Türkiye Klinikleri Biyoistatistik Dergisi
Turkiye Klinikleri Journal of Biostatistics Turkiye Klinikleri J Biostat. 2022;14(1):32-44

ORİJİNAL ARAŞTIRMA ORIGINAL RESEARCH DOI: 10.5336/biostatic.2021-85755

Gail Modeli ile Makine Öğrenmesi Algoritmalarının Meme Kanseri Risk Değerlendirmesinde Karşılaştırılması: Metodolojik Çalışma

Comparison of the Machine Learning Algorithms in Breast Cancer Risk Assessment with the Gail Model: Methodological Study

Berfu PARÇALI[a], Fezan MUTLU[a]

[a]Eskişehir Osmangazi Üniversitesi Sağlık Bilimleri Enstitüsü, Biyoistatistik ABD, Eskişehir, Türkiye

ÖZET Amaç: Meme kanserinin; meme dokusu içiresinde yer alan süt kanallarının doku hücrelerinde oluştuğu bilinmektedir. Süt kanallarını oluşturan bu hücrelerin kontrolsüz olarak artmasına ise duktal hiperplazi denir. Bir kadında, yaşamı süresince invazif (yayılma eğilimi olan) meme kanseri gelişme riskinin %13,3 olduğu bilinmektedir. Meme kanserinin oluşma riski yaşa bağlı olarak artmaktadır. Gail modeli; meme kanserinde temel faktörleri değerlendiren, genel olarak kabul görmüş kanser riski değerlendirme modelidir. Bu çalışmada, Gail modeli baz alınarak, makine öğrenmesi yöntemlerinin meme kanseri risk değerlendirmesinde karşılaştırılması amaçlanmıştır. **Gereç ve Yöntemler:** İlk olarak, veri setine Gail modeli uygulanmış ve risk faktörü belirlenmiş ve %80 eğitim, %20 test olmak üzere ayrı eğitim test veri seti oluşturulmuştur. Daha sonra oluşturulan bu veri setlerine k-en yakın komşu, yapay sinir ağları (YSA), destek vektör makinesi [support vector machine (SVM)] ve naive Bayes (NB) algoritmaları uygulanmış ve uygukanan yöntemlerin risk tahmin sonuçları karşılaştırılmıştır. Bulgular: Karşılaştırma sonuçlarına göre %80 eğitim, %20 test veri seti için sınıflandırma performansı en düşükten en yükseğe doğru sırasıyla SVM [eğri altında kalan alan (area under the curve "AUC") =0,911], NB (AUC=0,939) ve YSA (AUC=0,949) şeklindedir. **Sonuç:** Meme kanserinin erken aşamada teşhis edilmesi; tedavi yöntemlerinin sayısını, tedavinin başarıya ulaşma oranını ve hayatta kalma şansını artırmaktadır. Meme kanseri risk hesaplamasında makine öğrenmesi yöntemlerinin etkili olduğu görülmüştür.

Anahtar kelimeler: Gail modeli; meme kanseri; makine öğrenmesi; yapay sinir ağları; destek vektör makinesi

ABSTRACT Objective: Breast cancer; it occurs in the tissue cells of the milk ducts in the breast tissue. The uncontrolled increase in the cells forming the milk ducts is called ductal hyperplasia. It is known that the risk of developing invasive (with a tendency to spread) breast cancer in a woman during her lifetime is 13.3%, and the risk of developing breast cancer increases with age. The Gail model is a well accepted cancer risk assessment model which evaluates the main factors in breast cancer. The aim of this study is to compare machine learning methods in breast cancer risk assessment based on the Gail model. **Material and Methods:** Firstly, the risk factor was determined by the application of the Gail model into the data set, discrete training test data sets were presented which is 80% train and 20% test. Afterwards, k-nearest neighbor, artificial neural network (ANN), support vector machine (SVM) and naive Bayes (NB) algorithms applied on each set and risk estimation results were compared. **Results:** Classification performance from the lowest to the highest for 80% training and 20% test data set according to the comparison results is as follows; SVM [area under the curve (AUC)=0.911], NB (AUC=0.939) and ANN (AUC=0.949). **Conclusion:** Early diagnosis of breast cancer increases the number of possible treatments, the success rate of the treatments and the chance of survival. It has been seen that machine learning algorithms effective in breast cancer risk calculation.

