T.C.

ESKİŞEHİR OSMANGAZİ ÜNİVERSİTESİ

SAĞLIK BİLİMLERİ ENSTİTÜSÜ

BİYOİSTATİSTİK ANABİLİM DALI

GAIL MODELİ İLE MAKİNE ÖĞRENMESİ ALGORİTMALARININ MEME KANSERİ RİSK DEĞERLENDİRMESİNDE KARŞILAŞTIRILMASI

YÜKSEK LİSANS

BERFU PARÇALI

DANIŞMAN

PROF. DR. FEZAN MUTLU

Eskişehir

2020

T.C.

ESKİŞEHİR OSMANGAZİ ÜNİVERSİTESİ

SAĞLIK BİLİMLERİ ENSTİTÜSÜ

BİYOİSTATİSTİK ANABİLİM DALI

GAIL MODELİ İLE MAKİNE ÖĞRENMESİ ALGORİTMALARININ MEME KANSERİ RİSK DEĞERLENDİRMESİNDE KARŞILAŞTIRILMASI

YÜKSEK LİSANS

BERFU PARÇALI

DANIŞMAN

PROF. DR. FEZAN MUTLU

Eskişehir

2020

TEŞEKKÜR

Tez çalışmam sırasında bilgi, birikim ve tecrübeleri ile destek olan ve önerilerini göstermekten kaçınmayan değerli danışman hocam sayın Prof. Dr. Fezan MUTLU'ya teşekkür ve saygılarımı sunarım.

Çalışmalarım boyunca yardımlarını hiç esirgemeyen değerli arkadaşlarım Fatma Gül GEZER, Eylem GÜL, İsmail GÜR, Burak AKDEMİR, Betül AYDIN ve Ezgi ÖZKURT'a çok teşekkür ederim.

Çalışmalarım boyunca maddi ve manevi desteği ile beni asla yalnız bırakmayan annem Leyla PARÇALI ve kardeşim Fatma Ece PARÇALI 'ya sonsuz teşekkürler ederim.

17 / 12 / 2020

Berfu PARÇALI

ÖZET

Meme kanserinin erken aşamada teşhis edilmesi; tedavi yöntemlerinin sayısını, tedavinin başarıya ulaşma oranını ve hayatta kalma şansını arttırmaktadır. Gail Modeli, meme kanserinde temel faktörleri değerlendiren, kabul görmüş kanser riski değerlendirme modelidir. Bu çalışmada Gail Modeli baz alınarak makine öğrenmesi yöntemlerinin meme kanseri risk değerlendirmesinde karşılaştırılması amaçlanmıştır. İlk olarak veri setine Gail Modeli uygulanmış ve risk faktörü belirlenmiş, %70 eğitim %30 test ve %80 eğitim %20 test olmak üzere 2 ayrı eğitim test veri seti oluşturulmuştur. Daha sonra veri setlerine k-En Yakın Komşu, Yapay Sinir Ağları, Destek Vektör Makinesi ve Naive Bayes algoritmaları uygulanmış ve risk tahmin sonuçları karşılaştırılmıştır. Karşılaştırma sonuçlarına göre %70 eğitim %30 test veri seti için sınıflandırma performansı en düşükten en yükseğe doğru sırası ile k-NN (AUC=0.5375), NB (AUC=0.8542), SVM (AUC=0.9375) ve YSA(AUC=0.9875) şeklindedir. %80 eğitim %20 test veri seti için sınıflandırma performansı en düşükten en yükseğe doğru sırası ile k-NN (AUC=0.5892), SVM (AUC=0.9088), NB (AUC=0.9305) ve YSA (AUC=0.9718) şeklindedir.

Anahtar kelimeler: Gail Modeli, Makine Öğrenmesi, Meme Kanseri, Yapay Sinir Ağları, Destek Vektör Makinesi, k-En Yakın Komşu, Naive Bayes

SUMMARY

Early diagnosis of breast cancer increases the number of possible treatments, the success rate of the treatments and the chance of survival. The Gail Model is a well accepted cancer risk assessment model which evaluates the main factors in breast cancer. The aim of this work is compare machine learning methods in breast cancer risk assessment based on the Gail Model.SVM, k-NN, ANN, NB algorithm with the purpose of breast cancer risk assessment. Firstly, risk factor was determined using the Gail method on the dataset, then dataset was divided into training - testing sets using 70 - 30 and 80 - 20 splits which resulted in two seperate training and testing sets. Afterwards, on each set, k-NN, ANN, SVM and NB algorithms were applied and results were compared based on the classification performance. According to the comparison results, the classification performance for 70% training and 30% test data set was k-NN (AUC=0.5375), NB (AUC=0.8542), SVM (AUC=0.9375) and ANN (AUC=0.9875) directly from from lowest to highest. On the other hand, for 80% training and 20% test data set, classification performance wasis from lowest to highest, respectively, k-NN (AUC=0.5892), SVM (AUC=0.9088), NB (AUC=0.9305) and ANN (AUC=0.9718).

Keywords: Gail Model, Machine Learning, Breast Cancer, Artificial Neural Network, Support Vector Machine, k-Nearest Neighbor, Naive Bayes

İÇİNDEKİLER

TABLO DİZİNİ

ŞEKİL DİZİNİ

SİMGE VE KISALTMALAR DİZİNİ

AUC	:	ROC Eğrisi Altında Kalan Alan
BCDDP	:	Meme Kanseri Tanı ve Tespit Projesi
BCRAT	:	Gail 2
BM	:	Boltzmann Makinesi
BOADICEA Modeli	:	Breast And Ovarian Analysis Of Disease Incidence and Carrier Estimation Algorithm
BRCAPRO	:	Berry-Parmigiani-Aguilar Modeli
DBM	:	Derin Boltzmann Makineleri
DP	:	Dinamik Programlama
DQN	:	Derin Q-Ağları
DRAM	:	Dinamik Rastgele Erişimli Bellek
DSA	:	Derin Sinir Ağları
GMM	:	Gaussian Karışım Modeli
IBIS-I	:	Uluslararası Meme Müdahale Çalışması
KDG	:	Kısmi Doğrusal Gömme
k-NN	:	k-En Yakın Komşu
KSA	:	Konvolüsyon Sinir Ağları
MCMCGLMM	:	Markov Zinciri Monte Carlo Genelleştirilmiş Doğrusal Karma Model
NB	:	Naive Bayes
NSABP	:	Ulusal Cerrahi Adjuvan Meme ve Bağırsak Projesi
PCA	:	Temel Bileşen Analizi
RBF	:	Radyal Temel İşlevi

RF	:	Rastgele Orman
SEER	:	Sürveyans Epidemiyoloji ve Sonuçlar
SVM	:	Destek Vektör Makinesi
T1	:	İlk Andaki Yaş
T2	:	Tahmin Yaşı
YSA	:	Yapay Sinir Ağları

1. GİRİŞ VE AMAÇ

Meme kanseri hem Türkiye'de hem de dünya genelinde kadınlarda sık görülen bir kanser türüdür. Sık görülen bu kanser türünün, kanser sebebi ile ölümlerin üst sıralarında yer almadığı bilinmektedir (Özmen 2013).

Gelir düzeyi yüksek olan ülkelerde genellikle meme kanseri hastalarının %70'inden fazlası kanserin I ve II. evrelerinde teşhis edilmektedir ancak, gelir düzeyi düşük ve orta olan ülkelerin çoğunda bu oran sadece %20-50 civarındadır (Suarez, Perez-Castejon, Jimenez, & Domper, 2002). 2006 ve 2016 yılları arasında kanser vaka oranları %28 civarında artış göstermiş ve en hızlı artış az gelişmiş ülkelerde gerçekleşmiştir (Fitzmaurice vd., 2018).

Meme kanseri; meme dokusu içerisinde yer alan süt kanallarının doku hücrelerinde oluşmaktadır. Süt kanallarını oluşturan doku hücrelerinin kontrolsüz olarak artmasına duktal hiperplazi denir (Barton, 2005).

Meme kanseri gelişiminde etkili birçok risk faktörü bulunmaktadır. Bunlardan bazıları, vücut kitle indeksi, dens meme yapısı, menarş yaşı, menopoz yaşı, laktasyon, ilk tam gebelik yaşı, emzirme, düşük yapma, alkol kullanımı, aile öyküsü, östrojen ve progesteron seviyesi, bilinen ya da şüphe edilen BRCA 1/2, atipik hiperplazi veya lobüler karsinoma in situdur (Kelsey, Gammon, & John, 1993). Bir kadında, yaşamı süresince invaziv (yayılma eğilimi olan) meme kanseri gelişme riskinin %13,3 olduğu bilinmektedir ve meme kanserinin oluşma riski yaşa bağlı olarak artmaktadır (Phillips, Glendon, & Knight, 1999). Meme kanseri teşhislerinin yaklaşık %18'i 40'lı yaşlardaki kadınlar arasındadır ve meme kanseri olan kadınların %77'si teşhis konulduğunda 50 yaşının üzerindedir (Chapman vd., 2007).

Gail Modeli 1973 – 1980 yılları arasında meme taraması yapılmış olan 284.780 Amerikalı beyaz kadınların Meme Kanseri Tanı ve Tespit Projesi (BCDDP) verileri kullanılarak 1989 yılında geliştirilmiştir (Gail vd., 1989).

Karakayali ve Ekici (2007) çalışmalarında Gail Modeli'ni kullanmıştır. Sonuçlar değerlendirildiğinde, Gail Modeli'nin bireysel meme kanseri risk hesaplamasında güvenilir olduğu görülmüştür.

Ahmad ve arkadaşları (2013) bireylerde meme kanseri tekrarını tahmin etmek için makine öğrenmesi yöntemlerinden 3 tanesini karşılaştırmışlardır. Bu yöntemler,

Karar Ağaçları, Destek Vektör Makinesi ve Yapay Sinir Ağları'dır. Bu yöntemler içerisinde meme kanseri tekrarını en az hata oranı ve en yüksek doğrulukla Destek Vektör Makinesi yöntemi tahmin etmiştir. Karar Ağacı modeli ise diğer iki modele göre öngörülen doğruluğu, en düşük doğrulukla tahmin etmiştir (Ahmad, Eshlaghy, Poorebrahimi, Ebrahimi, & Razavi, 2013).

Asri ve arkadaşları (2016) çalışmalarında meme kanseri risk tahmini ve teşhisi için Destek Vektör Makinesi, Karar Ağacı, Naive Bayes ve k-En Yakın Komşular makine öğrenme algoritmalarının performanslarını karşılaştırmışlardır. Çalışmanın amacı, verileri her algoritmanın verimlilik ve etkinlik açısından sınıflandırmadaki doğruluğunu; doğruluk, kesinlik, duyarlılık ve özgüllük açısından değerlendirmektir. Değerlendirme sonuçlarında, Destek Vektör Makinesi'nin en düşük hata oranı ile en yüksek doğruluğu verdiği görülmüştür. (Asri, Mousannif, Al Moatassime, & Noel, 2016).

X. Wang ve arkadaşları (2018) meme kanseri riskini tahmin etmek için Gail modelinin performansını ardışık deneme analiziyle sistematik bir şekilde gözden geçirmiş ve meta-analiz değerlendirmesi çalışmışlardır. Çalışma sonucunda Gail Modeli'nin; Amerikalı ve Avrupalı kadınlarda meme kanseri insidansını tahmin etme performansı yüksekken, bireysel düzeyde ki risk tahmini için daha az olduğu görülmüştür. Gail Modeli'nin, Asyalı kadınlarda da meme kanseri riskini yüksek performansla tahmin ettiği doğrulanmıştır (X. Wang vd., 2018).

Saritas ve Yasar (2019) çalışmalarında, meme kanseri şüphesiyle kliniğe başvuran hastaların verilerine Yapay Sinir Ağları ve Naive Bayes sınıflandırma algoritmaları uygulamış ve hastalık tanısını tahmin etmeleri istenmiştir. YSA algoritması %86.95, NB algoritması ise %83.54 doğrulukla sınıflandırmıştır (Saritas & Yasar, 2019).

Stark ve arkadaşları (2019) çalışmalarında 5 yıllık meme kanseri riskini BCRAT ve makine öğrenmesi yöntemleriyle karşılaştırmışlardır. Karşılaştırma sonucunda BCRAT doğruluk yüzdesi %56.3, Lojistik Regresyon ve Doğrusal Diskriminant Analizi doğruluk yüzdesi %61.3 ve Yapay Sinir Ağları doğruluk yüzdesi %60.8 olduğu gözlenmiştir (Stark, Hart, Nartowt, & Deng, 2019).

Ming ve arkadaşları (2020) çalışmalarında bir örneklemde yaşam boyu meme kanseri riskini tahmin etmek için 3 farklı makine öğrenmesi algoritması Markov

Zinciri Monte Carlo Genelleştirilmiş Doğrusal Karma Model (MCMCGLMM), AdaBoost ve Rastgele Orman (RF) ve 0A modeli kullanmışlardır. BOADICEA % 63.9, AdaBoost %88.9, MCMCGLMM % 85.1, RF % 84.3 tahmin doğrulu ile sınıflandırmıştır (Ming vd., 2020).

Bu çalışmanın amacı; meme kanseri riskinin hesaplanmasında yaygın olarak kullanılan Gail Modeli ile makine öğrenmesi yöntemlerinin karşılaştırılması ve meme kanseri riskinin hesaplanmasında hangi makine öğrenmesi yönteminin daha etkili olduğunu saptamaktır.

2. GENEL BİLGİLER

2.1. Meme Kanserinde Risk Tahmin Modelleri

Meme kanserinin erken aşamada teşhis edilmesi; tedavi yöntemlerinin sayısını, tedavinin başarıya ulaşma oranını ve hayatta kalma şansını arttırmaktadır.

Meme kanserinin gelişme riskini hesaplamak için birçok istatistiksel model geliştirilmiştir (van Asperen et vd., 2004). Günümüzde, birçok araştırmacı tarafından çeşitli istatistiksel yöntemler kullanılarak geliştirilmiş çok sayıda risk tahmin modeli bulunmaktadır. Bu istatistiksel tahmin modellerinde yer alan değişkenler genellikle demografik özellikler ve biyomedikal verilerdir (Boyle vd., 2004).

Geliştirilen bu istatistiksel modeller arasında en çok kullanılan modellerin başında Gail ve Claus modelleri yer almaktadır. Ancak bu modeller de meme kanseri gelişme riskini tam anlamı ile değerlendirememektedir (Dumitrescu & Cotarla, 2005).

2.1.1. Gail modeli

Bir risk tahmini olan Gail Modeli'nin geçerliliği, meme kanserini önleme stratejisinde uygun klinik kararlar vermek için çok önemlidir. Gail Modeli'nin performans doğruluğunu değerlendirmek için en yaygın olarak değerlendirilen iki konu kalibrasyon ve genetik kökendir. Daha spesifik olmak gerekirse; kalibrasyon, tahmin modelinin popülasyon düzeyindeki performansını ölçmek için bir araçtır. Genetik köken ise modelin farklı popülasyonlardaki bireysel risklerini ayırt etme yeteneğini değerlendirmektedir (Anothaisintawee, Teerawattananon, Wiratkapun, Kasamesup, & Thakkinstian, 2012).

Meme kanseri riskini mümkün olan en doğru biçimde tespit edebilmek için, meme kanseriyle ilişkili çok sayıda risk faktörünü değerlendirmek önemlidir (Amir, Freedman, Seruga, & Evans, 2010).

Gail Modeli için en ideal yaş aralığı 35 yaş ve üzeri, düzenli mamografi taramalarına devam eden ve gen havuzu dar olan bireylerdir (van Asperen vd., 2004). Meme kanseri gelişmesinde yaş tek başına en önemli bağımsız faktördür. Çünkü yaş artıkça meme kanserine yakalanma riski de artmaktadır (Eroglu, Eryilmaz, Civcik, & Gurbuz, 2010). Gail Modeli, meme kanserinde değiştirilemeyen faktörleri değerlendiren, onaylanmış kanser riski değerlendirme modelidir. Bu model BRCA

1/2 mutasyon taşıyıcılarında ki kanser risklerini tahmin etmek için uygun değildir (Engel & Fischer, 2015).

1989 yılında geliştirilen Gail Modeli (Model 1), 1999 yılında Ulusal Cerrahi Adjuvan Meme ve Bağırsak Projesi (NSABP) istatistikçileri tarafından sadece invaziv meme kanseri gelişme mutlak riskini yansıtmak için BCRAT (Gail 2) Modeli olarak değiştirilmiştir (Costantino vd., 1999).

Değiştirilen modelin ilk modelden farkı şunlardır:

BCRAT Modeli'ndeki meme kanseri insidans oranları; Model 1'de bulunan invaziv kanser ve in sitü (yayılma eğilimi olmayan) kanser türlerinden sadece invaziv kanser türlerini içermektedir (Costantino vd., 1999).

BCRAT Modeli'nde; yaşlara özgü invaziv meme kanseri oranlarının Sürveyans Epidemiyoloji ve Sonuçlar (SEER) verileri kullanılmıştır, Model 1'de BCDDP verileri kullanılmıştır (Costantino vd., 1999).

En önemli değişiklik risk faktörlerine, atipik hiperplazi ile birlikte olan meme biyopsilerinin de eklenmesidir (Karakayali vd., 2007).

Son olarak Model 1'de Amerikalı beyaz kadınların meme kanseri insidans oranları bulunurken, BCRAT Modeli'nde Afro-Amerikan hastalar için karma meme kanseri insidans oranları da dahil edilmiştir (Amir vd., 2010)

BCRAT Modeli'nde, 5 yıl içerisinde meme kanserine yakalanma ihtimali %1,66'dan az olan birey düşük riskli, %1,66 veya daha fazla olan birey ise yüksek riskli olarak değerlendirilmektedir (Amir vd., 2010).

Gail Modeli'nde meme kanseri riski hesaplanmasında önemli olan faktörler şunladır:

- Yaş (20-70),
- Menarş yaşı,
- İlk doğum yaşı,
- Meme biyopsisi,
- Atipik duktal hiperplazi,
- Birinci derece yakınlardaki meme kanseri sayısı (Evans & Howell, 2007).

Gail Modeli ile yaş endeksli $(a+\tau)$ rölatif risk aşağıdaki formülle hesaplanmaktadır.

$$P\{\alpha,\tau,r(t)\} = \int_a^{a+\tau} h_1(t)r(t)e^{-\int_a^t h_1(u)r(u)du}\{S_2(t)/S_2(a)dt\} \quad (2.1)$$

Burada;

S_2: t yaşına kadar hayatta kalanların yarışan riskleri olasılığını ifade eder ve aşağıdaki formül ile hesaplanır.

$$S_2(t) = e^{-\int_0^t h_2(u)du} \quad (2.2)$$

$h_1(t)$: Risk faktörleri bilinmeyen bireyin yaş endeksli riskini temsil etmektedir.

$h_2(t)$: Bireyin yaş endeksli ölüm nedenleri riskini temsil etmektedir.

$r(t)$: Rölatif (Göreceli) riski temsil etmektedir.α: Bireyin riski hesaplandığı andaki yaşını temsil etmektedir (Gail vd., 1989).

