. Introduction. In: Support Vector Machines for Pattern Classification. 2st ed. Loughborough: Springer-Verlag London Limited; 2010, p:1, 2, 24, 25, 32, 33.

18. Burges, CJC. A Bound on the Generalization Performance of a Pattern Recognition Learning Machine. In: A Tutorial on Support Vector Machines for Pattern Recognation. 1st ed. Netherlands: Kluwer Academic Publishers; 1998, p:2, 121, 167.

19. Çomak E. Destek Vektör Makineleri Çoklu Sınıf Problemleri İçin Çözüm Önerileri. Selçuk Üniversitesi, Fen Bilimleri Enstitüsü, Yüksek Lisans Tezi, 2004, Konya (Doç. Dr. Ahmet ARSLAN).

20. Cox PG, Adhami R. Multi-Class Support Vector Machine Classifier Applied to Hyper-Spectral Data. Proceedings of the Thirty-Fourth Southeastern Symposium on System Theory. Piscataway: IEEE, 2002; 271-274.

21. Orsenigo C, Vercellis C. Discrete Support Vector Decision Trees via Tabu Search. Computational Statistics & Data Analysis. 2004; 47(2):311–322.

22. Zhou ZH. Ensembles Methods. In: Ensemble Methods Foundations and Algorithms. 1st ed. London: CRC Press; 2012, p:15.

23. Sutton CD. Classification and Regression Trees, Bagging and Boosting. Handbook of Statistics, Elsevier. 2005; 24:303-329

24. Breiman L. Random Forests. Machine Learning. 2001; 45(1):5–32.

25. Silahtaroğlu G. An Attribute-Centre Based Decision Tree Classification Algorithm. World Academy of Science, Engineering and Technology. 2009; 56:302-306.

26. Akman M. Veri Madenciliğine Genel Bakış ve Random Forests Yönteminin İncelenmesi: Sağlık Alanında Bir Uygulama. Ankara Üniversitesi, Sağlık Bilimleri Enstitüsü, Yüksek Lisans Tezi, 2010, Ankara (Doç. Dr. Yasemin GENÇ, Doç. Dr. Handan ANKARALI).

27. Palmer D, O 'Boyle N, Glen R, Mitchell J. Random Forest Models To Predict Aqueous Solubility. Journal of Chemical Information and Modeling 2006; 47(1):150-8.

28. Liaw A. Classification and Regression by Random Forest. R News. 2002; 2(3):18-22.

29. Segal M R. Machine Learning Benchmarks and Random Forest Regression. eScholarship: Center for Bioinformatics and Molecular Biostatistics. 2004; 14:1-14.

30. Korkem E. Mikroarray Gen Ekspresyon Veri Setlerinde Random Forest ve Naıve Bayes Sınıflama Yöntemleri Yaklaşımı. Hacettepe Üniversitesi, Sağlık Bilimleri Enstitüsü, Yüksek Lisans Tezi, 2013, Ankara (Doç. Dr. Erdem KARABULUT).

31. Haykin S. İntroduction. In: Neural Networks and Learning Machines. 2 st ed. Hamilton: Pearson Prentice Hall; 2009, p:2.

32. Kriesel D. Biological Neural Networks. In: Neural Networks. 1st ed. Bonn: ZETA2-EN; 2005, p:13.

33. Koç ML, Balas CE, Arslan A. Taş Dolgu Dalgakıranların Yapay Sinir Ağları ile Ön Tasarımı. İnşaat Mühendisleri Odası Yayını Teknik Dergi. 2004; 15(4): 3354.

34. Bangal CB. Automatıc Generatıon Control of Interconnected Power Systems Using Artificial Neural Network Techniques. Bharath Üniversitesi, Fen Bilimleri Enstitüsü, Doktora Tezi, 2009, Chennai (Dr. Jayant G. Ghodekar).

35. Karasu F. Petrokimya Sektöründe Talep Tahmininde Yapay Sinir Ağlarının Kullanılması " Petkim A.Ş. Örneği ". Dokuz Eylül Üniversitesi, Sosyal Bilimler Enstitüsü, Yüksek Lisans Tezi, 2012, İzmir (Danışman: Doç. Dr. Hilmi YÜKSEL).

36. Özün A. Yapay Sinir Ağları ile Risk Öngörüsü. Marmara Üniversitesi, Sosyal Bilimler Enstitüsü, Doktora Tezi, 2011, İstanbul (Danışman: Prof. Dr. Özhan TINGÖY).

37. Zhang G, Patuwo BE, Hu MY. Forecasting with artificial neural networks: The state of the art. International Journal of Forecasting. 1998; 14:35–62.