Keywords: Gail model; breast cancer; machine learning; artificial neural network; support vector machine

Meme kanseri, meme dokusu içerisinde yer alan süt kanallarının doku hücrelerinde oluşmaktadır. Süt kanallarını oluşturan doku hücrelerinin kontrolsüz olarak artmasına duktal hiperplazi denir.[1]

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.[2] Meme kanseri teşhislerinin yaklaşık %18'i 40'lı yaşlardaki kadınlar arasındadır ve meme kanseri olan kadınların %77'si teşhis konulduğunda 50 yaşının üzerindedir.[3]

Correspondence: Fezan MUTLU
Eskişehir Osmangazi Üniversitesi Sağlık Bilimleri Enstitüsü, Biyoistatistik ABD, Eskişehir, Türkiye
E-mail: fsahin@ogu.edu.tr

Peer review under responsibility of Turkiye Klinikleri Journal of Biostatistics.

Received: 18 Aug 2021 ***Received in revised form:*** 23 Jan 2022 ***Accepted:*** 24 Jan 2022 ***Available online:*** 27 Jan 2022

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

MEME KANSERİNDE RİSK TAHMİN MODELLERİ

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

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.[4]

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.[5]

GAİL MODELİ

Bir risk tahmini olan Gail modelinin geçerliliği, meme kanserini önleme stratejisinde uygun klinik kararlar vermek için çok önemlidir.

Meme kanseri riskini mümkün olan en doğru biçimde tespit edebilmek için meme kanseriyle ilişkili çok sayıda risk faktörünü değerlendirmek önemlidir.[6]

1989 yılında geliştirilen Gail modeli (Model 1), 1999 yılında Ulusal Cerrahi Adjuvan Meme ve Bağırsak Projesi istatistikçileri tarafından, sadece invaziv meme kanseri gelişme mutlak riskini yansıtmak için "Breast Cancer Risk Assessment Tool (BCRAT)" (Gail 2) modeli olarak değiştirilmiştir.[7]

En önemli değişiklik, risk faktörlerine, atipik hiperplazi ile birlikte olan meme biyopsilerinin de eklenmesidir.[8]

BCRAT modelinde, 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.[6]

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

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

Burada:

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

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

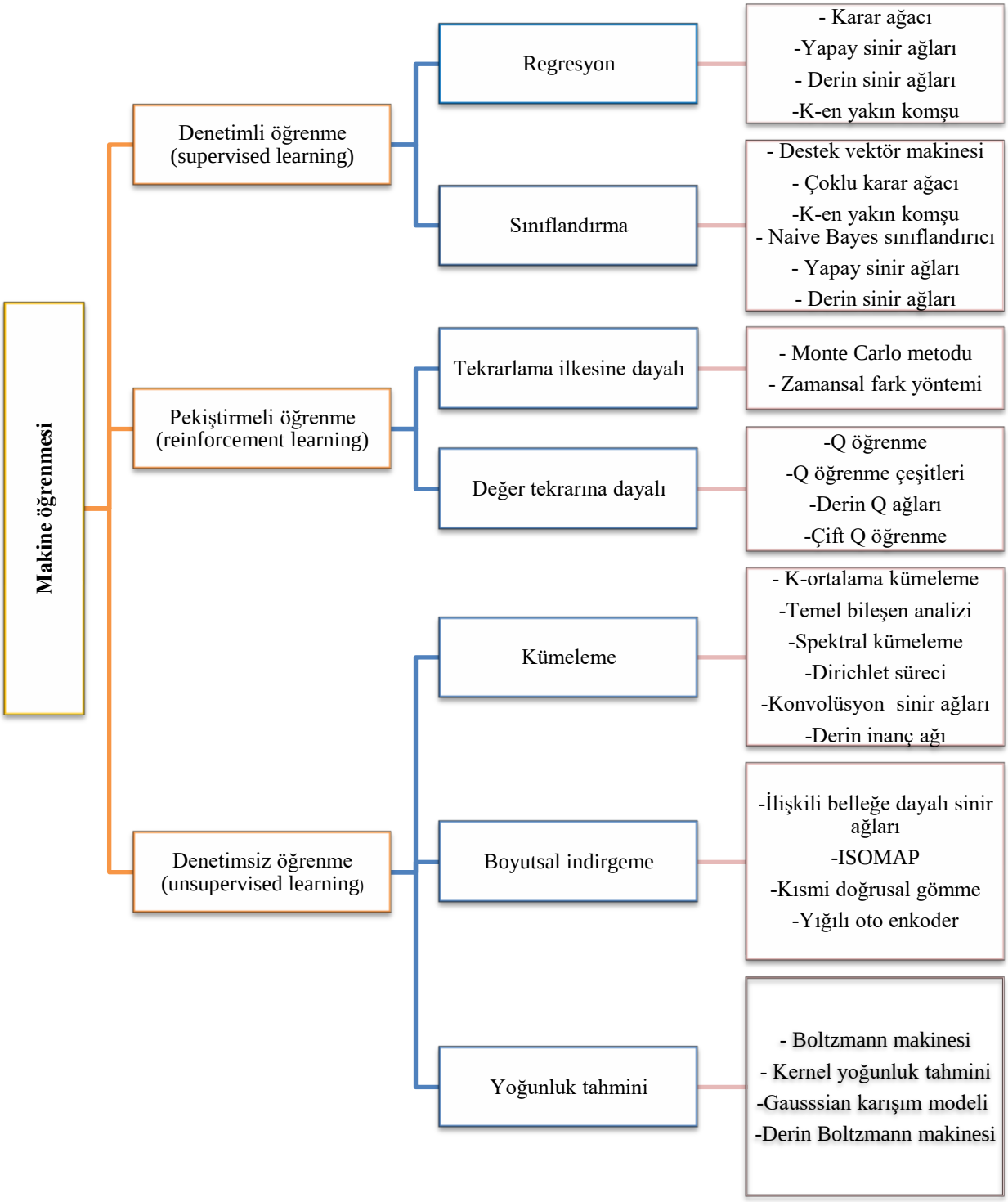
$h_1(t)$: Risk faktörleri bilinmeyen bireyin yaş endeksli riskini temsil etmektedir.

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

$r(t)$: Rölatif (göreceli) riski temsil etmektedir. α: Bireyin riski hesaplandığı andaki yaşını temsil etmektedir.[9]

MAKİNE ÖĞRENMESİ

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 zekânın bir alt dalıdır.[10] Makine öğrenmesi yöntemleri Şekil 1'de gösterilmiştir.[11]



ŞEKİL 1: Makine öğrenmesi yöntemleri.

İstatistiksel yöntemlerde ve yapay sinir ağlarında (YSA), 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.[12]

DENETİMLİ ÖĞRENME (SUPERVISED LEARNING)

Denetimli öğrenmenin temel amacı, girdiden çıktıya kadar doğru değerlerin denetmen (süpervizör) tarafından sağlanmasıdır. Denetimli öğrenmede, hedefler hakkında bilgi, eğitim veri kümesine (S) ait bir dizi hedef değişkenden elde edilir.[13]

REGRESYON

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

Regresyon; 2 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 ve regresyon analizindeki temel amaç, en az değişkeni kullanarak bağımlı değişken ile bağımsız değişkenler arasındaki ilişkiyi açıklayabilmek ve genel olarak kabul edilebilen bir model kurmaktır.[14] 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.[15]