2.1.2. Claus modeli

Yaygın olarak kullanılan bir diğer genetik model Claus tarafından 1991 geliştirilmiştir. Bu model, çok merkezli ve toplum temelli vaka kontrol çalışması olan Kanser ve Steroid Hormon Çalışması'nın (Cancer and Steroid Hormone Study) verileri kullanılarak geliştirilmiştir (Claus, Risch, & Thompson, 1991). Claus Modeli, yaşa bağlı penetrasyon ile bir otozomal dominant gen varsaymaktadır. Bu model genellikle, meme kanseri gelişimi bakımından riske sahip olan ailelerde ki meme kanseri riskini tahmin etmek amacıyla kullanılmaktadır (Engel & Fischer, 2015).

Claus Model'inde olasılıklar, vakaların ve kontrollerin (aynı zamanda probandlar olarak da tanımlanır) annelerinin ve kız kardeşlerinin ortak analizi olarak hesaplanır. Bir vakanın annesinin olasılığı, vakanın etkilendiği yaşa bağlı olarak hesaplanır. Vakanın kız kardeşlerinin olasılığı da hem vakanın etkilendiği yaşa hem de annenin meme kanseri durumuna bağlı olarak hesaplanır.

Meme kanseri olan bir annenin kansere yakalandığı yaş ya da kanser hastası olmayan bir anne durumunda, annenin mevcut yaşı veya ölüm anındaki yaşı modele dahil

edilir (Claus vd., 1991). Proband yakınları için, anne ve kız kardeşlerin olasılığı, probandın mevcut yaşına bağlı olarak hesaplanmaktadır. Bu analizlerde, kanser başlangıç yaş dağılımı ve meme kanseri riskinin, A (anormal) ve a (normal) alelleri ile aynı diallelik major konuma bağlı olduğu varsayılmaktadır (Claus vd., 1991).

Zamanı t_i aralıklarına ayırıp, her bir alt uç noktası t_{i-1} ve üst uç noktası t olan Meme Kanserinin Genetik Analizi ve φ_i, genotipi aa için i aralığında ki riski temsil ettiğini kabul edelim.

Bu durumda genotip aa için kümülatif risk fonksiyonu şöyle yazılabilir:

$$H(t) = \sum_{j=1}^{i-1} \varphi_j(t_j - t_{j-1}) + \varphi_i(t - t_{i-1}) \tag{2.3}$$

Bu nedenle, bir kadının i aralıkta meme kanserinden etkilenme olasılığı veya penetransı:

$$\lambda_i = F(t_i) - F(t_{i-1}) = e^{-\sum_{j=1}^{i} \varphi_j(t_j - t_{j-1})} - e^{-\sum_{j=1}^{i-1} \varphi_j(t_j - t_{j-1})} \tag{2.4}$$

formülü ile hesaplanmaktadır (Claus vd., 1991).

2.1.3. Berry-Parmigiani-Aguilar modeli (BRCAPRO)

BRCAPRO Modeli, BRCA 1 ve BRCA 2'nin mutasyon taşıma olasılığını tahmin etmek için Bayes prensiplerini kullanarak geliştirilen bir matematiksel modeldir. Model, kanserden etkilenen ve etkilenmeyen tüm birinci ve ikinci derece akrabaların soyunu içermektedir (W. S. Rubinstein, O'Neill, Peters, Rittmeyer, & Stadler, 2002). Etkilenen akrabalarda başlangıç yaşı ve etkilenmeyen akrabalarda kansersiz yaş, bu modelin önemli bileşenleridir.

BRCAPRO Modeli, beyaz ve Aşkenazi Yahudi popülasyonlarının mutasyon frekansları kullanılarak tasarlanmıştır ve diğer popülasyonlar için etkili bir model olmayabilir (Vogel vd., 2007).

2.1.4. Couch modeli

Meme kanseri olan 263 kadından, aile öyküsü ve DNA analizi için kan alınmış daha sonra BRCA 1 mutasyonlarını tanımlamak için konformasyona duyarlı jel elektroforezi ve DNA sıralaması kullanılmıştır. (Couch vd., 1997).Tek değişkenli analizde BRCA 1 mutasyonlarını öngören değişkenleri kullanarak, lojistik

regresyonu temel alan bir olasılık tahmin modeli geliştirilmiştir (Lindor vd., 2007). Geliştirilen bu modelin analiz sonuçları Tablo 2.1'de sunulmuştur (Couch vd., 1997).

Tablo 2.1. Couch tablosu

Meme Kanseri Teşhisinde Ortalama Yaş	Tahmin Edilen Yüzde Olasılık (%95 GA)	Meme Kanseri Teşhisinde Ortalama Yaş	Tahmin Edilen Yüzde Olasılık (%95 GA)
Sadece meme kanseri olan aileler		Sadece meme kanseri olan Aşkenazi Yahudi aileler	
<35	17.4 (6.5–38.8)	<35	47.9
35-39	11.7 (5.1–24.6)	35-39	36.7 (12.8–69.6)
40-44	7.7 (3.6–15.6)	40-44	26.8 (9.7–55.3)
45-49	5.0 (2.3–10.8)	45-49	18.7 (6.8–42.0)
50-54	3.2 (1.2–8.1)	50-54	12.7 (4.3–31.8)
55-59	2.1 (0.6–6.5)	55-59	8.4 (2.5–24.8)
>59	1.3 (0.3–5.5)	>59	5.5 (1.3–20.0)
Meme ve yumurtalık kanseri olan aileler		Meme ve yumurtalık kanseri olan Aşkenazi Yahudi aileler	
<35	55.0 (27.2–80.0)	<35	84.2
35-39	43.5 (22.4–67.2)	35-39	77.1 (40.1-94.4)
40-44	32.7 (17.0–53.5)	40-44	67.9
45-49	23.4 (11.4–42.1)	45-49	57.2 (24.9-84.3)
50-54	16.2 (6.7–34.2)	50-54	45.7
55-59	10.8 (3.5–28.8)	55-59	4.7 (10.8-70.0)
>59	7.1 (1.7–24.8)	>59	25.01
Aile üyelerinden birinde meme ve yumurtalık kanseri varsa		Aşkenazi Yahudi aile üyelerinden birinde meme ve yumurtalık kanseri varsa	
<35	77.0	<35	93.6
35-39	67.8 (37.1–88.3)	35-39	90.2
40-44	57.1 (28.4–81.7)	40-44	85.3
45-49	54.5	45-49	78.5
50-54	34.6 (12.1–67.0)	50-54	69.8
55-59	25.0	55-59	59.3
>59	17.03	>59	47.8
Meme ve yumurtalık kanseri olan ailelerde meme ve yumurtalık kanseri olan 1 üye varsa		Meme ve yumurtalık kanseri olan Aşkenazi Yahudi ailelerde meme ve yumurtalık kanseri olan 1 üye varsa	
<35	96.6	<35	98.8
35-39	92.4 (72.0–98.3)	35-39	96.8
40-44	88.5 (63.4–97.2)	40-44	98.1
45-49	82.9 (52.0–95.6)	45-49	95.5
50-54	75.4	50-54	93.0
55-59	65.9	55-59	89.4
>59	54.9	>59	81.3

2.1.5. Breast and ovarian analysis of disease incidence and carrier estimation algorithm (BOADICEA modeli)

BOADICEA Modeli, bir BRCA 1 veya BRCA 2'nin mutasyon taşıma olasılığını ve meme veya yumurtalık kanseri gelişme risklerini tahmin etmek için kullanılmaktadır (Antoniou vd., 2008).

2.1.6. Tyrer – Cuzick modeli

Tyrer-Cuzick Modeli, meme kanserine genel bakış çalışmalarından göreceli riskleri bir araya getirerek geliştirilmiştir. Bu model ilk defa bir önleme denemesi olan Uluslararası Meme Müdahale Çalışması'nın (IBIS-I) geçerliliğini değerlendirmek amacıyla kullanılmıştır (Brentnall vd., 2015). Tyrer-Cuzick Modeli; BRCA 1 / 2'nin mutasyon taşıma durumunu ve menarş yaşı, ilk doğum yaşı, menopoz yaşı, atipik hiperplazi, yerinde lobüler karsinom, boy ve vücut kitle indeksi bilgilerini kullanarak meme kanseri riskini hesaplamaktadır (Tyrer, Duffy, & Cuzick, 2004).

2.1.7. Modellerin karşılaştırılması

Tyrer-Cuzick Modeli ve Claus Modeli, invaziv meme kanseri riski ve duktal karsinom in sitü meme kanseri riskini ortaya koymaktadır. Ancak BCRAT, BRCAPRO ve BOADICEA Modelleri'nde ise yalnızca invaziv meme kanseri riskini ortaya koymaktadır. Claus Modeli ve BRCAPRO Modeli, yüksek riskli bir klinik popülasyonda genellikle BCRAT Modeli'den daha düşük risk tahminlerine sahiptir (Gail & Mai, 2010). Bilinen aile öyküsü olmayan meme kanserinden etkilenen birçok bireyin, risk faktörleri çoğu risk modellerine dahil edilmemiştir. Modellerin karşılaştırılması Tablo 2.2'de belirtilmiştir (Evans & Howell, 2007).

Tablo 2.2. Risk faktörleri ve risk faktörlerini içeren modeller

Risk Faktörleri ve Risk Faktörlerini İçeren Modeller						
	Rölatif Risk	**Gail Modeli**	**Claus Modeli**	**BRCAPRO Modeli**	**Cuzick-Tyrer Modeli**	**BOADICEA Modeli**
Kişisel Bilgiler						
Yaş (20-70)	30	Evet	Evet	Evet	Evet	Evet
Vücut Kitle İndeksi	2	Hayır	Hayır	Hayır	Evet	Hayır
Alkol Kullanımı	1.24	Hayır	Hayır	Hayır	Hayır	Hayır
Hormonal/Reprodüktif Faktörler						
Menarş Yaşı	2	Evet	Hayır	Hayır	Evet	Hayır
İlk Canlı Doğum Yaşı	3	Evet	Hayır	Hayır	Evet	Hayır
Menapoz Yaşı	4	Hayır	Hayır	Hayır	Evet	Hayır
Hormon Tedavisi Kullanımı	2	Hayır	Hayır	Hayır	Evet	Hayır
Doğum Konrol Hapı Kullanımı	1.24	Hayır	Hayır	Hayır	Hayır	Hayır
Emzirme	0.8	Hayır	Hayır	Hayır	Hayır	Hayır
Östrojen Seviyesi	5	Hayır	Hayır	Hayır	Hayır	Hayır
Kişisel Meme Hastalığı Öyküsü						
Meme Biyopsileri	2	Evet	Hayır	Hayır	Evet	Hayır
Atipik Duktal Hiperplazi	3	Evet	Hayır	Hayır	Evet	Hayır
Lobüler Karsinoma İn Sitü	4	Hayır	Hayır	Hayır	Evet	Hayır
Meme Yoğunluğu	6	Hayır	Hayır	Hayır	Hayır	Hayır
Aile Geçmişi						
Birinci Derece Yakınlar	3	Evet	Evet	Evet	Evet	Evet
İkinci Derece Yakınlar	1.5	Hayır	Evet	Evet	Evet	Evet
Üçüncü Derece Yakınlar	-	Hayır	Hayır	Hayır	Hayır	Evet
Kanser Başlangıç Yaşı	3	Hayır	Evet	Evet	Evet	Evet
Bilateral Meme Kanseri	3	Hayır	Hayır	Evet	Evet	Evet
Yumurtalık Kanseri	1.5	Hayır	Hayır	Evet	Evet	Evet
Erkek Meme Kanseri	3-5	Hayır	Hayır	Evet	Hayır	Evet

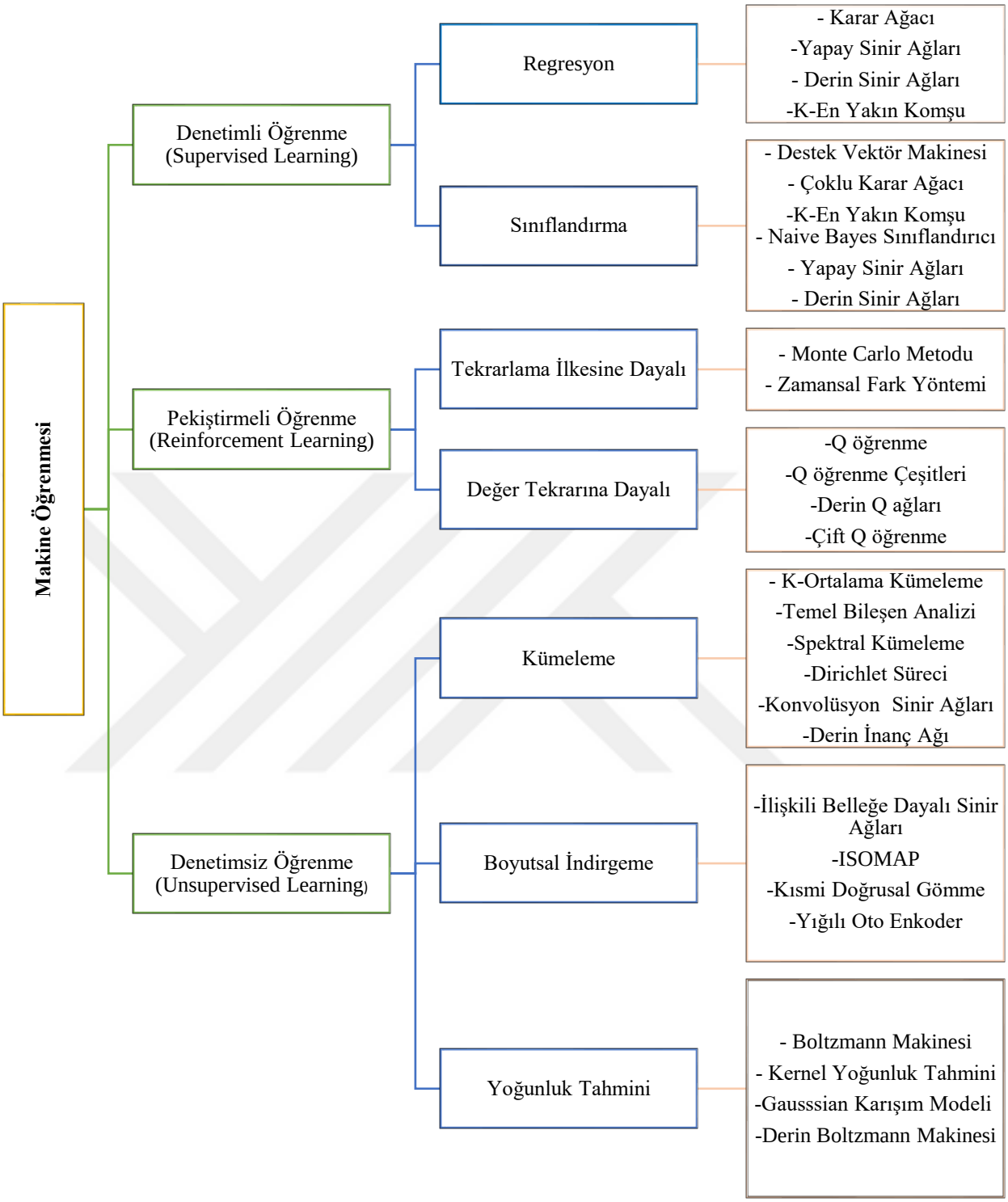
2.2. Makine Öğrenmesi

Makine öğrenmesi; matematiksel ve istatistiksel yöntemler kullanarak, elde olan verilerden tahminler yapan ve bu tahminlerle varılmak istenilen sonuçların tahminlerinde bulunan, modelleme ve algoritmalardan oluşan yapay zekanın bir alt dalıdır (Akay, 2018). Makine öğrenmesinin temel hedefi, doğru çıkarımlar yapmaktır (Özkan & Ülker, 2017). Makine öğrenmesi yöntemleri Şekil 2.1'de gösterilmiştir (Sharma & Wang, 2019).

Makine öğrenmesi ve veri madenciliği iç içedir. Veri madenciliği, çok disiplinli bir alan olup, sağlık alanı en önemli uygulama alanları arasında yer almaktadır.

Makine öğrenmesi, genellikle tarama testlerinden elde edilen verileri kullanarak çeşitli kanser türlerinin ön tanısı ve hasta semptomlarına göre risk ve önceliklerinin tespiti gibi geniş alanlarda uygulanmaktadır (Casanova vd., 2013).

İstatistiksel yöntemlerde ve Yapay Sinir Ağlarında, verilerden fonksiyon üretildikten sonra bu fonksiyonun anlaşılabilir bir kural olarak yorumlanması zordur. Bu nedenle, karar ağaçları oluşturulduktan sonra kökten yaprağa doğru inilerek her dal bir kural oluşturacak şekilde kurallar yazılmaktadır. Bu kural çıkarma algoritması veri madenciliği çalışmalarının doğru sonuçlar elde etmesini sağlamaktadır (Clark vd., 2014).



Şekil 2.1. Makine öğrenmesi yöntemleri

2.2.1. Denetimli öğrenme (Supervised learning)

Denetimli Öğrenme'nin temel amacı, girdiden çıktıya kadar doğru değerlerin denetmen(süpervizör) tarafından sağlanmasıdır. Denetimli Öğrenme'de, hedefler hakkında bilgi eğitim veri kümesine (S) ait bir dizi hedef değişkenden elde edilir (Hastie, Tibshirani, & Wainwright, 2015).

2.2.1.1. Regresyon

İstatistiksel anlamda regresyon sözcüğü ilk kez Francis Galton tarafından kullanılmıştır (Stigler, 1989). Galton uzun yıllar süren çalışmaları sonucunda 1877-1885 dönemlerinde regresyon analizi ile ilgili çalışmalarda bulunmuş ve modern regresyon metinlerinin temelini oluşturmuştur (Stigler, 1997).

Doğrusal regresyon birçok alanda tahmin için yaygın olarak kullanılmaktadır (sigorta veya kredi riski tahmini, kişiselleştirilmiş ilaç, pazar analizi vb.).

Regresyon; iki ya da daha fazla değişken arasındaki ilişkiyi ölçmek için kullanılan analiz metodudur. Sayısal tahmin değerlerini sınıf etiketleriyle eşleştirerek sınıflandırma görevleri için de kullanılabileceğinden güçlü bir metottur (Jagielski vd., 2018).

Regresyon analizindeki temel amaç, en az değişkeni kullanarak bağımlı değişken ile bağımsız değişkenler arasındaki ilişkiyi açıklayabilmek ve genel olarak kabul edilebilen bir model kurmaktır (Coşkun, Kartal, Coşkun, & Bircan, 2004). Açıklanan değişken Y ile ve açıklayıcı değişkenleri X_1, X_2, …, X_k ile belirtirsek, bu değişkenlerle ilgili genel bir model

$$y_i = \beta_0 + \sum_{j=1}^{k} x_{ij}\beta_j + e_i \qquad (2.5)$$

Değişkenlerle ilgili genel modelde β_j parametreleri doğrusaldır

e_i: Hata terimi

β_0: x=0 olduğunda bağımlı değişkenin alacağı değer (kesim noktası)

β_j: Regresyon Katsayısı (Seber & Lee, 2012).