38. Vhinen M. How to evaluate performance of prediction methods? Measures and their interpretation in variation effect analysis. BMC Genomics. 2012; 13(4):2-5.

39. Akosa JS. Predictive Accuracy: A Misleading Performance Measure for Highly Imbalanced Data Classified Negative. In Proceedings of the SAS Global Forum, Orlando, FL, USA, 2–5 April 2017.

40. Cianfrocca M, Goldstein LJ. Prognostic and Predictive Factors in Early-Stage Breast Cancer. The Oncologist. 2004; 9(6):606-616.

41. Canda T. Meme Kanserinde Prognostik Faktörler. Türkiye Ekopatoloji Dergisi. 1995; 1 (1-2): 13-23.

42. Evcimik T. Meme Kanserinde Prognostik Faktörlerin Sağkalıma Etkisi. İstanbul Göztepe Eğitim ve Araştırma Hastanesi 3.Genel Cerrahi Kliniği, Uzmanlık Tezi, 2008, İstanbul (Doç. Dr. M. Rafet YİĞİTBAŞI)

43. Woodward WA, Vinh-Hung V, Ueno NT, Cheng YC, Royce M, Tai P, et al. Prognostic value of nodal ratios in node-positive breast cancer. J Clin Oncol. 2006; 24:2910-6.

44. Cornwell LB, McMasters KM, Chagpar AB. The impact of lymphovascular invasion on lymph node status in patients with breast cancer. Am Surg. 2011; 77:874–877.

45. Jones RL, Salter J, A'Hern R, et al. The Prognostic Significance of Ki-67 Before and After Neoadjuvant Chemotherapy in Breast Cancer. Breast Cancer Research and Treatment. 2009; 116(1):53–68.

46. Kate RJ, Nadig R. Stage-Specific Predictive Models for Breast Cancer Survivability. International Journal of Medical Informatics. 2017; 97(1):304–311.

47. Chao CM, Yu YW, Cheng BW and Kuo YL. Construction the Model on the Breast Cancer Survival Analysis Use Support Vector Machine, Logistic Regression and Decision Tree. Journal of Medical Systems. 2014; 38(10):106-113.

48. Tapak L, Khorram NS, Amini P, Alafchi B, Hamidi O, Poorolajal J. Prediction of Survival and Metastasis in Breast Cancer Patients Using Machine Learning Classifiers. Clinical Epidemiology and Global Health. 2018; (In Press Corrected Proof).

49. Predictive Value of CD24 and CD44 for Neoadjuvant Chemotherapy Response and Prognosis in Primary Breast Cancer Patients. Journal of Medical and Dental Sciences. 2010; 57(2):165-175.

50. Khondoker M, Dobson R, Skirrow C, Simmons A and Stahl D. A Comparison of Machine Learning Methods for Classification Using Simulation with Multiple Real Data Examples from Mental Health Studies. Statistical Methods in Medical Research. 2016; 25(5):1804–1823.

51. Engelhardt A, Kanawade R, Knipfer C, Schmid M, Stelzle F and Adler W. Comparing Classification Methods for Diffuse Reflectance Spectra to Improve Tissue Specific Laser Surgery. BMC Medical Research Methodology. 2014; 14(1):1-15.

9. ÖZGEÇMİŞ

Adı Soyadı: Emre DİRİCAN

Doğum yeri ve tarihi: Malatya 1985

Uyruğu: T.C.

Medeni hali: Evli

Ünvanı: Araştırma Görevlisi

Yabancı dili: İngilizce

Çalıştığı Kurum: Hatay Mustafa Kemal Üniversitesi Tıp Fak. Biyoistatistik AD

Telefon: 0531 5057799

E-mail: emredirilbey@gmail.com

Adres: Hatay Mustafa Kemal Üniversitesi Tıp Fakültesi Temel Tıp Bilimleri Biyoistatistik AD / HATAY

Tlf: (0326) 229 10 00-17112

Görevler:

Görev Ünvanı	Görev Yeri	Yıl
Y.Lisans	İnönü Üniversitesi Tıp Fakültesi Biyoistatistik AD	2009-2012
Doktora	Dicle Üniversitesi Tıp Fakültesi Biyoistatistik AD	2013-…
Arş. Gör.	Mustafa Kemal Üniversitesi Tayfur Ata Sökmen Tıp Fakültesi Biyoistatistik AD	2015
Arş. Gör.	Dicle Üniversitesi Tıp Fakültesi Biyoistatistik AD	2015-2017
Arş. Gör.	Mustafa Kemal Üniversitesi Tayfur Ata Sökmen Tıp Fakültesi Biyoistatistik AD	2017-…