K-En Yakın Komşu

K-en yakın komşu [k-nearest neighbor (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.[16]

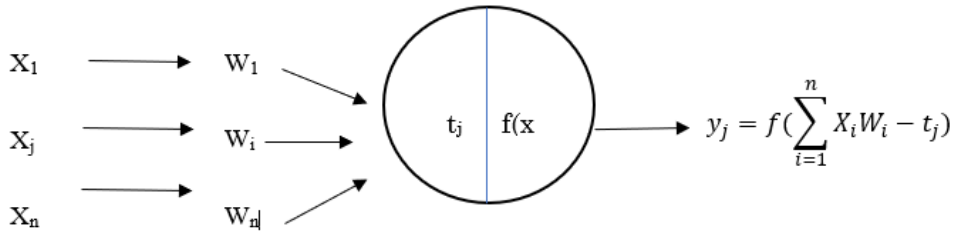
Bir öznitelik uzayında, A ve B noktaları arasındaki mesafeyi ölçmek için literatürde en çok kullanılan uzaklık Öklid uzaklığıdır. m boyutlu bir öznitelik uzayında, A ve B özellik vektörleri A=(x_1, x_2, …, x_m) ve B=(y_1, y_2, …, y_m) olarak gösterilsin. A ve B arasındaki mesafenin Öklid metriği:

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

formülü ile hesaplanmaktadır.[17]

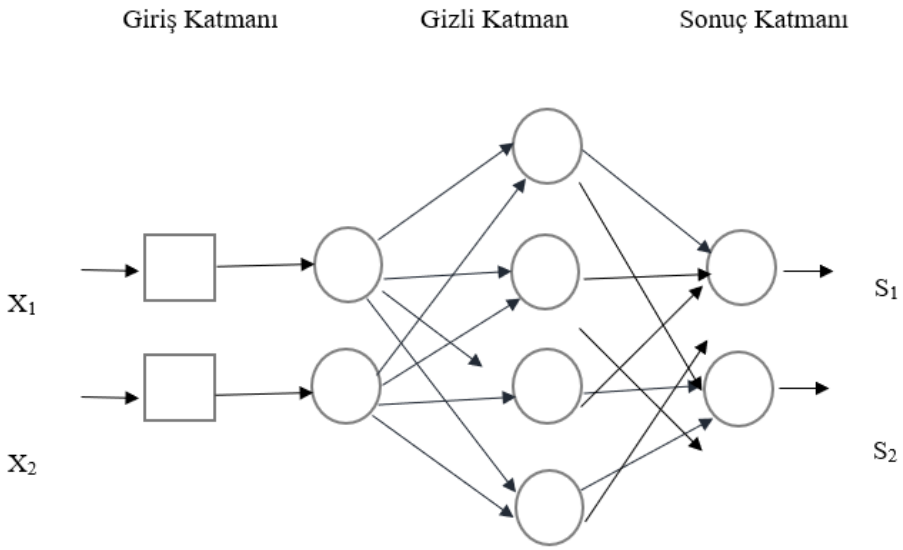
YSA

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.[18] YSA, yapısal olarak 5 unsurdan oluşmaktadır ve bu unsurlar Şekil 2'de gösterilmiştir: Girdiler, ağırlıklar, toplam fonksiyonu, aktivasyon fonksiyonu ve çıktıdır.[19]



ŞEKİL 2: 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.[20] Sinir ağı Şekil 3'te gösterilmiştir.[21]



ŞEKİL 3: Sinir ağı.

SINIFLANDIRMA

Makine öğrenmesinin temel birimi olan sınıflandırma yönteminin amacı, bilinmeyen bir veri parçasını bilinen bir gruba atamaktır.[22]

DESTEK VEKTÖR MAKİNESİ

Destek vektör makinesi [support vector machine (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.[23]

Radyal temel işlevi (radyal temel işlevi ya da Gauss Kernel), SVM eğitiminde en çok kullanılan çekirdektir.