2.2.1.1.1. K-en yakın komşu

k-En Yakın Komşu (k-NN) algoritması parametrik olmayan bir sınıflandırma yöntemidir. k-NN yönteminde, vektörün en yakın komşuları olan sınıflardan yararlanarak sınıflandırma yapılmaktadır. Veri setine katılacak olan yeni verinin, mevcut verilere göre uzaklığı hesaplanıp (Öklid Uzaklığı, Manhattan Uzaklığı ya da Minkowski Uzaklığı), k sayıda yakın komşuluğuna bakılır (Keller, Gray, & Givens, 1985).

Bir öznitelik uzayında A ve B noktaları arasındaki mesafeyi ölçmek için, literatürlerde en çok kullanılan uzaklık Öklid uzaklığıdır. m boyutlu bir öznitelik uzayında, A ve B özellik vektörleri A= (x_1, x_2, …, x_m) ve B= (y_1, y_2, …, y_m) olarak gösterilsin. A ve B arasındaki mesafenin Öklid metriği;

$$Dist(A,B) = \sqrt{\frac{\sum_{i=1}^{m}(x_i - y_i)^2}{m}} \tag{2.6}$$

Formülü ile hesaplanmaktadır (Hu, Huang, Ke, & Tsai, 2016).

2.2.1.1.2. Karar ağaçları

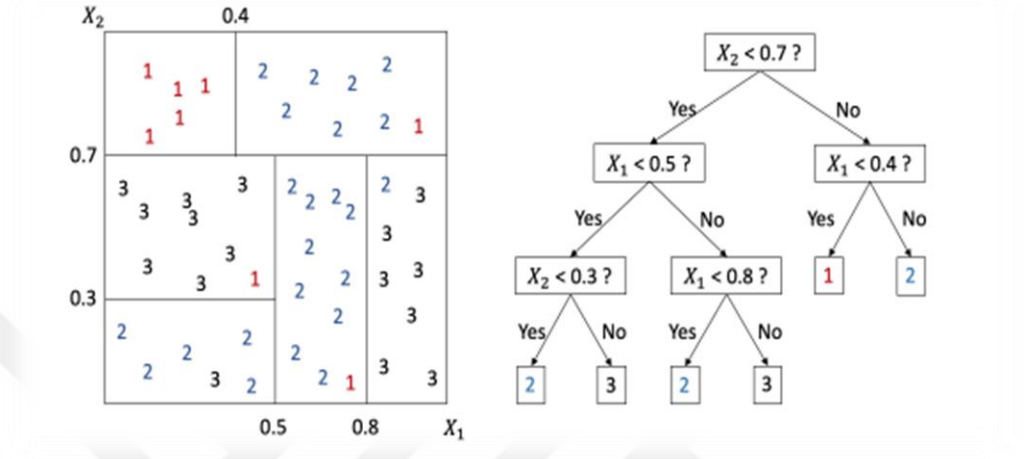
Karar Ağaçları, sınıflandırma ve regresyon için kullanılan parametrik olmayan bir Denetimli Öğrenme yöntemidir. Bu yöntemin amacı, verilerin özelliklerinden çıkarılan basit karar kurallarını öğrenerek hedef değişkenin değerini tahmin eden bir model oluşturmaktır (Scikit Learn, 2020,Şubat 28). Karar Ağaçları, büyük veri kümelerinden verimli bir şekilde öğrenilebilen, doğru tahminler üretebilen bir sınıflandırma türüdür. (Schietgat vd., 2010).

Karar Ağaçları, hiyerarşik şekilde düzenlenen ve grafiksel ağaç olarak gösterilen bir dizi sorudur. Belirli bir giriş nesnesi için bir Karar Ağacı, bilinen özellikler hakkında art arda sorular sorarak nesnenin bilinmeyen bir özelliğini tahmin eder. Daha sonra hangi sorunun sorulacağı önceki sorunun cevabına bağlıdır ve bu ilişki grafiksel olarak nesnenin izlediği ağaçta bir yol olarak temsil edilir. (Criminisi, Shotton, & Konukoglu, 2012).

Sorguların sayısı ve niteliği nedeniyle, sabit uzunluklu bir özellik vektörüne dayanan standart karar ağacı yapımı mümkün değildir. Bunun yerine, her bir düğümde yalnızca küçük bir rastgele sorgu örneği tutulur. Ağaç derinliğiyle birlikte

büyümek ve birden fazla ağaç üretmek için karmaşıklıklar kısıtlanır (Amit & Geman, 1997).

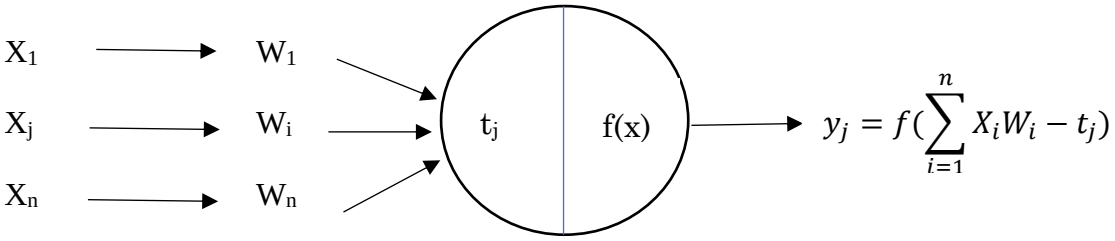
Karar ağacında, ağaçtaki her uç düğüm için genellikle ortalama veya çoğunluk oyu ile karar verilen uygun bir değer atanır. Şekil 2.2 sınıflandırma ağacının bir örneği olarak gösterilebilir (Zhou, 2019).



Şekil 2.2. İki boyutlu ortak değişken alan üzerindeki sınıflandırmayı öngören bir sınıflandırma ağacı

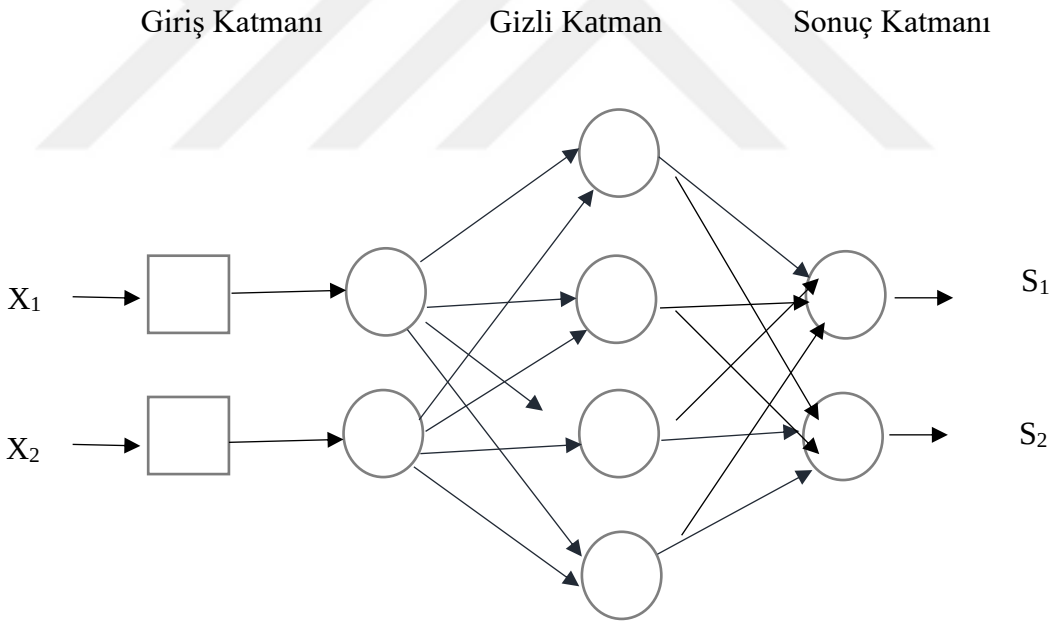
2.2.1.1.3. Yapay sinir ağları

Yapay Sinir Ağları (YSA), beynin çalışma şeklinden esinlenen sinir hücresi ağlarını taklit ederek hesaplama yapmaya çalışan, doğrusal olmayan bir modeldir (Churpek vd., 2016). YSA, yapay sinir hücrelerinin katmanlarla bağlanmasıyla oluşturulan veri tabanına bağlı sistemlerdir. YSA, insan beyninin öğrenme ve değişik koşullar altında hızlı karar verebilme gibi yeteneklerini, basitleştirilmiş modeller yardımıyla çözmeyi amaçlamaktadır (Hamzaçebi & Kutay, 2004). YSA, kısmi en küçük kareler gibi sonuç gözlemlenmeyen bir ara değişken seti tarafından modellenir. Burada ara değişken setlerine gizli katman veya gizli değişken denir. Belirtilen gizli değişkenler orijinal tahmin edicilerin doğrusal kombinasyonlarıdır ve kısmi en küçük kareler modelinin aksine, hiyerarşik bir şekilde tahmin edilemezler (Kuhn & Johnson, 2013). YSA, yapısal olarak 5 unsurdan oluşmaktadır. Bu unsurlar; Girdiler, Ağırlıklar, Toplam fonksiyonu, Aktivasyon fonksiyonu ve Çıktıdır (Kalogirou, 1999).



Şekil 2.3. Sinir ağı mimarisi

Burada, W_i ağırlık faktörleri, t_j sinapslar ve sınır değerlerdir. YSA da, ağırlık faktörünün etkisine bağlı olarak hücreye gelen uyarımlar (X_1, X_i, ... X_n) hücre içi denge durumu veya sınır değer (t_j) de dikkate alınarak doğrusal olmayan bir aktivasyon fonksiyonu yardımıyla çıktı şeklinde sonuçlara (y_j) dönüştürülmektedir (Koç, Balas, & Arslan, 2004). Sinir ağı Şekil 2.4' te gösterilmiştir (Bose & Garga, 1993).



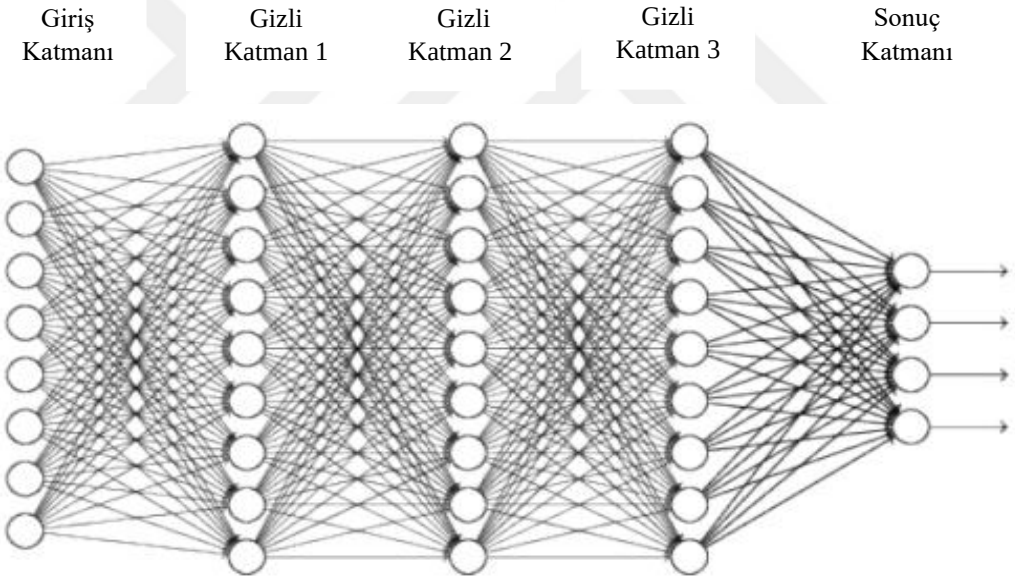
Şekil 2.4. Sinir ağı

2.2.1.1.4. Derin sinir ağları

Derin sinir ağları (DSA) isteğe bağlı olarak derin düşünce ağı ön eğitim algoritması kullanılarak başlatılan birçok gizli katmana sahip geleneksel çok katmanlı bir algılayıcıdır (Seide, Li, & Yu, 2011). DSA, konuşma ve görüntü tanıma gibi bilişsel görevlerde başarı göstermiş bir modeldir. (Jerry vd., 2017).

Büyük DSA modelleri çok güçlüdür, fakat büyük miktarda enerji tüketmektedir. Çünkü modelin harici DRAM(Dinamik Rastgele Erişimli Bellek)'de saklanması gereklidir ve her görüntünün, kelimenin veya konuşmanın örneği için her defasında DRAM'dan getirilmelidir (Han vd., 2016).

Tam bağlı bir DSA'nın, eğitimini hızlandırmak için yonga üzerindeki depolama kullanılmalı ve tüm düğümlerin birbirine bağlandığı dizide aynı düğümde veri hareketleri en aza indirilmelidir (Jerry vd., 2017). Şekil 2.5 derin sinir ağlarının bir örneği olarak gösterilebilir (Nielsen, 2015).



Şekil 2.5. Derin sinir ağları

YSA'da derin mimariler kullanan DSA'lar, tek bir katmandaki katman ve birim sayısı artırıldığında daha yüksek karmaşıklığa sahip işlevleri temsil edebilir. Yeterli ve uygun eğitim veri setleri ve modeller göz önüne alındığında, derin öğrenme yaklaşımları insanların işlem kolaylığı için haritalama işlevleri oluşturmalarına yardımcı olabilir (Liu vd., 2017).

2.2.1.2. Sınıflandırma

Makine Öğrenmesi'nin temel birimi olan sınıflandırma yönteminin amacı, bilinmeyen bir veri parçasını bilinen bir gruba atamaktır (Harrington, 2012). En yaygın kullanılan sınıflandırma algoritmalarından biri, popülerliği ve ikili sınıflandırma problemlerinde maksimum marj ayrımını garanti ettiğinden Destek Vektör Makinesi (SVM)'dir (Criminisi vd., 2012).

2.2.1.2.1. Destek vektör makinesi

Destek Vektör Makinesi (SVM) teorisi yapısal risk minimizasyonu fikrine dayanmaktadır. SVM'nin iyi genelleme yeteneği, genellikle büyük bir marjın varlığı ile açıklanmaktadır (Vapnik & Chapelle, 2000). SVM; optimal bir ayırma ve hiper düzlemi bulmak için orijinal giriş alanını daha yüksek boyutlu bir özellik alanına dönüştüren,v istatistiksel öğrenme teorisine dayanan makine öğrenmesi yöntemidir (Lin & Wang, 2002). Birçok uygulamada SVM'nin sıklıkla kullanılan diğer makine öğrenme yöntemlerinden daha yüksek performans sağladığı görülmüştür, aynı zamanda sınıflandırma problemlerini çözmek için de güçlü araçlar olarak tanıtılmıştır (Pang, Lee, & Vaithyanathan, 2002).

Radyal Temel İşlevi (RBF ya da Gauss Kernel), SVM eğitiminde en çok kullanılan çekirdektir.

$$K(x_i, x_j) = e^{\frac{-\|x_i - x_j\|^2}{2\sigma^2}} \tag{2.7}$$

Burada x_i, x_j girdi uzayının öznitelik vektörlerini temsil etmektedir. İki öznitelik vektörü arasındaki Öklid mesafesi $-\|x_i - x_j\|$ olarak tanımlanmaktadır. Serbest parametre, σ olarak tanımlanmıştır (Chang, Hsieh, Chang, Ringgaard, & Lin, 2010).

2.2.1.2.2. Naive bayes sınıflandırıcı

Makine Öğrenmesinde araştırmaların bir diğer amacı da algoritmaları ve etki alanı özelliklerini davranışla ilişkilendiren ilkeleri keşfetmektir. Bu amaçla, birçok araştırmacı ampirik düzenlilik arayışında doğal ve yapay alanlarla sistematik deneyler gerçekleştirmiştir (Langley, Iba, & Thompson, 1992). Naive Bayes (NB),

yaygın olarak kullanılan Bayes Formülü veya Bayes teorimi olarak bilinen olasılık teoremi ile sağlanır (Lewis, 1998).

$$P(A\backslash B) = \frac{P(B\backslash A) \times P(A)}{P(B)} \quad (2.8)$$

P(A): A olayının gerçekleşme olasılığı

P(B): B olayının gerçekleşme olasılığı

P(A\B): B olayı gerçekleştiğinde A olayının gerçekleşme olasılığı

P(B\A): A olayı gerçekleştiğinde B olayının gerçekleşme olasılığı (Zhang, 2016).

NB sınıflandırıcısı; özelliklerin, verilen sınıftan bağımsız olduğunu varsayarak öğrenmeyi büyük ölçüde basitleştirmektedir. Bağımsızlık, genel olarak zayıf bir varsayım olsa da NB pratikte genellikle daha karmaşık sınıflandırıcılarla rekabet ettiği bilinmektedir (Rish, 2001).

2.2.2. Denetimsiz öğrenme (Unsupervised learning)

Denetimsiz Öğrenmede, Denetimli Öğrenmenin aksine herhangi bir süpervizör bulunmamaktadır. Sadece girilen veriler bulunur. Denetimsiz öğrenmenin temel amacı, girdilerde bir düzen bulmaktır (Alpaydin, 2014).

2.2.2.1. Kümeleme (Clustering)

Kümeleme; birbirine benzer özellik gösteren verilerin, özelliklerine göre homojen alt gruplara dağıtılmasına denir. Kümelemede elde edilen kümelerin benzerlikleri fazla, kümeler arası benzerlikler ise azdır (Alpar, 2017).

2.2.2.1.1. K-ortalama kümeleme

K-ortalama kümeleme, kümeleme hatasını en aza indiren popüler bir kümeleme yöntemidir. Ancak, K-ortalama algoritması kısmi bir arama prosedürüdür ve performansı büyük ölçüde başlangıç koşullarına bağlı olduğundan dezavantajlara da sahiptir (Likas, Vlassis, & Verbeek, 2003).

2.2.2.1.2. Temel bileşen analizi (PCA-Principal component analysis)

Temel Bileşen Analizi (PCA), yaygın kullanılan çok değişkenli istatistik yöntemidir. Sinyal işlemede, sinyallerin yararlı bir belirleyicisidir. PCA'nın, kabaca sinyal alt uzayında bulunduğu varsayılmaktır. Sinyal modelleme, spektrum kestirimi ve dizi işlemenin çeşitli modern yöntemleri bu kavrama dayanmaktadır (Karhunen & Joutsensalo, 1995).

Bu yöntemde karşılıklı bağımlılık yapısı gösteren, ölçüm sayısı n olan p adet değişken; doğrusal, ortogonal ve birbirinden bağımsız olan k tane yeni değişkene dönüştürülmektedir (Yıldız, Çamurcu, & Doğan, 2010).

2.2.2.1.3. Spektral kümeleme

Spektral Kümeleme, günümüzde kullanılan en popüler modern kümeleme algoritmalarından biridir. Uygulaması basittir ve çoğu zaman k-ortalama algoritması gibi geleneksel kümeleme algoritmalarından daha iyi performans gösterir (Von Luxburg, 2007).