10. ORJİNALLİK RAPORU

Emre Dirican Doktora tez

ORIJINALLIK RAPORU

%4	%2	%1	%1
BENZERLIK ENDEKSI	İNTERNET KAYNAKLARI	YAYINLAR	ÖĞRENCI ÖDEVLERI

BİRİNCİL KAYNAKLAR

1	openaccess.ogu.edu.tr:8080 İnternet Kaynağı	<%1
2	studylibtr.com İnternet Kaynağı	<%1
3	Submitted to Konya Necmettin Erbakan University Öğrenci Ödevi	<%1
4	www.labtestsonline.org.tr İnternet Kaynağı	<%1
5	kutuphane.ksu.edu.tr İnternet Kaynağı	<%1
6	red.uao.edu.co İnternet Kaynağı	<%1
7	Submitted to TechKnowledge Turkey Öğrenci Ödevi	<%1
8	linknovate.com İnternet Kaynağı	<%1

9	Submitted to Bursa Teknik Āniversitesi Öğrenci Ödevi	<%1
10	www.tmd.ac.jp İnternet Kaynağı	<%1
11	www.coursehero.com İnternet Kaynağı	<%1
12	Submitted to Düzce Üniversitesi Öğrenci Ödevi	<%1
13	Submitted to Erciyes Āniversitesi Öğrenci Ödevi	<%1
14	nedir.com İnternet Kaynağı	<%1
15	Submitted to Liverpool John Moores University Öğrenci Ödevi	<%1
16	acikerisim.selcuk.edu.tr:8080 İnternet Kaynağı	<%1
17	Leili Tapak, Nasrin Shirmohammadi-Khorram, Payam Amini, Behnaz Alafchi, Omid Hamidi, Jalal Poorolajal. "Prediction of survival and metastasis in breast cancer patients using machine learning classifiers", Clinical Epidemiology and Global Health, 2018 Yayın	<%1
18	AKÇAY, Aytaç and DEMİREL, Ayşe.	<%1

	"Pyometralı köpeklerde bazı kan parametrelerinin optimal pozitiflik eşiğinin özgün oranlar ve ROC (receiver operating characteristic) eğrisi yöntemi ile belirlenmesi", TUBITAK, 2011. Yayın	
19	insis.vse.cz İnternet Kaynağı	<%1
20	Submitted to University of Witwatersrand Öğrenci Ödevi	<%1
21	www.turkiyeklinikleri.com İnternet Kaynağı	<%1
22	YILMAZ, Emel, ÇEÇEN, Dilek, TOĞAÇ, Hülya KIZIL, MUTLU, Senan, KARA, Havva and ASLAN, Arzu. "Ameliyat Sürecindeki Hastaların Konfor Düzeyleri ve Hemşirelik Bakımları", Manisa Celal Bayar Üniversitesi Sağlık Bilimleri Enstitüsü, 2018. Yayın	<%1
23	Submitted to Fırat Üniversitesi Öğrenci Ödevi	<%1
24	www.researchgate.net İnternet Kaynağı	<%1
25	sablon.sdu.edu.tr İnternet Kaynağı	<%1

26	eprints.sdu.edu.tr İnternet Kaynağı	<%1
27	Metin Yildiz, Muhammet Yorulmaz. "Eye gaze location detection based on iris tracking with web camera", 2018 26th Signal Processing and Communications Applications Conference (SIU), 2018 Yayın	<%1
28	Ugo Bruzzo, Mattia Pedrini, Francesco Sala, Richard J. Szabo. "Framed sheaves on root stacks and supersymmetric gauge theories on ALE spaces", Advances in Mathematics, 2016 Yayın	<%1
29	www.ibrahimcayiroglu.com İnternet Kaynağı	<%1
30	okul.selyam.net İnternet Kaynağı	<%1
31	K.D. Paulsen, S.P. Poplack, Dun Li, M.W. Fanning, P.M. Meaney. "A clinical prototype for active microwave imaging of the breast", IEEE Transactions on Microwave Theory and Techniques, 2000 Yayın	<%1
32	OZDEMIR, Mehtap, BAKAN, Nurten, SAHIN, Omer Torun, KURTCELEBI, Nevın, ERBESLER, Zeynel Abıdın and TUNCA, Savas	<%1