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

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.[24]

NAİVE BAYES SINIFLANDIRICI

Naive Bayes (NB), yaygın olarak kullanılan Bayes formülü veya Bayes teorimi olarak bilinen olasılık teoremi ile sağlanır.[25]

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

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.[26]

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.[27]

GEREÇ VE YÖNTEMLER

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 normal dağılım kullanılarak 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 dâhil olmak üzere toplamda 11 ırkın dağılımı elde edilmiştir.[28]

Türetilen verilerin menarş yaşı da ırklara göre belirlenmiştir.[29] Irklara göre menarş yaşı için ırkların menopoz ve menarş yaşları üzerine yapılan çalışmalar referans alınmıştır.[30]

Türetilen verilerin ilk canlı doğum yaşı, “Statista” sitesindeki ırkların ilk doğum yaşları dağılım oranlarına göre elde edilmiştir.[31]

Veri seti senaryosu Tablo 1’de gösterilmiştir.

Verilerin analizinde Python programı (Python Software Foundation, 9450 SW Gemini Dr., ECM# 90772, Beaverton, OR 97008, USA) kullanılmıştır.

TABLO 1: Veri seti senaryosu.

Değişken	Verinin açıklaması	Verinin özellikleri
r	Simülasyon sayısı	5
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

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.

MAKİNE ÖĞRENMESİ ALGORİTMALARININ UYGULANMASI

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

Türetilen veriler makine öğrenmesi algoritmalarına sunulmadan önce minimum-maksimum normalizasyon işlemi gerçekleştirilmiş ve veriler [0,1] aralığında yeniden ölçeklendirilmiştir.

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

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

Bu çalışmada, risk değişkeni Gail modeli ile hesaplanmış, daha sonra elde edilen risk değişkeni, makine öğrenmesi algoritmaları için hedef olarak belirlenmiştir.

Makine öğrenmesi algoritmaları hedef değişkeni (target), risk sütunu olarak belirlenmiştir. Eğitim ve test veri seti oluşturulmuştur. Sınıflandırma performansının karşılaştırılması için %80 eğitim ve %20 test olmak üzere, eğitim ve test veri seti oluşturulmuştur. Eğitim setine makine öğrenmesi algoritmaları ve "cross validation" (k=10, tekrar sayısı=10) uygulanmıştır.

BULGULAR

k-NN SINIFLANDIRMA SONUCUNA İLİŞKİN BULGULAR

Veri setine ilk önce k-NN algoritması uygulanmıştır. Algoritmada k değeri 5 seçilmiştir. Sınıflandırma sonuçları Tablo 2'de belirtilmiştir.

TABLO 2: k-NN algoritması sınıflandırma sonucu.