Spektral Kümeleme, cebirsel grafik teorisine dayanan bir kümeleme yöntemidir ve verilerin evrensel yapısı hakkında herhangi bir varsayımda bulunmaz. Spektral Kümeleme verileri evrensel optimum seviyeye yakınsak olabilir, özellikle dışbükey olmayan veri setleri için uygun olan gelişigüzel şeklin alanı için iyi performans gösterir (Jia, Ding, Xu, & Nie, 2014).

2.2.2.1.4. Dirichlet süreci

Dirichlet Süreçleri, popüler bir Bayesian parametrik olamayan model sınıfıdır. Dirichlet süreçleri; yoğunluk tahmini, yarı parametrik modelleme, sidestepping model seçimi ve ortalaması için kullanılmaktadır (Teh, 2010).

2.2.2.1.5. Konvolüsyon sinir ağları

Konvolüsyon Sinir Ağları (KSA), değişken ve karmaşık sinyallerle kullanım için özel olarak tasarlanmıştır ve diğer tekniklerden daha iyi performans göstermektedir. KSA; görüntü sınıflandırma, büyük ölçekli konuşma görevleri, trafik işareti tanıma ve diğer çeşitli uygulamalarda yaygın olarak kullanılmaktadır (Guo, Chen, & Shen, 2016).

2.2.2.1.6. Derin inanç ağı

Derin İnanç Ağları; stokastik gizli değişkenlerin çoklu katmanlarından oluşan olasılıksal üretken modellerdir. Gizli değişkenler tipik olarak ikili değerlere sahiptir ve gizli birimler ya da özellik dedektörleri olarak adlandırılır. En üstteki iki katman, aralarında simetrik olmayan bağlantılara sahiptir ve bu katmanlar ilişkilendirilebilir bir bellek oluşturur. Alt katmanlar, yukarıdaki katmandan aşağı katmana olacak şekilde yönlendirilmiş bağlantılar alırlar. En alt katmandaki birimlerin durumları bir veri vektörünü temsil etmektedir (Hinton, 2009).

2.2.2.2. Boyutsal indirgeme

Konuşma sinyalleri, dijital fotoğraflar veya fMRI taramaları gibi veriler genellikle yüksek bir boyutluluğa sahiptir. Bu tür verileri yeterince işleyebilmek için verilerinin boyutunun indirgenmesi gerekir. Boyut indirgenmesi, yüksek boyutlu verilerin indirgenmiş boyutun anlamlı bir temsiline dönüştürülmesidir. Verilerin gerçek boyutsallığı, verilerin gözlenen özelliklerini hesaba katmak için gereken minimum parametre sayısıdır.

Boyutsallığın indirgenmesi birçok alanda önemlidir, çünkü boyutsallığın zorluğunu ve yüksek boyutlu alanların diğer istenmeyen özelliklerini azaltır (Van Der Maaten, Postma, & Van den Herik, 2009).

2.2.2.2.1. İlişkili belleğe dayalı sinir ağları

İlişkili Belleğe Dayalı Sinir Ağları, X ve Y katmanı olmak üzere iki katman halinde düzenlenmiş nöronlardan oluşmaktadır. Bir katmandaki nöronlar, diğer katmandaki nöronlara tam olarak bağlanırken, aynı katmanda ki nöronlar arasında hiçbir bağlantı yoktur. İki katman arasındaki ileri ve geri bilgi akışlarının tekrarlanması yoluyla, depolanan iki kutuplu vektör çiftleri için iki yönlü bir ilişkili arama gerçekleştirir (Zhao, 2002).

2.2.2.2.2. Isomap

Isomap, ağırlıklı bir grafik üzerindeki jeodezik mesafelerinin klasik ölçeklendirme ile birleştirildiği, yaygın olarak kullanılan düşük boyutlu gömme yöntemlerinden biridir.

İki düğüm arasındaki en kısa yol boyunca kenar ağırlıklarının toplamı jeodezik mesafe olarak atanır. Jeodezik mesafe matrisinin en üstteki öz vektörleri, n-boyutlu Öklid uzayındaki koordinatları temsil eder (Choi & Choi, 2007).

2.2.2.2.3. Kısmi doğrusal gömme

Kısmi Doğrusal Gömme (KDG), yüksek boyutlu girdilerin en yakın düşük boyutlara yerleştirilmesini hesaplayan denetimsiz bir öğrenme algoritmasıdır.

KDG'nin kısmi boyutsallığı azaltmaya yönelik kümeleme yöntemlerinden farkı, girdilerini daha düşük boyuta sahip tek bir küresel koordinat sistemine eşler ve optimizasyonları kısmi minimum değeri içermez. KDG, lineer rekonstrüksiyonların (tekrar yapılanma) kısmi simetrilerinden yararlanarak, yüz görüntülerini veya metin belgeleri gibi doğrusal olmayan çeşitli küresel yapıları öğrenebilir (Roweis & Saul, 2000).

2.2.2.2.4. Yığılı oto enkoder

Derin sinir ağı mimarilerinin bir kolu olan Yığılı Oto Enkoder, birçok alanda önemli rol oynamaktadır. Yığılı Oto Enkoder'in temel amacı, başlangıç örnekleminden özellikler saptamak ve bu temel özellikler aracılığıyla modeli ifade etmektir. Yığılı Oto Enkoder; bir giriş katmanı, bir çıkış katmanı ve bunları bağlayan bir veya daha fazla gizli katman ile çok katmanlı algılayıcıya çok benzeyen, ileri beslemeli, tekrar etmeyen bir sinir ağıdır (Tao, Zhang, Yang, Wang, & Lu, 2015).

2.2.2.3. Yoğunluk tahmini

Yoğunluk tahmini, arkeoloji, bankacılık, klimatoloji, ekonomi, genetik, hidroloji ve fizyoloji gibi birçok alanda uygulanmaktadır (Sheather, 2004).

2.2.2.3.1. Boltzmann makinesi

Boltzmann Makinesi (BM); görünür bir değişken katmanının dağılımını modellemek için gizli ikili değişkenler veya birimler katmanı kullanan bir olasılık modelidir (Sutskever, Hinton, & Taylor, 2009).

Görüntü ve metin gibi yüksek boyutlu verileri içeren değişkenlere başarıyla uygulanmıştır. Bu nedenle, genellikle iki yaklaşım izlenmektedir. Öncelikle bir BM, girişlerin dağılımını modellemek için denetimsiz bir şekilde eğitilir (birden fazla BM de eğitilebilir ve bunlar birbiri üstüne kümelenebilir). Daha sonra, 2 yol izlenebilir. Ya gizli katmanı, giriş verisini gizli katman tarafından verilen ifade ile değiştirerek önişleme için kullanılır ya da BM'nin parametreleri bir ileri beslemeli sinir ağını başlatmak için kullanılır. Her iki durumda da BM eldeki denetimli öğrenme problemini çözmek için başka bir öğrenme algoritması (önceden işlenmiş girişleri

veya sinir ağı kullanan sınıflandırıcı) ile eşleştirilmelidir (Fiore, Palmieri, Castiglione, & De Santis, 2013).

2.2.2.3.2. Kernel yoğunluk tahmini

Kernel Yoğunluk Tahmini, verilerin istatistiksel analizinde kullanılan önemli bir araçtır. Parametrik olmayan bir tahmin türü olan yoğunluk tahmini, verilerin dağılımındaki çok modluluk, çarpıklık gibi yapıları değerlendirmek için kullanılabilir. Ayrıca Bayes posterior özetlenmesi, sınıflandırılması ve ayırt edici analizi için de kullanılabilir (Botev, Grotowski, & Kroese, 2010).

2.2.2.3.3. Gausssian karışım modeli

Gaussian Karışım Modeli (GMM), Gauss bileşen yoğunluğunun ağırlıklı toplamı olarak temsil edilen parametrik bir olasılık yoğunluk fonksiyonudur. GMM'ler biyometrik sistemlerde, özellikle de hoparlör tanıma sistemlerinde, geniş bir örnekleme sınıfını temsil ettikleri için sıklıkla kullanılır. GMM'nin güçlü özelliklerinden biri de keyfi olarak şekillendirilmiş yoğunluklara kolay yaklaşımlar oluşturma yeteneğidir.

Klasik tek modlu Gauss modeli, bir konuma (ortalama vektör) ve eliptik bir şekle (kovaryans matrisi) göre özellik dağılımlarını gösterir ve bir vektör öbeği veya en yakın komşu modeli, ayrı bir karakteristik şablonlar kümesinin dağılımını temsil eder (Reynolds, 2009).

2.2.2.3.4. Derin boltzmann makinesi

Derin Boltzmann Makineleri (DBM) de derin inanç ağları gibi, git gide karmaşıklaşan bir öğrenme potansiyeline sahiptir. Bu nedenle nesne ve konuşma tanıma sorunlarını çözmenin umut verici bir yolu olarak kabul edilir. Derin inanç ağlarından farklı olarak DBM, yaklaşık çıkarım prosedüründe aşağıdan yukarıya geçişe ek olarak, yukarıdan aşağıya geri bildirimi içerebilir ve DBM'nin belirsiz girdilerle ilgili belirsizliği daha iyi yaymasını ve dolayısıyla daha sağlam bir şekilde başa çıkmasını sağlar (Salakhutdinov & Hinton, 2009).

2.2.3. Pekiştirmeli öğrenme (Reinforcement learning)

Pekiştirmeli Öğrenme'nin temel amacı, sayısal karşılığı olan bir sinyali en üst düzeye çıkarmak için ne yapılacağını öğrenmek ve durumların eylemlerle nasıl eşleştirileceğini öğrenmektir. Pekiştirmeli Öğrenme'nin, Denetimli Öğrenmeden

farkı Denetimli Öğrenme'de problemler tek adım iken Pekiştirmeli Öğrenme'de ise birden çok adım bulunmaktadır (Faustino, 2011).

2.2.3.1. Tekrarlama ilkesine dayalı pekiştirmeli öğrenme

Tekrarlama ilkesi algoritması, Markov karar süreci için ortalama maliyet ve optimal kontrol problemine bir çözüm oluşturmak için kullanılan tekrarlı bir prosedürdür (Meyn, 1997).

2.2.3.1.1. Monte carlo metodu

Monte Carlo Metodu, [0, 1] aralığında rasgele düzgün dağılmış bir sayı üreteciyle, varyans minimizasyonu ve çapraz entropi yöntemleri adı verilen kümülatif dağılım fonksiyonu yaklaşımlarını kullanarak bir olasılık değişkeninin yapay değerlerini üretir. Örnekleme algoritmalarını sıralı bir şekilde oluşturur. Bu algoritmaların daha da geliştirilmesi, başarılı yolların yeniden örneklenmesi ile elde edilir ve ardışık öneme sahip yeniden örnekleme algoritmaları ortaya çıkar (R. Y. Rubinstein & Kroese, 2016).

2.2.3.1.2. Zamansal fark yöntemi

Zamansal Fark Yöntemleri, pekiştirmeli öğrenme problemlerini çözmek için kullanılan popüler bir yöntemdir. Zamansal Fark Yöntemlerinin temel fikri, öğrenmenin zamansal olarak ardışık tahminler arasındaki farka dayandırılmasıdır. Başka bir deyişle öğrenmenin amacı, süpervizörün mevcut girdi modeli için mevcut tahminini bir sonraki adımda bir sonraki tahminle daha yakından eşleştirmektir (Tesauro, 1995).

2.2.3.2. Değer tekrarına dayalı pekiştirmeli öğrenme

Değer Tekrarına Dayalı Pekiştirmeli Öğrenme, planlamayı öğrenebilir ve planlamaya dayalı akıl yürütmeyi içeren sonuçları tahmin etmeye uygundur (Tamar, Wu, Thomas, Levine, & Abbeel, 2016).

2.2.3.2.1. Q öğrenme

Q-öğrenme, Pekiştirmeli Öğrenmenin modelsiz bir şeklidir. Ayrıca asenkron dinamik programlama (DP) yöntemi olarak da görülebilir. Etmenlerin, etki alanlarının haritalarını oluşturmadan, eylemlerin sonuçlarını deneyimleyerek Markovian etki alanlarında en iyi şekilde hareket ederek öğrenme yeteneklerini geliştirmelerini sağlar (Watkins & Dayan, 1992).

2.2.3.2.2. Derin q-ağları

Bir tür pekiştirmeli öğrenme modeli olan Derin Q-Ağları (DQN), bir ortamla etkileşime girerken en uygun eylemleri elde eden bir etmeni eğitmeyi hedefler. Bu model, birçok Atari 2600 oyununda profesyonel insan oyuncuları aşabilme yeteneği ile bilinir. İnsanüstü performansına rağmen, modelin derinlemesine anlaşılması ve DQN etmeninin karmaşık davranışlarının yorumlanması, uzun süreli model eğitim süreci nedeniyle zor uygulanmaktadır (Wang, Gou, Shen, & Yang, 2018).

2.2.3.2.3. Çift-q öğrenme

Standart Q-öğrenmede oluşan fazla tahminlerin üstesinden gelmek için Van Hassel tarafından 2010 yılında Çift-Q öğrenme algoritması ileri sürülmüştür (Hasselt, 2010). Çift-Q öğrenmesinin temel amacı, beklenen Q değerleri üzerindeki yakınlığın doğruluğunu artırmak için iki tahmin edici uygulamaktır (Huang, Lin, & Zhang, 2017).

3. GEREÇLER VE YÖNTEMLER

3.1. Gereçler

Bu çalışmada, 500 satır ve 9 sütundan oluşan veri seti ile R programında Gail Modeli ile makine öğrenmesi yöntemleri karşılaştırılmıştır. Veri seti senaryosu R programında elde edilmiştir. Her değişken için 500 veri türetilmiştir.

Türetilen verilerin meme taraması için ilk başvuru yaşı (T1), 23 ve 83 yaşları arasında dağılım göstermekledir.

Türetilen verilerin risk tahmin yaşı (T2) ise ilk başvuru yaşının 5 yıl sonrası olarak türetilmiştir.

Türetilen verilerin ırk dağılımı, Amerika Birleşik Devletleri Nüfus Sayım Bürosu (United States Census Bureau) sitesindeki ırk dağılım oranları baz alınarak Beyaz, Afro-Amerikan, Latin Amerikalı gibi ırklar dahil olmak üzere toplamda 11 ırkın dağılımı elde edilmiştir (United States Census Bureau, 2020, July 20).

Türetilen verilerin menarş yaşı da ırklara göre belirlenmiştir (Palmer, Rosenberg, Wise, Horton, & Adams-Campbell, 2003). Irklara göre menarş yaşı için ırkların menapoz ve menarş yaşları üzerine yapılan çalışmalar referans alınmıştır (Ahuja, 2016).

Türetilen verilerin ilk canlı doğum yaşı, Statista sitesindeki ırkların ilk doğum yaşları dağılım oranlarına göre elde edilmiştir (Statista, 2020, July 20).

Verilerin analizinde R Studio Version 3.6.3 programı kullanılmıştır.

Tablo 3.1. Veri seti senaryosu

Değişken	**Verinin Açıklaması**	**Verinin Özellikleri**
ID	Hastanın kimliği	1, 2…, 499, 500
Risk	Meme kanseri riski	0, 1
T1	İlk andaki yaş	23, 24…, 82, 83
T2	Tahmin yaşı	T1+5
N_Biop	Biyopsi sayısı	0, 1, 2, 3
HypPlas	Hiperplazi	0, 1, 99
AgeMen	Menarş yaşı	9, 10…,15
Age1st	İlk doğum yaşı	18, 19…, 33
N_Rels	Hasta Akraba Sayısı	0, 1, 2, 3
Race	Irk	1, 2…, 10, 11

3.2. Yöntem

Bu çalışmada Gail Modeli ile makine öğrenmesi yöntemlerinin meme kanseri risk değerlendirmesinde karşılaştırılması amaçlanmıştır. İlk olarak veri setine Gail Modeli uygulanmış ve risk faktörü belirlenmiştir (Risk faktörü Gail Modeline göre riski 1.66'dan büyük olanlar için 1, küçük olanlar için 0 olarak belirlenmiştir). Daha sonra aynı veri setine makine öğrenmesi algoritmaları uygulanmış ve risk tahmin sonuçları karşılaştırılmıştır.

3.2.1. Gail modelinin uygulanması

Veri setine aşağıdaki adımlar uygulanarak istenilen sonuçlar elde edilmiştir.

1. **Adım:** Veri seti R programında türetilmiştir.
2. **Adım:** Aşağıdaki kodlar ile veri setine Gail Modeli uygulanmıştır.

VERİ SETİNE GAIL MODELİNİN UYGULANMASI

```
install.packages("BCRA")          #Uygun paketin yüklenmesi
library(BCRA)                     #Uygun kütüphanenin seçilmesi
absolute.risk(data, 1,0)          #invazif meme kanseri riskinin hesaplanması
```

3.2.2. Makine öğrenmesi algoritmalarının uygulanması

Bu çalışmada, Gail Modeli'nin sonuçlarına göre risk grubu "0,1" olmak üzere 2 gruba ayrılmıştır. Sıfır değeri; Gail Modeli'ne göre "az riskli, risksiz", 1 değeri ise "riskli, yüksek riskli" olarak tanımlanmıştır. Risk Faktörü için ayrı bir sütun oluşturulmuş ve Gail Modeli'nin sonuçlarından elde edilen "0,1" değerleri bu sütuna aktarılmıştır.

Türetilen veriler makine öğrenmesi algoritmalarına sunulmadan önce min-max normalizasyon işlemi gerçekleştirilmiş ve veriler [0,1] aralığında yeniden ölçeklendirilmiştir (Jain, Nandakumar, & Ross, 2005).

$$Normalizasyon = \frac{x - \min(x)}{\max(x) - \min(x)} \tag{3.1}$$

Normalizasyon işleminin amacı, verideki aşırı salınımları engellemek ve sistem performansını arttırmaktır. Normalizasyon ile elde edilen yeni veri setinde T1 değişkeni, T2 değişkeni (T1+5) ile aynı ölçekte olduğundan T2 değişkeni makine öğrenmesi algoritmalarına dahil edilmemiştir.

Makine öğrenmesi algoritmaları hedef değişkeni (target), Risk sütunu olarak belirlenmiştir. Eğitim ve test veri seti oluşturulmuştur. Sınıflandırma performansının karşılaştırılması için %70 eğitim ve %30 test, %80 eğitim ve %20 test olmak üzere 2 ayrı eğitim ve test veri seti oluşturulmuştur.

EĞİTİM VE TEST SETLERİNİN OLUŞTURULMASI

```
set.seed(111)
orneklem <- sample(1:nrow(veri), size nrow(veri)*0.70,replace = FALSE)
eğitim <- veri[orneklem,]
test <-  veri[-orneklem,]
```

3.2.2.1. k-en yakın komşu algoritması

Veri setine aşağıdaki adımlar uygulanarak istenilen sonuçlar elde edilmiştir.

1. **Adım:** k değerinin, optimum değeri için algoritma oluşturulmuş ve doğruluk yüzdesine göre k optimal değeri seçilmiştir.
2. **Adım:** k değerleri için grafik çizdirilmiştir.
3. **Adım:** k-NN algoritması uygulanmış ve sınıflandırma sonuçları elde edilmiştir.