	Taner. "The Comparison of Sevoflurane-Remifentanyl and Propofol-Remifentanyl in Robotic Prostatectomies", Derman Tıbbi Yayıncılık, 2013. Yayın	
33	www.scribd.com İnternet Kaynağı	<%1
34	tez.sdu.edu.tr İnternet Kaynağı	<%1
35	Valanarasu T.. "Asymptotic initial-value method for a system of singularly perturbed second-order ordinary differential equations of convection-diffusion type", International Journal of Computer Mathematics, 11/1/2004 Yayın	<%1
36	Apreutesei, N.. "High regularity of the solutions of the telegraph system subjected to nonlinear boundary conditions", Nonlinear Analysis, 20051115 Yayın	<%1
37	Pelin Yildirim, Kokten Ulas Birant, Vladimir Radevski, Alp Kut, Derya Birant. "Comparative analysis of ensemble learning methods for signal classification", 2018 26th Signal Processing and Communications Applications Conference (SIU), 2018 Yayın	<%1

38	M. Aydin, E. Celik. "Assamese character recognition with Artificial Neural Networks", 2013 21st Signal Processing and Communications Applications Conference (SIU), 2013 Yayın	<%1
39	Kyung Ha Seok, Jooyong Shim, Daehyeon Cho, Gyu-Jeong Noh, Changha Hwang. "Semiparametric mixed-effect least squares support vector machine for analyzing pharmacokinetic and pharmacodynamic data", Neurocomputing, 2011 Yayın	<%1
40	FRANÇOIS GALLAIRE. "Three-dimensional transverse instabilities in detached boundary layers", Journal of Fluid Mechanics, 01/2007 Yayın	<%1

11. EKLER

Ek-1: Doktora tez savunma sınavına girebilmek için gerekli makalenin bilgileri aşağıda verilmiştir.

Dicle Tıp Dergisi / Dicle Medical Journal (2017) 44 (1) : 81 – 89 (DOI:10.5798/dicletip.298613)

Toplam Kolesterol, Ldl, Hdl, Trigliserit Seviyelerinin Yaşa göre Değişiminin Farklı Regresyon Modelleriyle İncelenmesi

Emre Dirican[1], Cemil Çolak[2], Zeki Akkuş[1]

[1] Dicle Üniversitesi, Tıp Fakültesi, Biyoistatistik Anabilim Dalı, Diyarbakir-Türkiye

[2] İnönü Üniversitesi Tıp Fakültesi Biyoistatistik Anabilim Dalı, Malatya, Türkiye

Geliş: 09.11.2016; Revizyon: 31.01.2017; Kabul Tarihi: 07.02.2017

Özet

Amaç: Bu çalışmada hiperlipidemi hastalarının Toplam Kolesterol, LDL, Trigliserit, HDL seviyelerinin farklı regresyon modelleriyle tahmini ve yaşa göre değişiminin belirlenmesi amaçlanmıştır. Bu amaçla, doğrusal ve doğrusal olmayan regresyon modelleri ile analiz yapılmıştır.