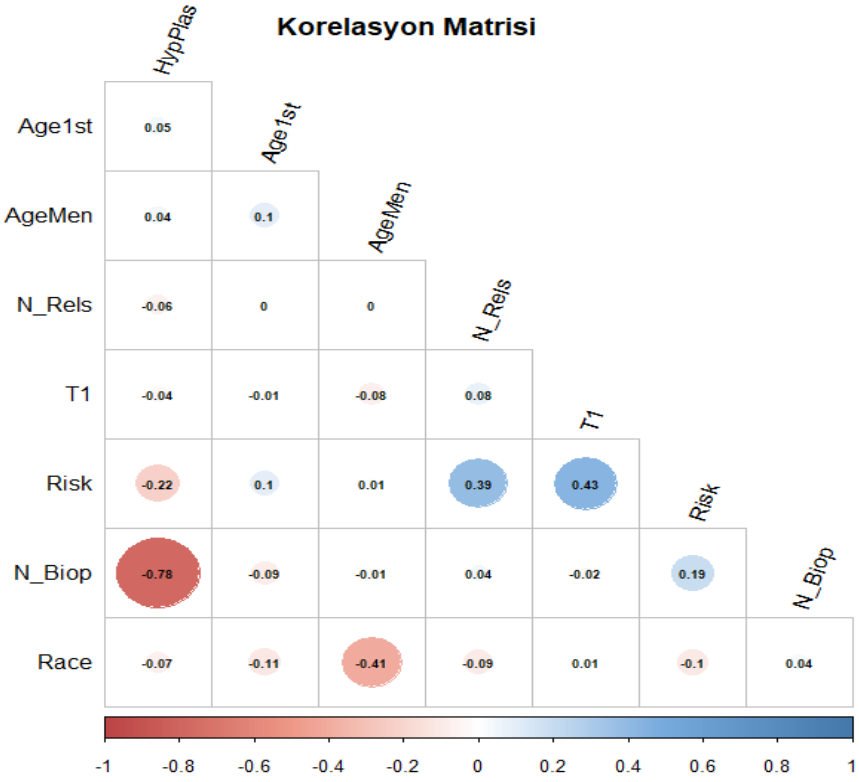
k-NN	0	1
0	%21,1	%16,8
1	%78,9	%83,2

k-NN: k en yakın komşu.

k-NN sınıflandırma istatistiksel olarak anlamlı bulunamamıştır (p=0,253). Sınıflandırma sonucuna göre yanlış negatif [false negatives (FN)]: %78,9, yanlış pozitif [false positives (FP)]: %16,8'dir.

YSA SINIFLANDIRMA SONUCUNA İLİŞKİN BULGULAR

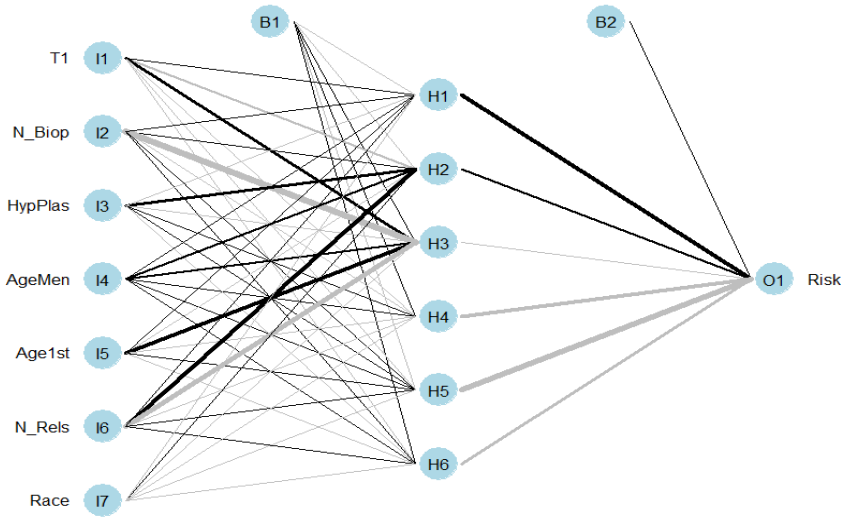
YSA'da optimal sonuç elde etmek ve ilişkili değişkenleri tespit etmek için Şekil 4'te yer alan korelasyon matrisi oluşturulmuştur.