VERİ SETİNE k-NN ALGORİTMASININ UYGULANMASI

```
library(class)
library(caret)      #ilgili kütüphanelerin yüklenmesi

knn.k<-knn(train=train,test=test.,cl=train_labels$Target,
k=k-degeri )     #k-nn algoritması
knn.cm <-confusionMatrix(table(knn.k,test.data$Target)) #Confusion matrisi
```

3.2.2.2. Yapay sinir ağları algoritması

Veri setine aşağıdaki adımlar uygulanarak istenilen sonuçlar elde edilmiştir.

1. **Adım:** Sinir ağında optimal sonuç elde etmek ve ilişkili değişkenleri tespit etmek için korelasyon grafiği oluşturulmuştur.
2. **Adım:** YSA algoritması uygulanmış ve sınıflandırma sonuçları elde edilmiştir.
3. **Adım:** YSA grafiği çizdirilmiştir.

VERİ SETİNE YSA ALGORİTMASININ UYGULANMASI

```
library(e1071)
library(caret)
library(NeuralNetTools)
library(nnet)
library(boot)                          #ilgili kütüphanelerin yüklenmesi
library(corrplot)
library(neuralnet)
library(dplyr)
corrTrain <- cor(egitim[,])
corrplot.mixed(corrTrain)              # Korelasyon Grafiğinin çizdirilmesi
ysa= neuralnet(Risk~., data=egitim, hidden=6,act.fct = "logistic",
linear.output = FALSE, algorithm = "rprop+")       #YSA algoritması
plotnet(ysa)
plot(ysa)                                          #YSA grafiği
```

3.2.2.3. Destek vektör makinesi algoritması

Veri setine aşağıdaki adımlar uygulanarak istenilen sonuçlar elde edilmiştir.

1. **Adım:** SVM algoritması uygulanmış ve sınıflandırma sonuçları elde edilmiştir.

VERİ SETİNE SVM ALGORİTMASININ UYGULANMASI

```
svm.sonuc <- svm(Risk~., data=egitim,
          kernel="radial",
          cost=100, scale = FALSE,
          gamma = if (is.vector(svm.train)) 1 else 1 / ncol(egitim),
          coef0 = 0,)                                    #SVM algoritması
print(svm.sonuc)
svm.tahmin= predict(svm.sonuc, svm.test,type="class" )
svm.cm= confusionMatrix(table(svm.tahmin,svm.test$Risk))
svm.cm                                                   #Confusion Matrisi
```

3.2.2.4. Naive bayes algoritması

Veri setine aşağıdaki adımlar uygulanarak istenilen sonuçlar elde edilmiştir.

1. **Adım:** Naive Bayes algoritması uygulanmış ve sınıflandırma sonuçları elde edilmiştir.

VERİ SETİNE NAIVE BAYES ALGORİTMASININ UYGULANMASI

```
library(e1071)
parametre <- function(resp, dt) {
 ad <- names(dt)
 frml_ <- as.formula(paste(resp, "~", paste(ad[!ad %in% resp], collapse = " + ")))
 frml_}
frml_<- parametre ("Risk", veri[,])
NB.sonuc=naiveBayes(frml_, data=egitim,
    threshold = 0.001, eps = 0.1)                    #Naive Bayes algoritması
print(NB.sonuc)
nb.tahmin <- predict(NB.sonuc, nb.test)
nb.cm <- confusionMatrix(table(nb.tahmin, nb.test$Risk))
ncm
```

3.3. Performans Değerlendirme Ölçütleri

Sınıflandırma işleminde veri seti, eğitim seti ve test seti olmak üzere 2 gruba ayrılmıştır.

Makine Öğrenmesi yöntemlerinin sınıflandırma başarısını değerlendirmek için doğruluk, Kappa, duyarlılık, özgüllük ve ROC eğrisi ölçütleri kullanılmıştır.

Doğruluk (Accuracy): Sınıflandırma modellerini değerlendirmek için kullanılan bir ölçü birimidir.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \times 100 \qquad (3.1)$$

TP: Doğru sınıflandırılan yüksek riskli birey

TN: Doğru sınıflandırılan düşük riskli birey

FP: Hatalı sınıflandırılan düşük riskli birey

FN: Hatalı sınıflandırılan yüksek riskli birey

Kappa: Hem değerlendiriciler arası hem de araştırmacılar arası güvenilirlik testi için yararlı olan güçlü bir istatistiktir. Korelasyon katsayıları gibi, -1 ile +1 arasında değişebilir. Tüm korelasyon istatistiklerinde olduğu gibi, Kappa da standartlaştırılmış bir değerdir ve bu nedenle birden çok çalışmada aynı şekilde yorumlanır (McHugh, 2012).

$$\kappa = \frac{Pr(a) - Pr(e)}{1 - Pr(e)} \tag{3.2}$$

Pr(a): Gerçek uyum

Pr(e): Rasgele uyum

Duyarlılık (Sensitivity): Gerçekte riskli olan bireylerin riskli olarak tanımlanması yeteneğini ifade etmektedir (Lalkhen & McCluskey, 2008).

$$Sensivity = \frac{TP}{TP + FN} \tag{3.3}$$

TP: Doğru sınıflandırılan yüksek riskli birey

FN: Hatalı sınıflandırılan yüksek riskli birey

Özgüllük (Specificity): Gerçekte düşük riskli olan bireylerin düşük riskli olarak tanımlanması yeteneğini ifade etmektedir (Lalkhen & McCluskey, 2008).

$$Specificity = \frac{TN}{TN + FP} \tag{3.4}$$

TN: Doğru sınıflandırılan düşük riskli birey

FP: Hatalı sınıflandırılan düşük riskli birey

ROC Eğrisi: Duyarlılığın özgüllüğe olan oranıyla hesaplanmaktadır (Hanley & McNeil, 1982).

$$ROC = \frac{Sensivity}{Specificity} \tag{3.5}$$

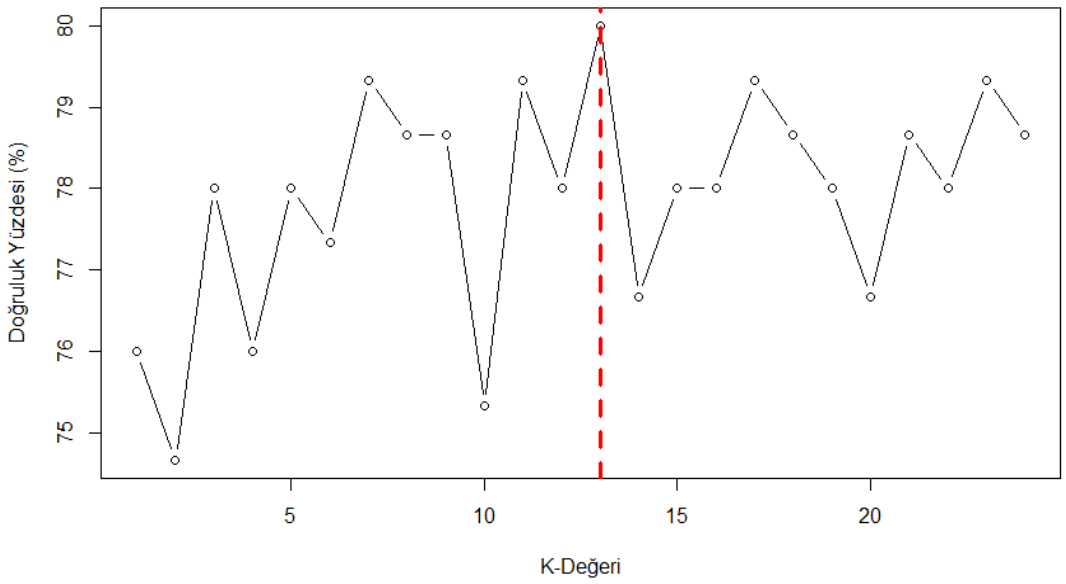
4. BULGULAR

Bu çalışmada Risk değişkeni Gail Modeli ile hesaplanmış, daha sonra elde edilen Risk değişkeni makine öğrenmesi algoritmaları için hedef olarak belirlenmiştir. %70 eğitim ve %30 test, %80 eğitim ve %20 test olmak üzere 2 ayrı eğitim ve test veri seti oluşturulmuştur.

4.1. %70 Eğitim ve %30 Test Veri Seti İçin Bulgular

4.1.1. k-NN sınıflandırma sonucuna ilişkin bulgular

Veri setine ilk önce k-NN algoritması uygulanmıştır. k değerinin optimal sonucu için doğruluk yüzdesi elde edilen Şekil 4.1'de belirtilmiştir**.**



Şekil 4.1. %70 eğitim ve %30 test veri seti için k-değeri optimal doğruluk yüzdesi sonucu

Şekil 4.1'e göre doğruluk yüzdesi en yüksek olan k değeri 13'tür. Algoritmaya k değeri 13 seçilerek devam edilmiştir. Tablo 4.1'de k-NN algoritmasının parametreleri verilmiştir. Sınıflandırma sonuçları Tablo 4.2'de belirtilmiştir.

Tablo 4.1. k-NN algoritma parametreleri

k	13
prob	TRUE
l	0

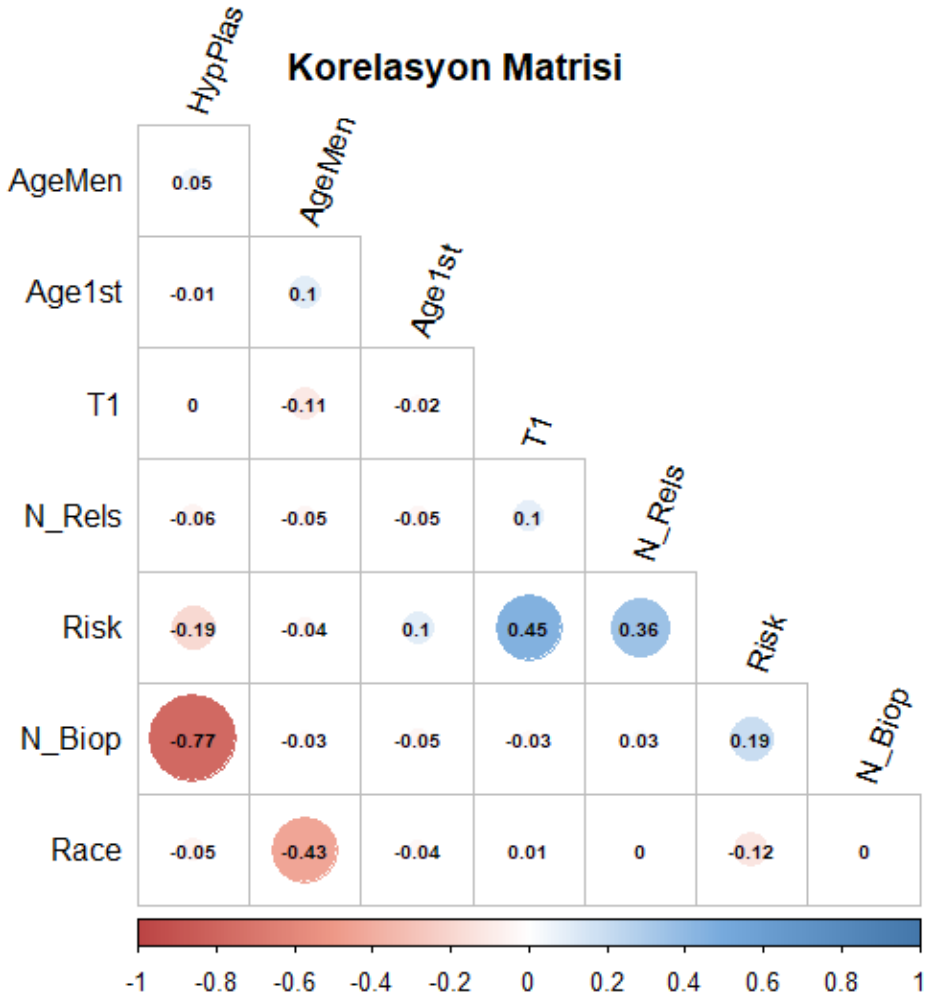
Tablo 4.2. %70 eğitim ve %30 test veri seti için k-NN algoritması sınıflandırma sonucu

k-NN	0	1
0	3	3
1	27	113

k-NN sınıflandırması istatistiksel olarak anlamsızdır (*p*= 0.5487). Sınıflandırma sonucuna göre toplamda 30 değer hatalı sınıflandırılmıştır (FN:3, FP:27).

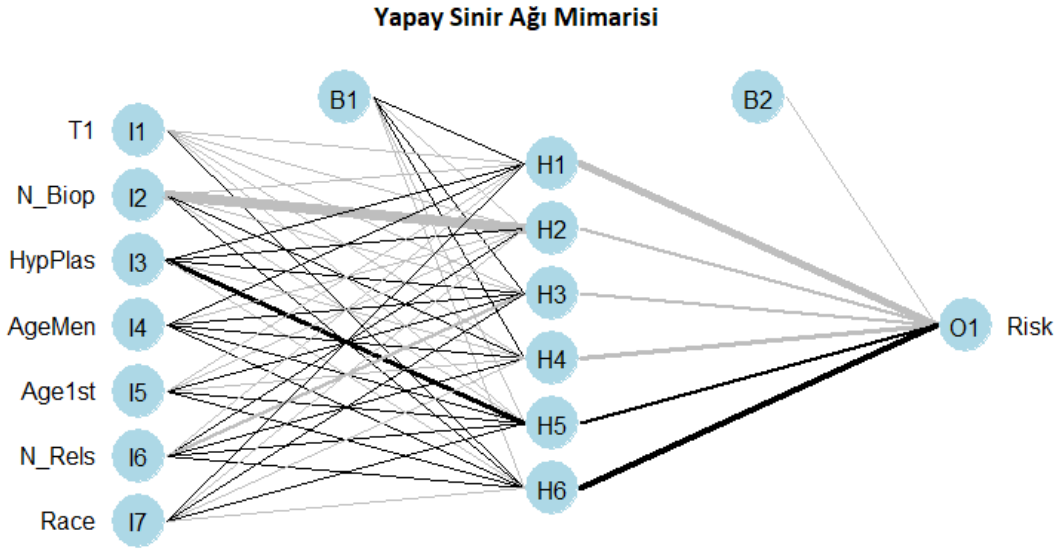
4.1.2. YSA sınıflandırma sonucuna ilişkin bulgular

Yapay Sinir Ağında optimal sonuç elde etmek ve ilişkili değişkenleri tespit etmek için Şekil 4.2' de yer alan korelasyon matrisi oluşturulmuştur.

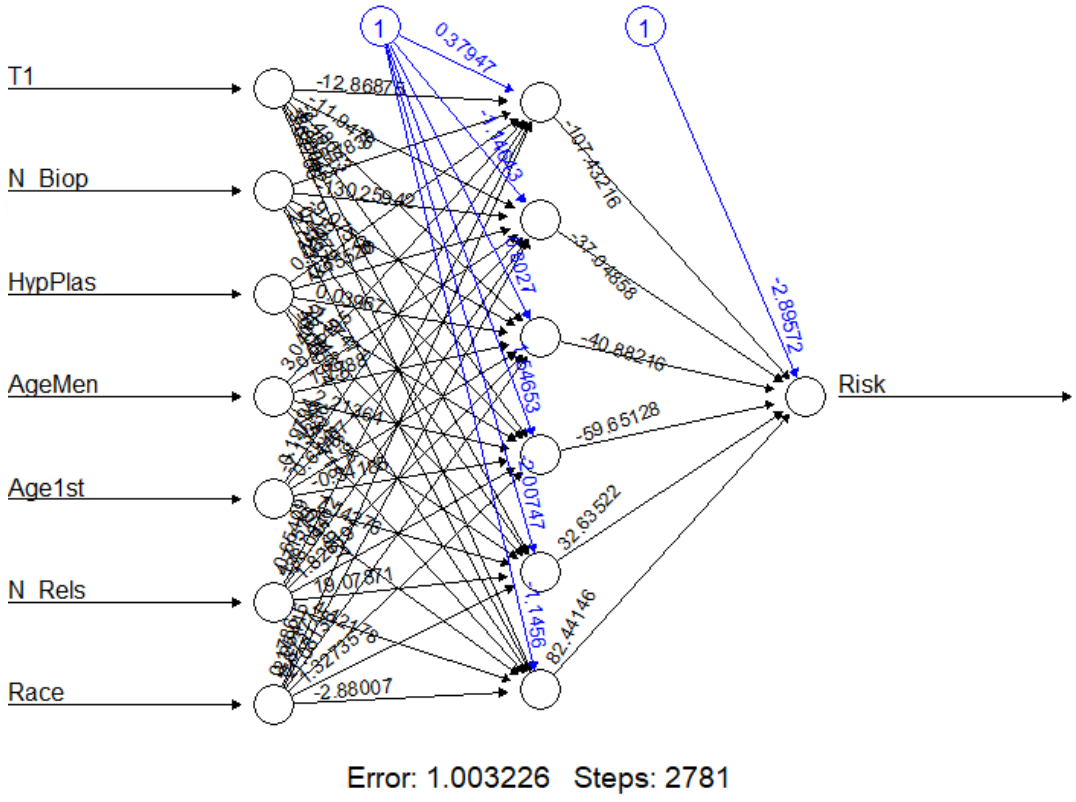


Şekil 4.2. %70 eğitim ve %30 test veri seti için korelasyon matrisi

Korelasyon matrisine göre; Risk ve Biyopsi Sayısı değişkenleri arasında pozitif yönde düşük düzeyde (r=0.19, p<0.001), Risk ve Hiperplazi değişkenleri arasında negatif yönde düşük düzeyde (r=-0.19, p<0.001) ilişki vardır. Risk ve Hasta Akraba Sayısı değişkenleri arasında pozitif yönde orta düzeyde (r=0.36, p<0.001), Adet Yaşı ve Irk değişkenleri arasında pozitif yönde orta düzeyde (r=-0.43, p<0.001), Risk ve Yaş değişkenleri arasında pozitif yönde orta düzeyde (r=0.45, p<0.001) ilişki vardır. Biyopsi sayısı ve Hiperplazi değişkenleri arasında negatif yönde yüksek düzeyde (r=-0.77, p<0.001) ilişki vardır.



Şekil 4.3. %70 eğitim ve %30 test veri seti için yapay sinir ağı mimarisi



Şekil 4.4. %70 eğitim ve %30 test veri seti için yapay sinir ağı matematiksel gösterimi

Tablo 4.3. YSA algoritma parametreleri

Giriş katmanındaki nöron sayısı	7
Gizli katmanların sayısı	1
Gizli katmandaki nöron sayısı	6
Çıktı katmanındaki nöron sayısı	1
Öğrenme Algoritması	RPROP+ (Geri beslemeli)
Aktivasyon Fonksiyonu	Logistic
Öğrenme Oranı	minus=0.5, plus=1.2
Ağ Çıkışı	0: Az Riski, 1: Riskli

Tablo 4.4. %70 eğitim ve %30 test veri seti için YSA algoritması sınıflandırma sonucu

YSA	0	1
0	30	3
1	0	117

Yapay Sinir Ağları sınıflandırması istatistiksel olarak anlamlıdır (p<0.001). Sınıflandırma sonucuna göre toplamda 3 değer hatalı sınıflandırılmıştır (FN:0, FP:3).