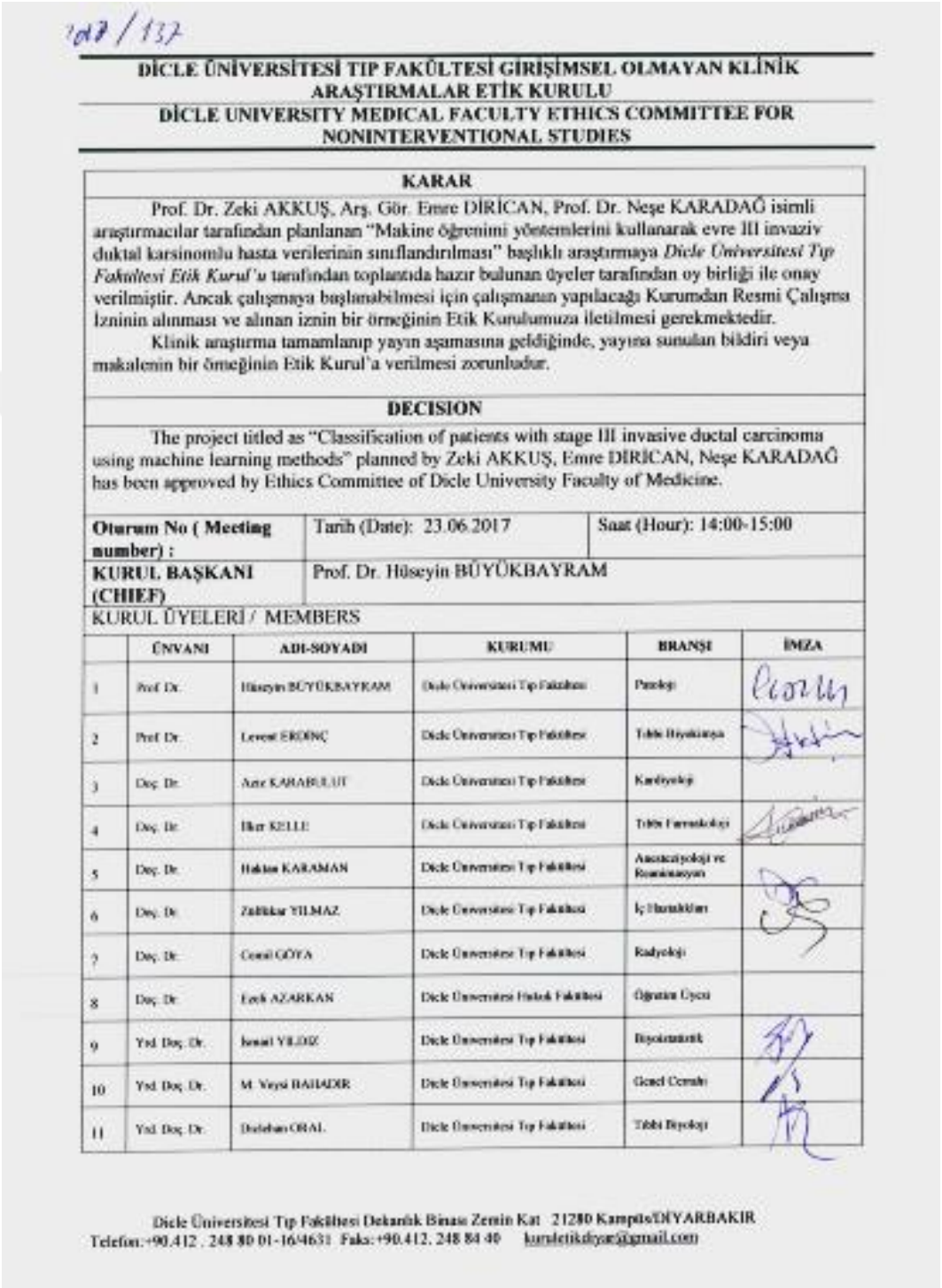
Yöntemler: Çalışmamız, İnönü Üniversitesi Turgut Özal Tıp Merkezi Kardiyoloji bölümüne müracaat eden 1278 hiperlipidemili hasta üzerinde gerçekleştirilmiştir. Retrospektif çalışmamızda veri toplama aracı olarak hasta dosyaları ve vaka kayıtları kullanılmıştır. Kayıtlardan toplanan veriler paket programlarla analiz edilerek modellemeler yapılmıştır.

Bulgular: Araştırmamızda 1278 hiperlipidemili hastanın 671'i (%52,4) erkek ve 607'si (%47,6) kadındır. Büyüme eğrilerinin uyumunun incelenmesinde açıklayıcılık katsayısı (R^2), Hata Kareler Ortalaması (HKO), Akaike Bilgi Kriteri (AIC) ve Schwarz Bilgi Kriteri (SIC) hesapları dikkate alındığında, kadın hastalarda; Kolesterol değerleri için en iyi büyüme modeli Kuadratik model, LDL için Gompertz model, Trigliserit için Lojistik model, HDL için Ustel modeldir. Erkek hastalarda ise; Kolesterol değerleri için en iyi büyüme modeli Gompertz model, LDL için Doğrusal model, Trigliserit için Üstel model, HDL için Doğrusal modeldir.

Sonuç: Doğrusal olmayan hiperlipidemik değerlerin tahmininde değişik modellerin farklı performans gösterdiği belirlenmiştir. Hiperlipidemi değerleri izleminde bu modellerden elde edilen büyüme eğrilerinin kullanılmasının, Kolesterol, LDL, HDL ve Trigliserit değerlerinin tahmin edilebilmesinde ayrıca normal değerlerden sapmaların değerlendirilmesinde yararlı olacağı düşünülmektedir.

Anahtar kelimeler: Doğrusal regresyon, doğrusal olmayan regresyon, uyum iyiliği kriterleri, en küçük kareler yöntemi, en çok olabilirlik yöntemi.

Ek-2: Etik Kurul Onayı



2017 / 137

DİCLE ÜNİVERSİTESİ TIP FAKÜLTESİ GİRİŞİMSEL OLMAYAN KLİNİK ARAŞTIRMALAR ETİK KURULU

DİCLE UNIVERSITY MEDICAL FACULTY ETHICS COMMITTEE FOR NONINTERVENTIONAL STUDIES

KARAR

Prof. Dr. Zeki AKKUŞ, Arş. Gör. Emre DİRİCAN, Prof. Dr. Neşe KARADAĞ isimli araştırmacılar tarafından planlanan "Makine öğrenimi yöntemlerini kullanarak evre III invaziv duktal karsinomlu hasta verilerinin sınıflandırılması" başlıklı araştırmaya *Dicle Üniversitesi Tıp Fakültesi Etik Kurul'u* tarafından toplantıda hazır bulunan üyeler tarafından oy birliği ile onay verilmiştir. Ancak çalışmaya başlanabilmesi için çalışmanın yapılacağı Kurumdan Resmi Çalışma İzninin alınması ve alınan iznin bir örneğinin Etik Kurulumuza iletilmesi gerekmektedir.

Klinik araştırma tamamlanıp yayın aşamasına geldiğinde, yayına sunulan bildiri veya makalenin bir örneğinin Etik Kurul'a verilmesi zorunludur.

DECISION

The project titled as "Classification of patients with stage III invasive ductal carcinoma using machine learning methods" planned by Zeki AKKUŞ, Emre DİRİCAN, Neşe KARADAĞ has been approved by Ethics Committee of Dicle University Faculty of Medicine.

Oturum No (Meeting number) :	Tarih (Date): 23.06.2017	Saat (Hour): 14:00-15:00
KURUL BAŞKANI (CHIEF)	Prof. Dr. Hüseyin BÜYÜKBAYRAM	

KURUL ÜYELERİ / MEMBERS

	ÜNVANI	ADI-SOYADI	KURUMU	BRANŞI	İMZA
1	Prof. Dr.	Hüseyin BÜYÜKBAYRAM	Dicle Üniversitesi Tıp Fakültesi	Patoloji	
2	Prof. Dr.	Levent ERDİNÇ	Dicle Üniversitesi Tıp Fakültesi	Tıbbi Biyokimya	
3	Doç. Dr.	Azer KARABULUT	Dicle Üniversitesi Tıp Fakültesi	Kardiyoloji	
4	Doç. Dr.	İlker KELLİ	Dicle Üniversitesi Tıp Fakültesi	Tıbbi Farmakoloji	
5	Doç. Dr.	Hakkı KARAMAN	Dicle Üniversitesi Tıp Fakültesi	Anesteziyoloji ve Reanimasyon	
6	Doç. Dr.	Zülfikar YILMAZ	Dicle Üniversitesi Tıp Fakültesi	İç Hastalıkları	
7	Doç. Dr.	Cemil GÖYA	Dicle Üniversitesi Tıp Fakültesi	Radyoloji	
8	Doç. Dr.	Ezel AZARKAN	Dicle Üniversitesi Hukuk Fakültesi	Öğretim Üyesi	
9	Yrd. Doç. Dr.	İsmail YILDIZ	Dicle Üniversitesi Tıp Fakültesi	Biyoistatistik	
10	Yrd. Doç. Dr.	M. Veysi BAHADIR	Dicle Üniversitesi Tıp Fakültesi	Genel Cerrahi	
11	Yrd. Doç. Dr.	Dolunay ORAL	Dicle Üniversitesi Tıp Fakültesi	Tıbbi Biyoloji	

Dicle Üniversitesi Tıp Fakültesi Dekanlık Binası Zemin Kat 21280 Kampüs/DİYARBAKIR
Telefon: +90.412 . 248 80 01-1634631 Faks: +90.412. 248 84 40 kurultikdiyar@gmail.com

Ek-3: Tez çalışmasında kullanılan bir kısım R kodları.

Gerçek veri seti data1 ve y bağımlı değişken olarak ifade edilmek üzere;

- *DVM için doğrusal çekirdek için,*

```
set.seed(3003)
intrain <- createDataPartition(y = data1$y, p= 0.7, list = FALSE)
training <- data1[intrain,]
testing <- data1[-intrain,]
trctrl <- trainControl(method = "repeatedcv", number = 10, repeats = 5)
grid <- expand.grid(C = c(0,0.01, 0.05, 0.1, 0.25, 0.5, 0.75, 1, 1.25, 1.5, 1.75, 2,5))
svm_Linear_Grid <- train(y ~., data = training, method = "svmLinear", trControl=trctrl, preProcess = c("center", "scale"), tuneGrid = grid, tuneLength = 10)
train_pred_grid <- predict(svm_Linear_Grid, newdata = training)
confusionMatrix(train_pred_grid, training$y)
test_pred_grid <- predict(svm_Linear_Grid, newdata = testing)
confusionMatrix(test_pred_grid, testing$y )
```