4.1.3. SVM sınıflandırma sonucuna ilişkin bulgular

Tablo 4.5'te SVM algoritma parametreleri belirtilmiştir.

Tablo 4.5. SVM algoritması parametreleri

Kernel	Radial
Maliyeti (cost)	100
Gamma	1/train data sayısı
Tolerans	0.001
Epsilon	0.1

Tablo 4.6. %70 eğitim ve %30 test veri seti için SVM algoritması sınıflandırma sonucu

SVM	0	1
0	29	11
1	1	109

Tablo 4.6'da SVM sınıflandırma sonuçları verilmiştir. Destek Vektör Makinesi sınıflandırması istatistiksel olarak anlamlıdır (p<0.001). Sınıflandırma sonucuna göre toplamda 12 değer hatalı sınıflandırılmıştır. (FN:1, FP:11).

4.1.4. NB sınıflandırma sonucuna ilişkin bulgular

Tablo 4.7'de NB algoritma parametreleri belirtilmiştir.

Tablo 4.7. %70 eğitim ve %30 test veri seti için NB algoritması parametreleri

Laplace	0
Eşik Değeri	0.001
Epsilon	0.1

Tablo 4.8. %70 eğitim ve %30 test veri seti için NB algoritması sınıflandırma sonucu

NB	0	1
0	23	7
1	7	113

Naive Bayes sınıflandırma istatistiksel olarak anlamlıdır (p<0.001). Sınıflandırma sonucuna göre toplamda 14 değer hatalı sınıflandırılmıştır (FN:7, FP:7).

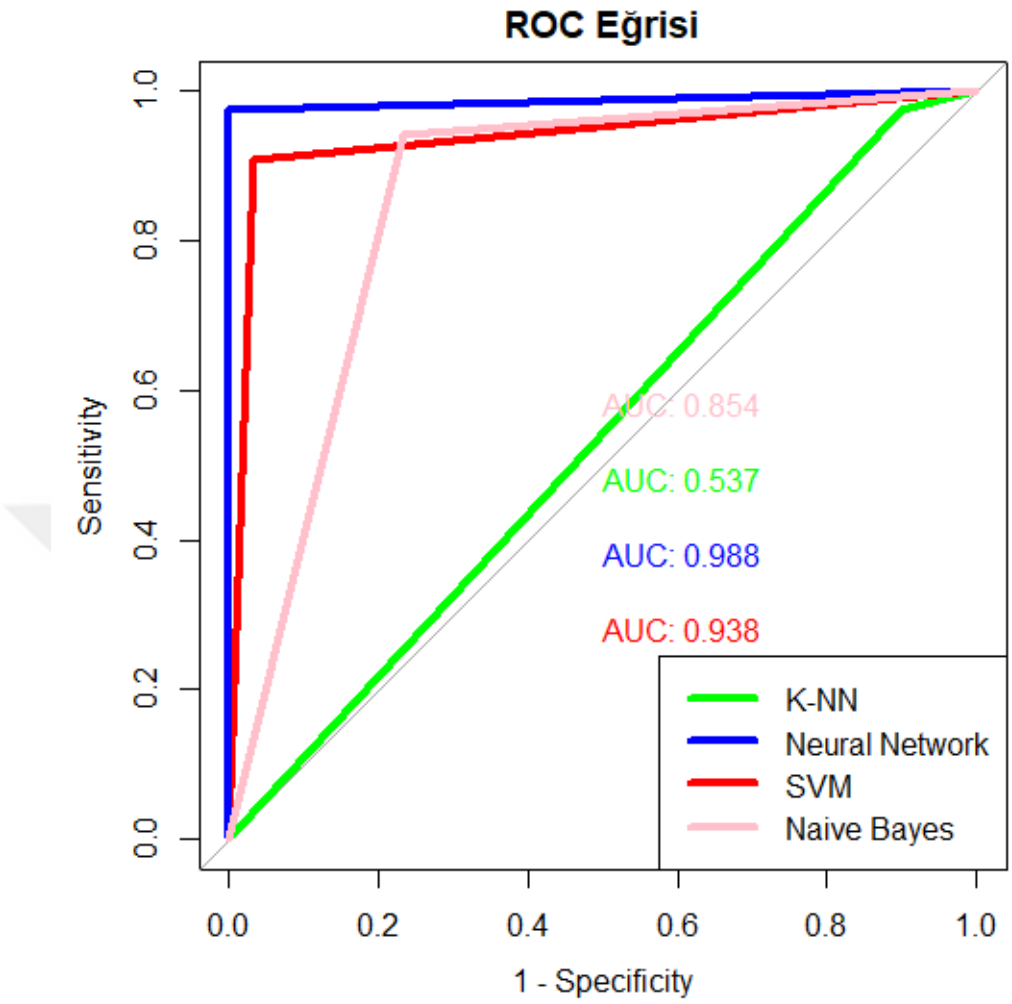
4.1.5. ROC eğrisi sonucuna ilişkin bulgular

Tablo 4.9'da performans değerlendirme sonuçlarına göre sınıflandırma karşılaştırmaları verilmiştir. Sınıflandırma karşılaştırma sonucuna göre en yüksek sınıflandırma sonucu ve Yapay Sinir Ağları algoritmaları ile elde edilmiştir. Doğruluk yüzdesi ve Kappa ölçütleri baz alındığında en yüksek performanstan en düşüğe doğru sırasıyla sınıflandırma sonuçları; Yapay Sinir Ağları, Destek Vektör Makinesi, Naive Bayes ve k-En Yakın Komşu şeklindedir.

Tablo 4.9. %70 eğitim ve %30 test veri seti için sınıflandırma karşılaştırmaları

Model	Doğruluk Yüzdesi (Accuracy)	*p*-değeri	Kappa	Duyarlılık (Sensitivity)	Özgüllük (Specificity)	ROC Eğrisi (AUC)
k-NN	0.8000	0.5487	0.1071	0.1000	0.9750	0.5375
YSA	**0.9800**	***p*<0.001**	**0.9398**	**1.0000**	**0.9750**	**0.9875**
SVM	0.9200	*p*<0.001	0.7778	0.9667	0.9083	0.9375
NB	0.9067	*p*<0.001	0.7083	0.7667	0.9417	0.8542

Şekil 4.5. %70 eğitim ve %30 test veri seti için ROC eğrisi



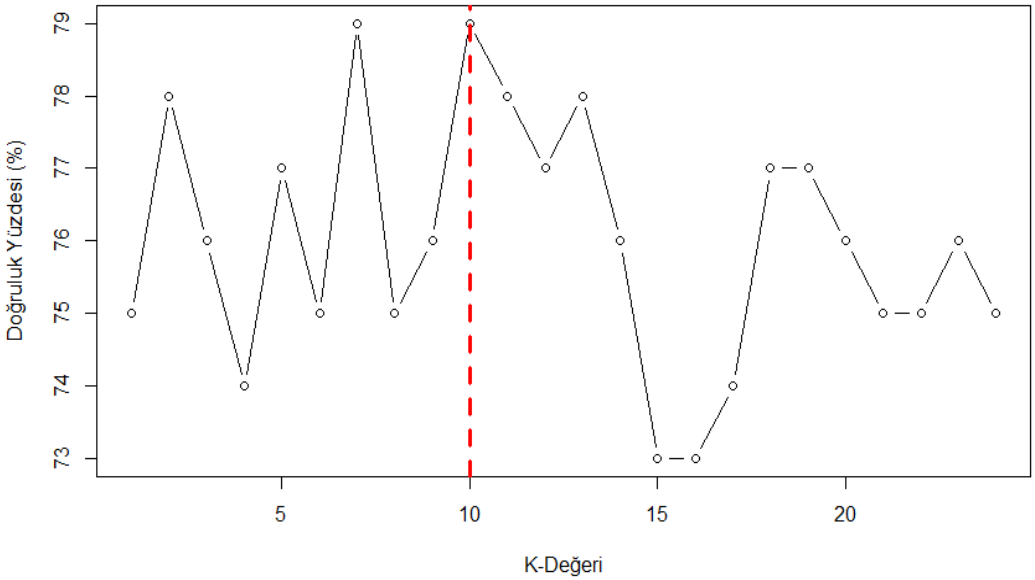
Şekil 4.5'te ROC eğrisi gösterilmiştir. Şekle göre YSA yöntemine ait ROC eğrisi altında kalan alan 0.988'dir ve bu alan $0.90<AUC_{YSA}<1.0$ olduğundan YSA yönteminin sınıflandırması mükemmeldir. Aynı şekilde SVM ait ROC eğrisi altında kalan 0.938'dir ve bu alan $0.90<AUC_{SVM}<1.0$ olduğundan SVM yönteminin sınıflandırması mükemmeldir. NB yöntemine ait ROC eğrisi altında kalan alan 0.854'tür ve $0.80<AUC_{NB}<0.90$ olduğundan NB yönteminin sınıflandırması iyidir. Kullanılan yöntemler arasında en düşük performanslı sınıflandırma yönteminin k-NN (AUC=0.537), en yüksek performanslı sınıflandırma yönteminin ise YSA (AUC=0.988) olduğu görülmüştür.

4.2. %80 Eğitim ve %20 Test Veri Seti İçin Bulgular

4.2.1. k-NN sınıflandırma sonucuna ilişkin bulgular

Veri setine ilk önce k-NN algoritması uygulanmıştır. k değerinin optimal sonucu için doğruluk yüzdesi elde edilen Şekil 4.6'da belirtilmiştir.

Şekil 4.6. %80 eğitim ve %20 test veri seti için k-değeri optimal doğruluk yüzdesi sonucu



Şekil 4.6'ya göre doğruluk yüzdesi en yüksek olan k değeri 10'dur. Algoritmaya k değeri 10 seçilerek devam edilmiştir. Algoritma parametreleri %70 eğitim ve %30 test veri seti ile aynı olup, k değeri 10 olarak değiştirilmiştir ve sınıflandırma sonuçları Tablo 4.10'da belirtilmiştir.

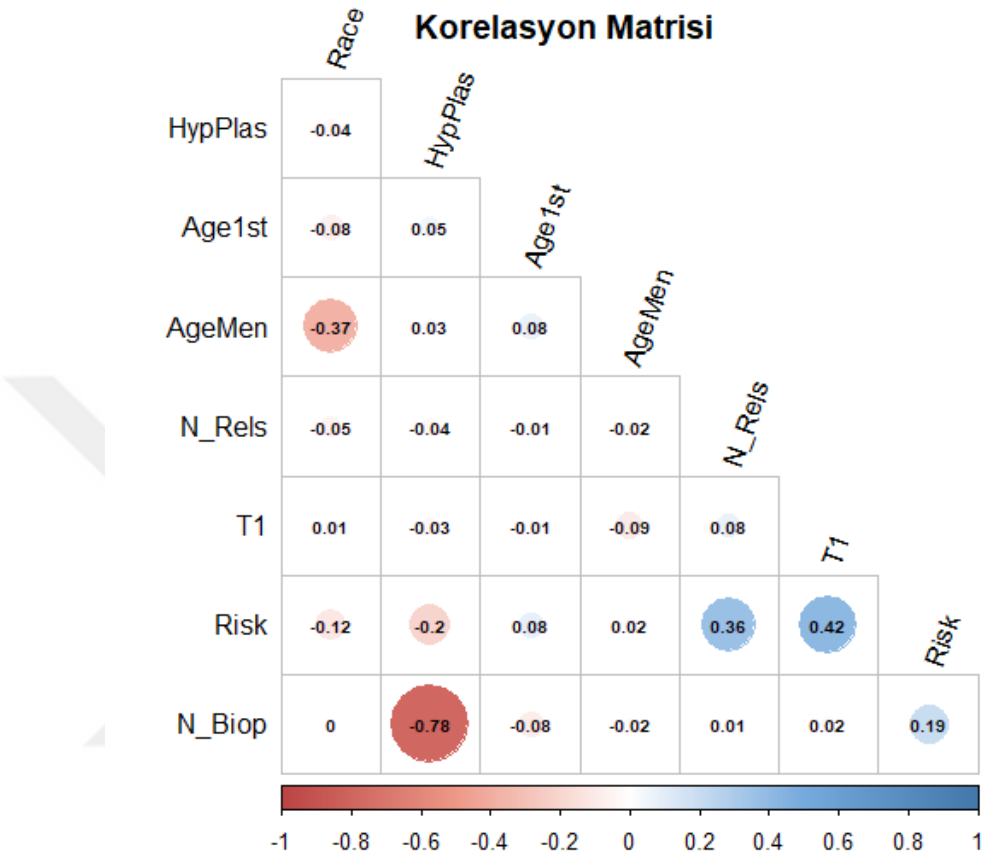
Tablo 4.10. %80 eğitim ve %20 test veri seti için k-NN algoritması sınıflandırma sonucu

k-NN	0	1
0	5	3
1	18	74

k-En Yakın Komşu sınıflandırma istatistiksel olarak anlamsızdır (p=0.3679). Sınıflandırma sonucuna göre sonucuna göre toplamda 21 değer hatalı sınıflandırılmıştır (FN:3, FP:18)

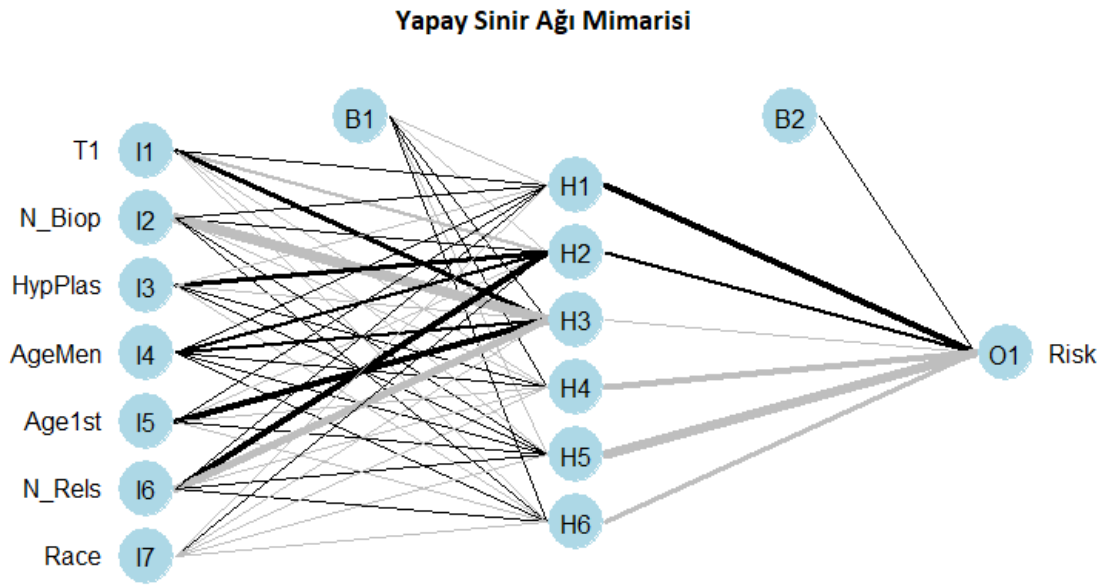
4.2.2. YSA sınıflandırma sonucuna ilişkin bulgular

Yapay Sinir Ağında optimal sonuç elde etmek ve ilişkili değişkenleri tespit etmek için Şekil 4.7' de yer alan korelasyon matrisi oluşturulmuştur.

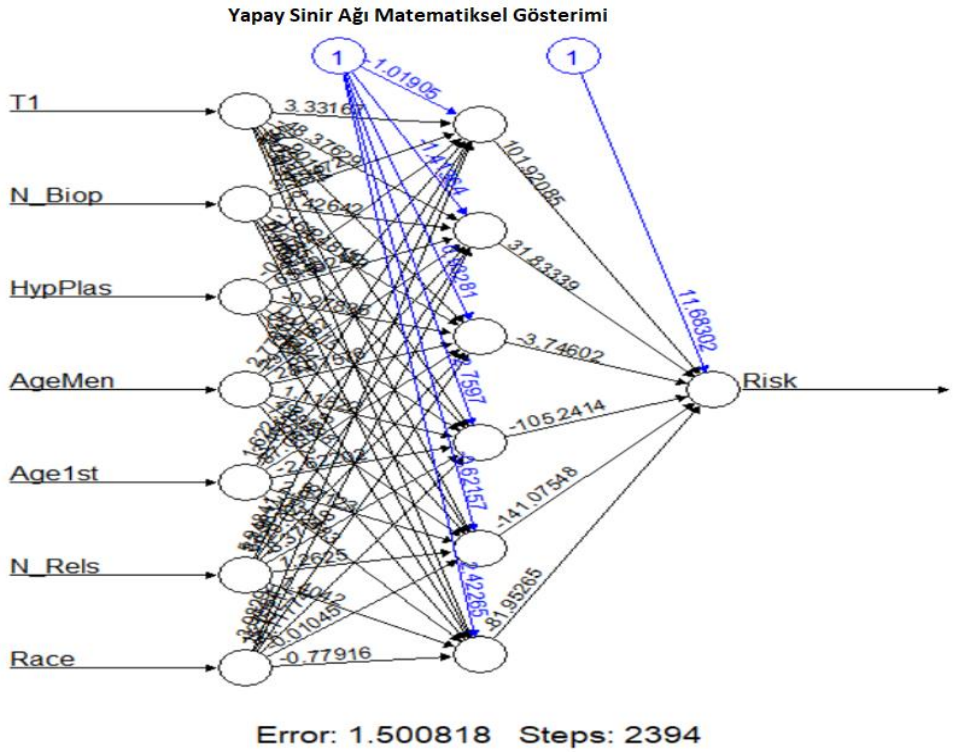


Şekil 4.7. %80 eğitim ve %20 test veri seti için korelasyon matrisi

Korelasyon matrisine göre, Risk ve Biyopsi Sayısı değişkenleri arasında pozitif yönde düşük düzeyde (r=0.19, p<0.001), Risk ve Hiperplazi değişkenleri arasında negatif yönde düşük düzeyde (r=-0.20, p<0.001) ilişki vardır. Risk ve Hasta Akraba Sayısı değişkenleri arasında pozitif yönde orta düzeyde ilişki (r=0.36, p<0.001), Adet Yaşı ve Irk değişkenleri arasında negatif yönde orta düzeyde ilişki (r=-0.37, p<0.001), Risk ve Yaş değişkenleri arasında pozitif yönde orta düzeyde ilişki vardır (r=0.42, p<0.001). Biyopsi sayısı ve Hiperplazi (r=-0.79, p<0.001) değişkenleri arasında negatif yönde yüksek düzeyde ilişki vardır.



Şekil 4.8. %80 eğitim ve %20 test veri seti için yapay sinir ağı mimarisi



Şekil 4.9. %80 eğitim ve %20 test veri seti için yapay sinir ağı mimarisi matematiksel gösterimi

Şekil 4.8 ve Şekil 4.9' da Yapay Sinir Ağı mimarisi elde edilmiştir. Sinir ağı 6 nöronlu tek katmandan oluşmaktadır. Algoritma parametreleri %70 eğitim ve %30 test veri seti ile aynı olup sınıflandırma sonucu Tablo 4.11'de belirtilmiştir.