- *DVM için polinomiyal çekirdek için,*

```
set.seed(3003)
intrain <- createDataPartition(y = data1$y, p= 0.7, list = FALSE)
training <- data1[intrain,]
testing <- data1[-intrain,]
svfit=svm(y~.,data=data1, kernel="polynomial", gama=?, cost=?,degree=?)
tran<-predict(svfit, newdata = svm.train)
confusionMatrix(tran, svm.train$y)
testsv<-predict(svfit, newdata = svm.validate)
confusionMatrix(testsv, svm.validate$y)
```

- *DVM için sigmoid çekirdek için,*

```
set.seed(3003)
intrain <- createDataPartition(y = data1$y, p= 0.7, list = FALSE)
training <- data1[intrain,]
testing <- data1[-intrain,]
svfit=svm(y~.,data=data1, kernel="sigmoid", gama=?, cost=?)
tran<-predict(svfit, newdata = svm.train)
confusionMatrix(tran, svm.train$y)
testsv<-predict(svfit, newdata = svm.validate)
confusionMatrix(testsv, svm.validate$y)
```

- *DVM için radyal tabanlı çekirdek için,*

```
set.seed(5000)
intrain <- createDataPartition(y = data1$y, p= 0.7, list = FALSE)
training <- data1[intrain,]
testing <- data1[-intrain,]
trctrl <- trainControl(method = "repeatedcv", number = 10, repeats = 5)
grid_radial <- expand.grid(sigma = c(0,0.01, 0.02, 0.025, 0.03, 0.04,
0.05, 0.06, 0.07,0.08, 0.09, 0.1, 0.25, 0.5, 0.75,0.9), C = c(0,0.01, 0.05, 0.1, 0.25, 0.5, 0.75, 1, 1.5, 2,5))
svm_Radial_Grid <- train(y ~., data = training, method = "svmRadial“, trControl=trctrl, preProcess = c("center", "scale"), tuneGrid = grid_radial, tuneLength = 10)
train_pred_Radial_Grid <- predict(svm_Radial_Grid, newdata = training)
confusionMatrix(train_pred_Radial_Grid, training$y)
test_pred_Radial_Grid <- predict(svm_Radial_Grid, newdata = testing)
confusionMatrix(test_pred_Radial_Grid, testing$y)
```

- *AUC radyal tabanlı çekirdek için;*

```
gc_ctrl1 <- trainControl(method = "repeatedcv", number = 10, repeats = 5, classProbs = TRUE, summaryFunction = twoClassSummary, savePredictions = TRUE)

gc_train1 <- train(y~., training, method = "svmLinear", tuneLength = 10, trControl = gc_ctrl1, preProc = c("center", "scale"), metric = "ROC", verbose = FALSE)

gc_prob <- predict(gc_train1, newdata = training %>% select(-y), type = "prob")

gc_pROC <- roc(response = training$y, predictor = gc_prob[, "0 or sağ "])

gc_pROC$auc
```

- *RF kodları*

```
set.seed(5000)

ind<-sample.split(Y=data1$y, SplitRatio = 0.7)

traindf<-data1[ind,]

testdf<-data1[!ind,]

modelrandom<-randomForest(y~., traindf, mtry=3 or 4, ntree=100 or 500)

modelrandom

rf.roc<-roc(traindf$y, modelrandom$votes[,2])

auc(rf.roc) or plot(rf.roc)

rf.roc<-roc(testdf$y, modelrandom$votes[,2])

auc(rf.roc)
```

- *YSA doğrusal aktivasyon fonksiyonu için*,

```
set.seed(5000)

levels(data1$y)<-c("first_class", "second_class")

intrain <- createDataPartition(y = data1$y, p= 0.7, list = FALSE)

training <- data1[intrain,]

testing <- data1[-intrain,]
```

```
fitControl<-trainControl(method = "repeatedcv", number = 10, repeats = 5, classProbs = TRUE, summaryFunction = twoClassSummary)

nnetGrid <- expand.grid(size = seq(from = 1, to = 10, by = 1), decay = seq(from = 0.1, to = 0.6, by = 0.1))

nnetFit <- train(y ~ ., data = training, method = "nnet", metric = "ROC", trControl = fitControl, tuneGrid = nnetGrid, verbose = FALSE)

train_pred_nnet <- predict(nnetFit, newdata = training)

confusionMatrix(train_pred_nnet, training$y)

test_pred_nnet <- predict(nnetFit, newdata = testing)

confusionMatrix(test_pred_nnet, testing$y)
```

- *YSA tanh aktivasyon fonksiyonu için,*

```
data1<-as.data.frame(data1)

intrain <-c(1:212)

train.x = data.matrix(data1[intrain, 1:15])

train.x<-normalize(train.x, method = "standardize", range = c(0, 1))

train.y = data1[intrain, 16]

model <- mx.mlp(train.x, train.y, hidden_node=9, out_node=2, activation = "tanh", out_activation="softmax", num.round=20, array.batch.size=15, learning.rate=0.07, momentum=0.09, eval.metric=mx.metric.accuracy, array.layout = “rowmajor”)

preds = predict(model, train.x)

pred.label = max.col(t(preds))-1

table(pred.label, train.y)
```