Tablo 4.11. %80 eğitim ve %20 test veri seti için YSA algoritması sınıflandırma sonucu

YSA	0	1
0	22	1
1	1	76

Yapay Sinir Ağları sınıflandırması istatistiksel olarak anlamlıdır (p<0.001). Sınıflandırma sonucuna göre toplamda 2 değer hatalı sınıflandırılmıştır (FN:1, FP:1).

4.2.3. SVM sınıflandırma sonucuna ilişkin bulgular

Algoritma parametreleri %70 eğitim ve %30 test veri seti ile aynı olup sınıflandırma sonucu Tablo 4.12'de belirtilmiştir.

Tablo 4.12. %80 eğitim ve %20 test veri seti için SVM algoritması sınıflandırma sonucu

SVM	0	1
0	20	4
1	3	73

Destek Vektör Makinesi sınıflandırması istatistiksel olarak anlamlıdır (p<0.001). Sınıflandırma sonucuna göre toplamda 7 değer hatalı sınıflandırılmıştır (FN:4, FP:3).

4.2.4. NB sınıflandırma sonucuna ilişkin bulgular

Algoritma parametreleri %70 eğitim ve %30 test veri seti ile aynı olup sınıflandırma sonucu Tablo 4.13'te belirtilmiştir.

Tablo 4.13. %80 eğitim ve %20 test veri seti için NB algoritması sınıflandırma sonucu

NB	0	1
0	21	4
1	2	73

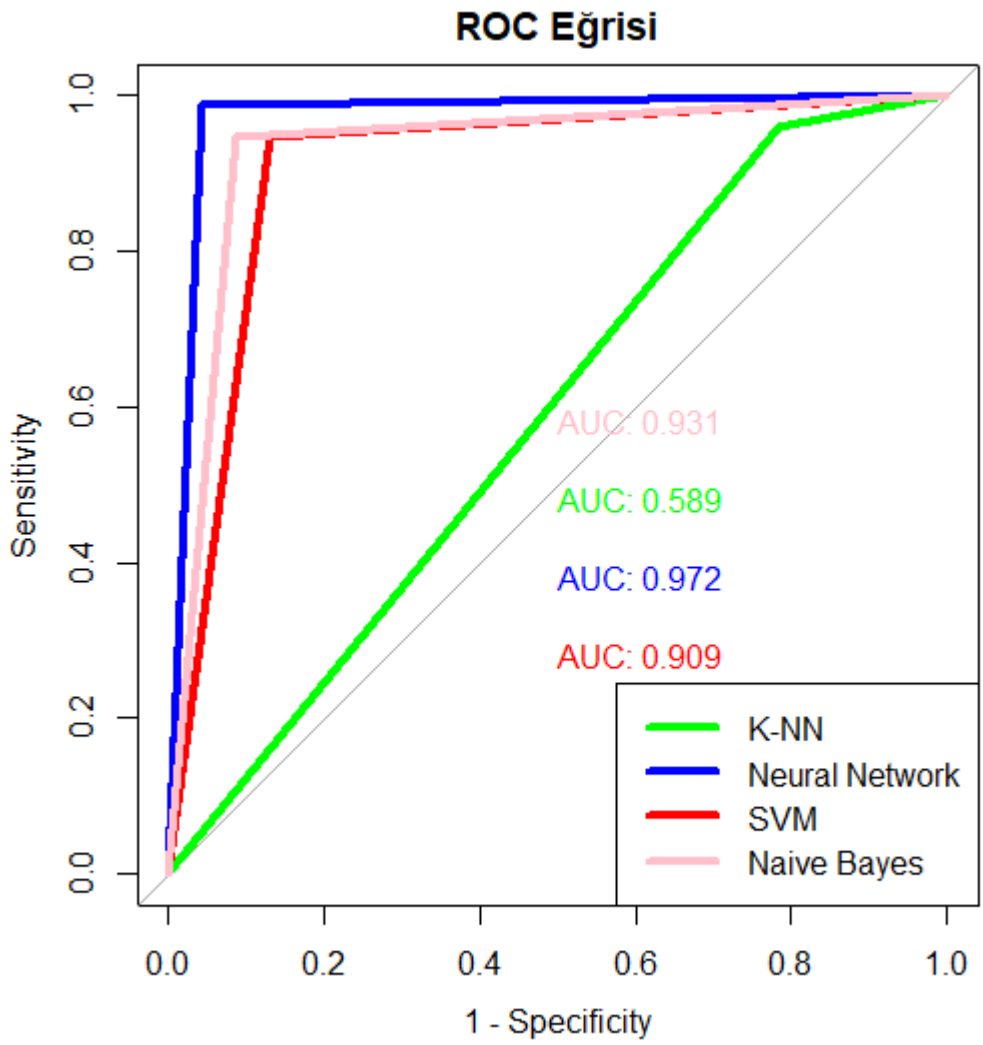
Naive Bayes sınıflandırması istatistiksel olarak anlamlıdır (p<0.001). Sınıflandırma sonucuna göre toplamda 6 değer hatalı sınıflandırılmıştır (FN:4, FP:2)

4.2.5. ROC eğrisi sonuçlarına ilişkin bulgular

Tablo 4.14'de performans değerlendirme sonuçlarına göre sınıflandırma karşılaştırmaları verilmiştir. Sınıflandırma karşılaştırma sonucuna göre en yüksek sınıflandırma sonucu ve Yapay Sinir Ağları algoritmaları ile elde edilmiştir. Doğruluk yüzdesi ve Kappa ölçütleri baz alındığında en yüksek performanstan en düşüğe doğru sırasıyla sınıflandırma sonuçları; Yapay Sinir Ağları, Naive Bayes, Destek Vektör Makinesi ve k-En Yakın Komşu şeklindedir.

Tablo 4.14. %80 eğitim ve %20 test veri seti için sınıflandırma karşılaştırmaları

Model	Doğruluk Yüzdesi (Accuracy)	*p*-değeri	Kappa	Duyarlılık (Sensitivity)	Özgüllük (Specificity)	ROC Eğrisi (AUC)
k-NN	0.7900	0.3679	0.2313	0.2174	0.9610	0.5892
YSA	**0.9800**	***p*<0.001**	**0.9435**	**0.9565**	**0.9870**	**0.9718**
SVM	0.9300	*p*<0.001	0.8053	0.8696	0.9481	0.9088
NB	0.9400	*p*<0.001	0.8356	0.9130	0.9481	0.9305



Şekil 4.10. %80 eğitim ve %20 test veri seti için ROC eğrisi

Şekil 4.10'da ROC eğrisi gösterilmiştir. Şekle göre YSA yöntemine ait ROC eğrisi altında kalan alan 0.972, NB yöntemine ait ROC eğrisi altında kalan alan 0.931, SVM yöntemine ait ROC eğrisi altında kalan alan 0.909 ve k-NN yöntemine ait ROC eğrisi altında kalan alan 0.589'dir. Sınıflandırma sonuçlarına göre $0.90<AUC_{YSA}<1.0$, $0.90<AUC_{NB}<1.0$, $0.90<AUC_{SVM}<1.0$ olduğundan YSA, NB ve SVM yöntemlerinin sınıflandırmaları mükemmeldir. Kullanılan yöntemler arasında en düşük performanslı sınıflandırma yönteminin k-NN (AUC=0.67) olduğu görülmüştür.

5. TARTIŞMA

Bazazeh ve Shubair (2016) çalışmalarında, meme kanseri tespiti ve teşhisi için Wisconsin meme kanseri veri setini, yaygın olarak kullanılan makine öğrenmesi yöntemlerinden üçünü kullanarak (Destek Vektör Makinesi, Rastgele Orman ve Bayesian Ağı) karşılaştırmışlardır. Karşılaştırma sonucunda Destek Vektör Makinesi %96.6, Rastgele Orman %99.9, Bayesian Ağı %99.1 doğrulukla sınıflandırmıştır (Bazazeh & Shubair, 2016).

Amrane ve arkadaşları (2018) çalışmalarında, Wisconsin meme kanseri veri seti ile meme kanseri sınıflandırmasını k-En Yakın Komşu ve Naive Bayes algoritmalarıyla karşılaştırmışlardır. Karşılaştırma sonucunda k-NN %97.51 ve NB %96.19 doğrulukla sınıflandırmıştır (Amrane, Oukid, Gagaoua, & Ensarİ, 2018).

Saritas ve Yasar (2019) çalışmalarında, meme kanseri şüphesiyle kliniğe başvuran hastaların verilerine Yapay Sinir Ağları ve Naive Bayes sınıflandırma algoritmaları uygulamış ve hastalık tanısını tahmin etmeleri istenmiştir. YSA algoritması % 86.95 ve NB algoritması ise % 83.54 doğrulukla sınıflandırmıştır.

Stark ve arkadaşları (2019) çalışmalarında 5 yıllık meme kanseri riskini BCRAT, Yapay Sinir Ağları, Lojistik Regresyon ve Doğrusal Diskriminant Analizi yöntemleriyle karşılaştırmışlardır. Karşılaştırma sonucunda BCRAT doğruluk yüzdesinin % 56.3, Lojistik Regresyon ve Doğrusal Diskriminant Analizi doğruluk yüzdelerinin % 61.3 ve Yapay Sinir Ağları doğruluk yüzdesinin % 60.8 olduğu gözlenmiştir.

Tseng ve arkadaşları (2019) meme kanseri metastazını tahmin etmek için makine öğrenmesi algoritmalarından Rasgele Orman, Naive Bayes, Destek Vektör Makinesi ve Lojistik Regresyon yöntemlerini karşılaştırmışlardır. Karşılaştırma sonucuna göre Rasgele Orman AUC=0.746, Naive Bayes AUC=0.648, Destek Vektör Makinesi AUC=0.645 ve Lojistik Regresyon AUC=0.581 olduğu gözlenmiştir (Tseng vd., 2019).

Ganggayah ve arkadaşları (2019), çalışmalarında 1993 ile 2016 yılları arasında Malezya'daki Malaya Üniversitesi Tıp Merkezi'nden elde edilen meme kanseri veri seti ile meme kanseri hayatta kalma oranını belirlemede; Karar Ağacı, Rastgele Orman, Yapay Sinir Ağları, Extreme Boost, Lojistik Regresyon ve Destek Vektör Makinesi yöntemlerini karşılaştırmışlardır. Karşılaştırma sonucunda Karar

ağacı doğruluk yüzdesinin %72, Yapay Sinir Ağları doğruluk yüzdesinin % 84, Lojistik Regresyon ve Destek Vektör Makinesinin doğruluk yüzdelerinin % 85, Rastgele Orman doğruluk yüzdesinin %86, Extreme Boost doğruluk yüzdesinin %87 olduğu gözlenmiştir (Ganggayah, Taib, Har, Lio, & Dhillon, 2019) .

Ming ve arkadaşları (2020) çalışmalarında Markov Zinciri Monte Carlo Genelleştirilmiş Doğrusal Karma Model (MCMCGLMM), AdaBoost ve Rastgele Orman (RF) ve BOADICEA modeli kullanmışlardır. BOADICEA % 63.9, AdaBoost %88,9, MCMCGLMM % 85.1, RF % 84.3 tahmin doğrulu ile sınıflandırmıştır.

Literatürde meme kanseri tanısının ya da kanser riskinin makine öğrenmesi yöntemleriyle sınıflandırıldığı çalışmalar olup, bu çalışma da benzer niteliktedir. Bu çalışmada meme kanseri risk tahmini için k-En Yakın Komşular, Yapay Sinir Ağları, Destek Vektör Makinesi ve Naive Bayes makine öğrenme algoritmalarının performanslarını karşılaştırmışlardır. Çalışmanın amacı, verileri her algoritmanın verimlilik ve etkinlik açısından sınıflandırmadaki doğruluğunu; doğruluk, Kappa, duyarlılık, özgüllük ve ROC eğrisi açısından değerlendirmektir.

6. SONUÇ VE ÖNERİLER

Bu tez çalışmasında R Studio Version 3.6.3 programı kullanılarak 500 adet veri türetilmiştir. Türetilen veri senaryosunda Gail Modeli ile meme kanseri riski hesaplanmış, daha sonra riski hesaplanan verilerin makine öğrenmesi yöntemlerine göre sınıflandırılması amaçlanmıştır. Sınıflandırma sonuçları %70 eğitim %30 test ve %80 eğitim %20 test olmak üzere 2 ayrı eğitim ve test veri setinde karşılaştırılmıştır.

Sınıflandırma performansını ölçmek için; doğruluk yüzdesi, Kappa, duyarlılık, özgüllük ve ROC eğrisi ölçütleri kullanılmıştır. Bu ölçütler doğrultusunda, %70 eğitim %30 test veri seti için karşılaştırma sonuçları Tablo 4.9'da gösterilmiştir. Bu karşılaştırmaya göre en yüksek sınıflandırma sonucu Yapay Sinir Ağları algoritması ile elde edilmiştir. Sırasıyla sınıflandırma sonuçları en yüksek performanstan en düşüğe doğru Yapay Sinir Ağları, Destek Vektör Makinesi, Naive Bayes ve k-En Yakın Komşu şeklindedir.

%80 eğitim %20 test veri seti için karşılaştırma sonuçları Tablo 4.14'te gösterilmiştir. Bu karşılaştırmaya göre en yüksek sınıflandırma sonucu ve Yapay Sinir Ağları algoritmaları ile elde edilmiştir. Doğruluk yüzdesi ve Kappa ölçütleri baz alındığında en yüksek performanstan en düşüğe doğru sırasıyla sınıflandırma sonuçları; Yapay Sinir Ağları, Naive Bayes, Destek Vektör Makinesi ve k-En Yakın Komşu şeklindedir.

Değerlendirme sonuçlarında; %70 eğitim %30 test veri seti için sınıflandırma sonuçları sırasıyla k-En Yakın Komşular (AUC=0.5375), Naive Bayes (AUC=0.8542), Destek Vektör Makinesi (AUC=0.9375) ve Yapay Sinir Ağları'dır (AUC=0.9875). Yapay Sinir Ağları'nın en düşük hata oranı ile en yüksek doğruluğu verdiği görülmüştür.

%80 eğitim %20 test veri seti için sınıflandırma sonuçları sırasıyla k-En Yakın Komşular (AUC=0.5892), Destek Vektör Makinesi (AUC=0.9088) Naive

Bayes (AUC=0.9305), ve Yapay Sinir Ağları'dır (AUC=0.9718). Yapay Sinir Ağları'nın en düşük hata oranı ile en yüksek doğruluğu verdiği görülmüştür.

Makine öğrenmesi ile istatistiksel modeller arasındaki en büyük fark, amaçlarıdır. Makine öğrenmesi modelleri, mümkün olan en doğru tahminleri yapmak

için tasarlanmıştır. İstatistiksel modeller ise, değişkenler arasındaki ilişkiler hakkında çıkarım yapmak için tasarlanmıştır. İstatistiksel modeller, tahmin, sınıflandıırıcı, sınıflandırma ve regresyon, parametre, model ve ortak değişkenler kullanırken makine öğrenmesi, öğrenme, varsayım, örnekler, denetimli ve denetimsiz öğrenme, ağırlık, verinin öne çıkan özelliklerini kullanır.

Veri seti yeni makine öğrenmesi algoritmaları ile denenebilir, meme kanseri riski diğer risk tahmin edici modellerle hesaplanabilir, hastane verileri kullanılabilir, meme kanseri üzerinde yapılacak yeni çalışmalar ışığında, meme kanseri riskinde önem teşkil eden diğer faktörler de belirlenip makine öğrenmesi algoritmalarına aktarılarak sınıflandırma başarıları artırılabilir.

KAYNAKLAR DİZİNİ

Ahmad, L. G., Eshlaghy, A., Poorebrahimi, A., Ebrahimi, M., & Razavi, A. (2013). Using three machine learning techniques for predicting breast cancer recurrence. *J Health Med Inform, 4*(124), 3.

Ahuja, M. (2016). Age of menopause and determinants of menopause age: A PAN India survey by IMS. *Journal of mid-life health, 7*(3), 126.

Akay, E. Ç. (2018). Ekonometride Yeni Bir Ufuk: Büyük Veri ve Makine Öğrenmesi. *Sosyal Bilimler Araştırma Dergisi, 7*(2), 41-53.

Alpar, R. (2017). *Uygulamalı çok değişkenli istatistiksel yöntemler*: Detay yayıncılık.

Alpaydin, E. (2014). *Introduction to machine learning*: MIT press.

Amir, E., Freedman, O. C., Seruga, B., & Evans, D. G. (2010). Assessing women at high risk of breast cancer: a review of risk assessment models. *JNCI: Journal of the National Cancer Institute, 102*(10), 680-691.

Amit, Y., & Geman, D. (1997). Shape quantization and recognition with randomized trees. *Neural computation, 9*(7), 1545-1588.

Amrane, M., Oukid, S., Gagaoua, I., & Ensarİ, T. (2018). *Breast cancer classification using machine learning.* Paper presented at the 2018 Electric Electronics, Computer Science, Biomedical Engineerings' Meeting (EBBT).

Anothaisintawee, T., Teerawattananon, Y., Wiratkapun, C., Kasamesup, V., & Thakkinstian, A. (2012). Risk prediction models of breast cancer: a systematic review of model performances. *Breast cancer research and treatment, 133*(1), 1-10.

Antoniou, A. C., Cunningham, A., Peto, J., Evans, D., Lalloo, F., Narod, S., . . . Southey, M. (2008). The BOADICEA model of genetic susceptibility to breast and ovarian cancers: updates and extensions. *British journal of cancer, 98*(8), 1457-1466.

Asri, H., Mousannif, H., Al Moatassime, H., & Noel, T. (2016). Using machine learning algorithms for breast cancer risk prediction and diagnosis. *Procedia Computer Science, 83*, 1064-1069.

Barton, M. B. (2005). Breast cancer screening: benefits, risks, and current controversies. *Postgraduate medicine, 118*(2), 27-46.

Bazazeh, D., & Shubair, R. (2016). *Comparative study of machine learning algorithms for breast cancer detection and diagnosis.* Paper presented at the 2016 5th International Conference on Electronic Devices, Systems and Applications (ICEDSA).

Bose, N. K., & Garga, A. K. (1993). Neural network design using Voronoi diagrams. *IEEE Transactions on Neural Networks, 4*(5), 778-787.

Botev, Z. I., Grotowski, J. F., & Kroese, D. P. (2010). Kernel density estimation via diffusion. *The annals of Statistics, 38*(5), 2916-2957.

KAYNAKLAR DİZİNİ (Devam Ediyor)

Boyle, P., Mezzetti, M., La Vecchia, C., Franceschi, S., Decarli, A., & Robertson, C. (2004). Contribution of three components to individual cancer risk predicting breast cancer risk in Italy. *European journal of cancer prevention, 13*(3), 183-191.