- *YSA sigmoid aktivasyon fonksiyonu için,*

```
data1<-as.data.frame(data1)

indexes = createDataPartition(data1$y, p = 0.7, list = F)

train <- data1[indexes, ]
```

```
test <- data1[-indexes, ]

train.x <- data.matrix(train[, 1:15])

train.y <- as.numeric(train[, 16])

test.x <- data.matrix(test[, 1:15])

test.y <- as.numeric(test[, 16])

mx.set.seed(0)

model <- mx.mlp(train.x, train.y, hidden_node = 9, out_node = 2, out_activation = "softmax", activation = "sigmoid", num.round = 10, array.batch.size = ?, learning.rate = ?, momentum = ?, eval.metric = mx.metric.accuracy, ctx = mx.cpu(), array.layout = "rowmajor")

pred <- predict(model, train.x, array.layout = "rowmajor")

pred.label <- max.col(t(pred))-1

cfm <- confusionMatrix(as.factor(pred.label), as.factor(train.y))

pred <- predict(model, test.x, array.layout = "rowmajor")

pred.label <- max.col(t(pred))-1

cfm <- confusionMatrix(as.factor(pred.label), as.factor(test.y))
```

- *Simülasyonla elde edilen veriler ve DVM; (15 bağımsız değişken, 0,4 risk oranı ve 1000 gözlem)*

```
mu=c(0,0,0,0,0,0,0,0,0,0,0,0,0,0,0)

a1<-mvrnorm(n=1000, mu=mu, Sigma = diag(15))

A1<-data.frame(a1)

a2<-rbinom(1000, 1, 0.4)

A2<-data.frame(a2)

A3<-cbind(A2,A1)

intrain <- createDataPartition(y = A3$a2, p= 0.7, list = FALSE)

training <- A3[intrain,]

testing <- A3[-intrain,]

trctrl <- trainControl(method = "cv", number = 10)
```

svm_Radial <- train(a2 ~., data = training, method = "svmRadial", trControl=trctrl, preProcess=c("center","scale"), tuneLength = 10)

train_pred_Radial <- predict(svm_Radial, newdata = training)

confusionMatrix(train_pred_Radial, training$a2)

test_pred_Radial <- predict(svm_Radial, newdata = testing)

confusionMatrix(test_pred_Radial, testing$a2)

- *Simülasyonla elde edilen veriler ve RF*

mu=c(0,0,…, 0)

a1<-mvrnorm(n=?,mu=?, Sigma = diag(?))

A1<-data.frame(a1)

a2<-rbinom(n, size, prob)

a2<-as.factor(a2)

A2<-data.frame(a2)

A3<-cbind(A2,A1)

ind<-sample.split(Y=A3$a2, SplitRatio = 0.7)

traindf<-A3[ind,]

testdf<-A3[!ind,]

modelrandom<-randomForest(a2~.,traindf,mtry=?, ntree=?)

modelrandom

modelrandom<-randomForest(a2~.,testdf, mtry=?, ntree=?)

modelrandom

- *Simülasyonla elde edilen veriler ve YSA*

mu=c(0,0,…, 0)

a1<-mvrnorm(n=?,mu=?, Sigma = diag(?))

A1<-data.frame(a1)

```
a2<-rbinom(n, size, prob)

a2<-as.factor(a2)
A2<-data.frame(a2)

A3<-cbind(A2,A1)

levels(A3$a2) <- c("first_class", "second_class")

intrain <- createDataPartition(y = A3$a2, p= 0.7, list = FALSE)

training <- A3[intrain,]

testing <- A3[-intrain,]

fitControl <- trainControl(method = "cv", number = 10, classProbs = TRUE,
summaryFunction = twoClassSummary)

nnetGrid <- expand.grid(size = seq(from = 1, to = 10, by = 3), decay = seq(from =
0.1, to = 0.5, by = 0.2))

nnetFit <- train(a2 ~ ., data = training, method = "nnet", metric = "ROC", trControl =
fitControl, tuneGrid = nnetGrid, verbose = FALSE)

train_pred_nnet <- predict(nnetFit, newdata = training)

confusionMatrix(train_pred_nnet, training$a2)

test_pred_nnet <- predict(nnetFit, newdata = testing)

confusionMatrix(test_pred_nnet, testing$a2)
```