Brentnall, A. R., Harkness, E. F., Astley, S. M., Donnelly, L. S., Stavrinos, P., Sampson, S., . . . Wilson, M. (2015). Mammographic density adds accuracy to both the Tyrer-Cuzick and Gail breast cancer risk models in a prospective UK screening cohort. *Breast cancer research, 17*(1), 147.

Casanova, R., Hsu, F.-C., Sink, K. M., Rapp, S. R., Williamson, J. D., Resnick, S. M., . . . Initiative, A. s. D. N. (2013). Alzheimer's disease risk assessment using large-scale machine learning methods. *PloS one, 8*(11).

Chang, Y.-W., Hsieh, C.-J., Chang, K.-W., Ringgaard, M., & Lin, C.-J. (2010). Training and testing low-degree polynomial data mappings via linear SVM. *Journal of Machine Learning Research, 11*(4).

Chapman, C., Murray, A., Chakrabarti, J., Thorpe, A., Woolston, C., Sahin, U., . . . Robertson, J. (2007). Autoantibodies in breast cancer: their use as an aid to early diagnosis. *Annals of oncology, 18*(5), 868-873.

Choi, H., & Choi, S. (2007). Robust kernel isomap. *Pattern recognition, 40*(3), 853-862.

Churpek, M. M., Yuen, T. C., Winslow, C., Meltzer, D. O., Kattan, M. W., & Edelson, D. P. (2016). Multicenter comparison of machine learning methods and conventional regression for predicting clinical deterioration on the wards. *Critical care medicine, 44*(2), 368.

Clark, I. A., Niehaus, K. E., Duff, E. P., Di Simplicio, M. C., Clifford, G. D., Smith, S. M., . . . Holmes, E. A. (2014). First steps in using machine learning on fMRI data to predict intrusive memories of traumatic film footage. *Behaviour research and therapy, 62*, 37-46.

Claus, E. B., Risch, N., & Thompson, W. D. (1991). Genetic analysis of breast cancer in the cancer and steroid hormone study. *American journal of human genetics, 48*(2), 232.

Costantino, J. P., Gail, M. H., Pee, D., Anderson, S., Redmond, C. K., Benichou, J., & Wieand, H. S. (1999). Validation studies for models projecting the risk of invasive and total breast cancer incidence. *Journal of the National Cancer Institute, 91*(18), 1541-1548.

Coşkun, S., Kartal, M., Coşkun, A., & Bircan, H. (2004). Lojistik regresyon analizinin incelenmesi ve diş hekimliğinde bir uygulaması. *Cumhuriyet Üniversitesi Diş Hekimliği Fakültesi Dergisi, 7*(1), 42-50.

Couch, F. J., DeShano, M. L., Blackwood, M. A., Calzone, K., Stopfer, J., Campeau, L., . . . Jablon, L. (1997). BRCA1 mutations in women attending clinics that evaluate the risk of breast cancer. *New England Journal of Medicine, 336*(20), 1409-1415.

KAYNAKLAR DİZİNİ (Devam Ediyor)

Criminisi, A., Shotton, J., & Konukoglu, E. (2012). Decision forests: A unified framework for classification, regression, density estimation, manifold learning and semi-supervised learning. *Foundations and Trends® in Computer Graphics and Vision, 7*(2–3), 81-227.

Dumitrescu, R., & Cotarla, I. (2005). Understanding breast cancer risk-where do we stand in 2005? *Journal of cellular and molecular medicine, 9*(1), 208-221.

Engel, C., & Fischer, C. (2015). Breast cancer risks and risk prediction models. *Breast care, 10*(1), 7-12.

Eroglu, C., Eryilmaz, M. A., Civcik, S., & Gurbuz, Z. (2010). Meme Kanseri Risk Değerlendirmesi: 5000 Olgu. *International Journal of Hematology & Oncology/UHOD: Uluslararasi Hematoloji Onkoloji Dergisi, 20*(2).

Evans, D. G. R., & Howell, A. (2007). Breast cancer risk-assessment models. *Breast cancer research, 9*(5), 213.

Faustino, P. F. P. (2011). *Dynamic equilibrium through reinforcement learning.*

Fiore, U., Palmieri, F., Castiglione, A., & De Santis, A. (2013). Network anomaly detection with the restricted Boltzmann machine. *Neurocomputing, 122*, 13-23.

Fitzmaurice, C., Akinyemiju, T. F., Al Lami, F. H., Alam, T., Alizadeh-Navaei, R., Allen, C., . . . Anderson, B. O. (2018). Global, regional, and national cancer incidence, mortality, years of life lost, years lived with disability, and disability-adjusted life-years for 29 cancer groups, 1990 to 2016: a systematic analysis for the global burden of disease study. *JAMA oncology, 4*(11), 1553-1568.

Gail, M. H., Brinton, L. A., Byar, D. P., Corle, D. K., Green, S. B., Schairer, C., & Mulvihill, J. J. (1989). Projecting individualized probabilities of developing breast cancer for white females who are being examined annually. *JNCI: Journal of the National Cancer Institute, 81*(24), 1879-1886.

Gail, M. H., & Mai, P. L. (2010). Comparing breast cancer risk assessment models. In: Oxford University Press.

Ganggayah, M. D., Taib, N. A., Har, Y. C., Lio, P., & Dhillon, S. K. (2019). Predicting factors for survival of breast cancer patients using machine learning techniques. *BMC medical informatics and decision making, 19*(1), 48.

Guo, X., Chen, L., & Shen, C. (2016). Hierarchical adaptive deep convolution neural network and its application to bearing fault diagnosis. *Measurement, 93*, 490-502.

Hamzaçebi, C., & Kutay, F. (2004). Yapay sinir ağlari ile türkiye elektrik enerjisi tüketiminin 2010 yilina kadar Tahmini. *Gazi Üniversitesi Mühendislik-Mimarlık Fakültesi Dergisi, 19*(3).

Han, S., Liu, X., Mao, H., Pu, J., Pedram, A., Horowitz, M. A., & Dally, W. J. (2016). EIE: efficient inference engine on compressed deep neural network. *ACM SIGARCH Computer Architecture News, 44*(3), 243-254.

Hanley, J. A., & McNeil, B. J. (1982). The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology, 143*(1), 29-36.

KAYNAKLAR DİZİNİ (Devam Ediyor)

Harrington, P. (2012). *Machine learning in action*: Manning Publications Co.

Hasselt, H. V. (2010). *Double Q-learning.* Paper presented at the Advances in neural information processing systems.

Hastie, T., Tibshirani, R., & Wainwright, M. (2015). *Statistical learning with sparsity: the lasso and generalizations*: Chapman and Hall/CRC.

Hinton, G. E. (2009). Deep belief networks. *Scholarpedia, 4*(5), 5947.

Hu, L.-Y., Huang, M.-W., Ke, S.-W., & Tsai, C.-F. (2016). The distance function effect on k-nearest neighbor classification for medical datasets. *SpringerPlus, 5*(1), 1-9.

Huang, H., Lin, M., & Zhang, Q. (2017). *Double-Q learning-based DVFS for multi-core real-time systems.* Paper presented at the 2017 IEEE International Conference on Internet of Things (iThings) and IEEE Green Computing and Communications (GreenCom) and IEEE Cyber, Physical and Social Computing (CPSCom) and IEEE Smart Data (SmartData).

Jagielski, M., Oprea, A., Biggio, B., Liu, C., Nita-Rotaru, C., & Li, B. (2018). *Manipulating machine learning: Poisoning attacks and countermeasures for regression learning.* Paper presented at the 2018 IEEE Symposium on Security and Privacy (SP).

Jain, A., Nandakumar, K., & Ross, A. (2005). Score normalization in multimodal biometric systems. *Pattern recognition, 38*(12), 2270-2285.

Jerry, M., Chen, P.-Y., Zhang, J., Sharma, P., Ni, K., Yu, S., & Datta, S. (2017). *Ferroelectric FET analog synapse for acceleration of deep neural network training.* Paper presented at the 2017 IEEE International Electron Devices Meeting (IEDM).

Jia, H., Ding, S., Xu, X., & Nie, R. (2014). The latest research progress on spectral clustering. *Neural Computing and Applications, 24*(7-8), 1477-1486.

Kalogirou, S. A. (1999). Applications of artificial neural networks in energy systems. *Energy Conversion and Management, 40*(10), 1073-1087.

Karakayali, F. Y., Ekici, Y., Sevmiş, Ş., Pehlivan, S., Arat, Z., & Moray, G. (2007). Meme kanseri için risk belirlenmesinde Gail modeli. *Turkish Journal of Surgery, 23*(4), 129-135.

Karhunen, J., & Joutsensalo, J. (1995). Generalizations of principal component analysis, optimization problems, and neural networks. *Neural Networks, 8*(4), 549-562.

Keller, J. M., Gray, M. R., & Givens, J. A. (1985). A fuzzy k-nearest neighbor algorithm. *IEEE transactions on systems, man, and cybernetics*(4), 580-585.

Kelsey, J. L., Gammon, M. D., & John, E. M. (1993). Reproductive factors and breast cancer. *Epidemiologic reviews, 15*(1), 36.

Koç, M. L., Balas, C. E., & Arslan, A. (2004). Taş dolgu dalgakıranların yapay sinir ağları ile ön tasarımı. *Teknik Dergi, 15*(74).

Kuhn, M., & Johnson, K. (2013). *Applied predictive modeling* (Vol. 26): Springer.

KAYNAKLAR DİZİNİ (Devam Ediyor)

Lalkhen, A. G., & McCluskey, A. (2008). Clinical tests: sensitivity and specificity. *Continuing Education in Anaesthesia Critical Care & Pain, 8*(6), 221-223.

Langley, P., Iba, W., & Thompson, K. (1992). *An analysis of Bayesian classifiers.* Paper presented at the Aaai.

Lewis, D. D. (1998). *Naive (Bayes) at forty: The independence assumption in information retrieval.* Paper presented at the European conference on machine learning.

Likas, A., Vlassis, N., & Verbeek, J. J. (2003). The global k-means clustering algorithm. *Pattern recognition, 36*(2), 451-461.

Lin, C.-F., & Wang, S.-D. (2002). Fuzzy support vector machines. *IEEE Transactions on Neural Networks, 13*(2), 464-471.

Lindor, N. M., Lindor, R. A., Apicella, C., Dowty, J. G., Ashley, A., Hunt, K., . . . Hopper, J. L. (2007). Predicting BRCA1 and BRCA2 gene mutation carriers: comparison of LAMBDA, BRCAPRO, Myriad II, and modified Couch models. *Familial cancer, 6*(4), 473-482.

Liu, W., Wang, Z., Liu, X., Zeng, N., Liu, Y., & Alsaadi, F. E. (2017). A survey of deep neural network architectures and their applications. *Neurocomputing, 234*, 11-26.

McHugh, M. L. (2012). Interrater reliability: the kappa statistic. *Biochemia medica: Biochemia medica, 22*(3), 276-282.

Meyn, S. P. (1997). The policy iteration algorithm for average reward Markov decision processes with general state space. *IEEE Transactions on Automatic Control, 42*(12), 1663-1680.

Ming, C., Viassolo, V., Probst-Hensch, N., Dinov, I. D., Chappuis, P. O., & Katapodi, M. C. (2020). Machine learning-based lifetime breast cancer risk reclassification compared with the BOADICEA model: impact on screening recommendations. *British journal of cancer*, 1-8.

Nielsen, M. A. (2015). *Neural networks and deep learning* (Vol. 2018): Determination press San Francisco, CA, USA:.

Özkan, İ., & Ülker, E. (2017). Derin Öğrenme ve Görüntü Analizinde Kullanılan Derin Öğrenme Modelleri. *Gaziosmanpaşa Bilimsel Araştırma Dergisi, 6*(3), 85-104.

Özmen, V. (2013). Türkiye'de Meme Kanseri. *Turkiye Klinikleri General Surgery-Special Topics, 6*(2), 1-6.

Palmer, J. R., Rosenberg, L., Wise, L. A., Horton, N. J., & Adams-Campbell, L. L. (2003). Onset of natural menopause in African American women. *American journal of public health, 93*(2), 299-306.

Pang, B., Lee, L., & Vaithyanathan, S. (2002). *Thumbs up?: sentiment classification using machine learning techniques.* Paper presented at the Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10.

KAYNAKLAR DİZİNİ (Devam Ediyor)

Phillips, K.-A., Glendon, G., & Knight, J. A. (1999). Putting the risk of breast cancer in perspective. In: Mass Medical Soc.

Reynolds, D. A. (2009). Gaussian Mixture Models. *Encyclopedia of biometrics, 741*.

Rish, I. (2001). *An empirical study of the naive Bayes classifier*. Paper presented at the IJCAI 2001 workshop on empirical methods in artificial intelligence.

Roweis, S. T., & Saul, L. K. (2000). Nonlinear dimensionality reduction by locally linear embedding. *Science, 290*(5500), 2323-2326.

Rubinstein, R. Y., & Kroese, D. P. (2016). *Simulation and the Monte Carlo method* (Vol. 10): John Wiley & Sons.

Rubinstein, W. S., O'Neill, S. M., Peters, J. A., Rittmeyer, L. J., & Stadler, M. P. (2002). Mathematical modeling for breast cancer risk assessment. *Breast Cancer, 16*(8).

Salakhutdinov, R., & Hinton, G. (2009). *Deep boltzmann machines*. Paper presented at the Artificial intelligence and statistics.

Saritas, M. M., & Yasar, A. (2019). Performance analysis of ANN and Naive Bayes classification algorithm for data classification. *International Journal of Intelligent Systems and Applications in Engineering, 7*(2), 88-91.

Schietgat, L., Vens, C., Struyf, J., Blockeel, H., Kocev, D., & Džeroski, S. (2010). Predicting gene function using hierarchical multi-label decision tree ensembles. *BMC bioinformatics, 11*(1), 2.

Scikit Learn. (2020,Şubat 28). Decision Trees. Retrieved from https://scikit-learn.org/stable/modules/tree.html

Seber, G. A., & Lee, A. J. (2012). *Linear regression analysis* (Vol. 329): John Wiley & Sons.

Seide, F., Li, G., & Yu, D. (2011). *Conversational speech transcription using context-dependent deep neural networks*. Paper presented at the Twelfth annual conference of the international speech communication association.

Sharma, S. K., & Wang, X. (2019). Towards massive machine type communications in ultra-dense cellular IoT networks: Current issues and machine learning-assisted solutions. *IEEE Communications Surveys & Tutorials*.

Sheather, S. J. (2004). Density estimation. *Statistical science*, 588-597.

Stark, G. F., Hart, G. R., Nartowt, B. J., & Deng, J. (2019). Predicting breast cancer risk using personal health data and machine learning models. *PloS one, 14*(12), e0226765.

Statista. (2020, July 20). Age of mothers at first birth in the U.S. by Hispanic origin 2018. Retrieved from https://www.statista.com/statistics/260386/mean-age-of-mothers-at-first-birth-in-the-united-states-in-by-hispanic-origin/

Stigler, S. M. (1989). Francis Galton's account of the invention of correlation. *Statistical Science*, 73-79.

Stigler, S. M. (1997). Regression towards the mean, historically considered. *Statistical methods in medical research, 6*(2), 103-114.

KAYNAKLAR DİZİNİ (Devam Ediyor)

Suarez, M., Perez-Castejon, M., Jimenez, A., & Domper, M. (2002). Early diagnosis of recurrent breast cancer with EDG-PET in patients with progressive elevation of serum tumor markers. *The Quarterly Journal of Nuclear Medicine and Molecular Imaging, 46*(2), 113.

Sutskever, I., Hinton, G. E., & Taylor, G. W. (2009). *The recurrent temporal restricted boltzmann machine.* Paper presented at the Advances in neural information processing systems.

Tamar, A., Wu, Y., Thomas, G., Levine, S., & Abbeel, P. (2016). *Value iteration networks.* Paper presented at the Advances in Neural Information Processing Systems.

Tao, S., Zhang, T., Yang, J., Wang, X., & Lu, W. (2015). *Bearing fault diagnosis method based on stacked autoencoder and softmax regression.* Paper presented at the 2015 34th Chinese Control Conference (CCC).

Teh, Y. W. (2010). Dirichlet Process. In.

Tesauro, G. (1995). Temporal difference learning and TD-Gammon. *Communications of the ACM, 38*(3), 58-68.

Tseng, Y.-J., Huang, C.-E., Wen, C.-N., Lai, P.-Y., Wu, M.-H., Sun, Y.-C., . . . Lu, J.-J. (2019). Predicting breast cancer metastasis by using serum biomarkers and clinicopathological data with machine learning technologies. *International journal of medical informatics, 128*, 79-86.

Tyrer, J., Duffy, S. W., & Cuzick, J. (2004). A breast cancer prediction model incorporating familial and personal risk factors. *Statistics in medicine, 23*(7), 1111-1130.

van Asperen, C. J., Jonker, M., Jacobi, C., van Diemen-Homan, J., Bakker, E., Breuning, M., . . . De Bock, G. (2004). Risk estimation for healthy women from breast cancer families: new insights and new strategies. *Cancer Epidemiology and Prevention Biomarkers, 13*(1), 87-93.

Van Der Maaten, L., Postma, E., & Van den Herik, J. (2009). Dimensionality reduction: a comparative. *J Mach Learn Res, 10*(66-71), 13.

Vapnik, V., & Chapelle, O. (2000). Bounds on error expectation for support vector machines. *Neural computation, 12*(9), 2013-2036.

Vogel, K. J., Atchley, D. P., Erlichman, J., Broglio, K. R., Ready, K. J., Valero, V., . . . Arun, B. (2007). BRCA1 and BRCA2 genetic testing in Hispanic patients: mutation prevalence and evaluation of the BRCAPRO risk assessment model. *Journal of Clinical Oncology, 25*(29), 4635-4641.

Von Luxburg, U. (2007). A tutorial on spectral clustering. *Statistics and computing, 17*(4), 395-416.

Wang, J., Gou, L., Shen, H.-W., & Yang, H. (2018). Dqnviz: A visual analytics approach to understand deep q-networks. *IEEE transactions on visualization and computer graphics, 25*(1), 288-298.

KAYNAKLAR DİZİNİ (Devam Ediyor)

Wang, X., Huang, Y., Li, L., Dai, H., Song, F., & Chen, K. (2018). Assessment of performance of the Gail model for predicting breast cancer risk: a systematic review and meta-analysis with trial sequential analysis. *Breast Cancer Research*, *20*(1), 18.

Watkins, C. J., & Dayan, P. (1992). Q-learning. *Machine learning, 8*(3-4), 279-292.

Yıldız, K., Çamurcu, Y., & Doğan, B. (2010). Veri madenciliğinde temel bileşenler analizi ve Negatifsiz matris çarpanlarına ayırma tekniklerinin karşılaştırmalı analizi. *Akademik Bilişim*, 10-12.

Zhang, Z. (2016). Naïve Bayes classification in R. *Annals of translational medicine, 4*(12).

Zhao, H. (2002). Global stability of bidirectional associative memory neural networks with distributed delays. *Physics Letters A, 297*(3-4), 182-190.

Zhou, Y. (2019). Asymptotics and Interpretability of Decision Trees and Decision Tree Ensembles.