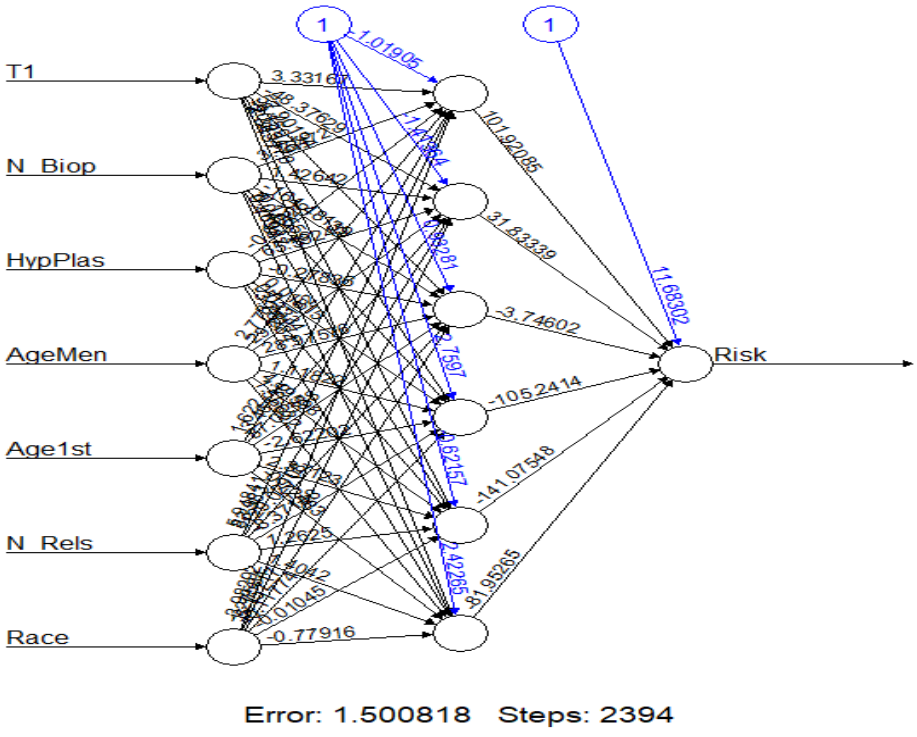
ŞEKİL 4: Veri seti için korelasyon matrisi.

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=-0,22, p<0,001) ilişki vardır. Risk ve hasta akraba sayısı değişkenleri arasında pozitif yönde orta düzeyde ilişki (r=0,39, p<0,001), âdet yaşı ve ırk değişkenleri arasında negatif yönde orta düzeyde ilişki (r=-0,41, p<0,001), risk ve yaş değişkenleri arasında pozitif yönde orta düzeyde ilişki vardır (r=0,43, p<0,001). Biyopsi sayısı ve hiperplazi (r=-0,78, p<0,001) değişkenleri arasında negatif yönde yüksek düzeyde ilişki vardır.

Şekil 5 ve Şekil 6'da YSA mimarisi elde edilmiştir. Sinir ağı, 7 nöronlu tek katmandan oluşmaktadır. Sınıflandırma sonucu Tablo 3'te belirtilmiştir.



ŞEKİL 5: Yapay sinir ağı mimarisi.



ŞEKİL 6: Yapay sinir ağı mimarisi matematiksel gösterimi.

TABLO 3: YSA algoritması sınıflandırma sonucu.

YSA	0	1
0	%63,3	%5,8
1	%36,7	%94,2

YSA: Yapay sinir ağları.

YSA sınıflandırması istatistiksel olarak anlamlıdır (p<0,001). Sınıflandırma sonucuna göre FN: %5,8, FP: %36,7'dir.

SVM SINIFLANDIRMA SONUCUNA İLİŞKİN BULGULAR

Sınıflandırma sonucu Tablo 4'te belirtilmiştir.

TABLO 4: SVM algoritması sınıflandırma sonucu.

SVM	0	1
0	%75,0	%8,5
1	%25,0	%91,5

SVM: Destek vektör makinesi.

SVM sınıflandırması istatistiksel olarak anlamlıdır (p<0,001). Sınıflandırma sonucuna göre FN: %8,5, FP: %25,0'dır.

NB SINIFLANDIRMA SONUCUNA İLİŞKİN BULGULAR

Sınıflandırma sonucu Tablo 5'te belirtilmiştir.

TABLO 5: NB algoritması sınıflandırma sonucu.

NB	0	1
0	%72,7	%7,8
1	%27,3	%92,2

NB: Naive Bayes.

NB sınıflandırması istatistiksel olarak anlamlıdır (p<0,001). Sınıflandırma sonucuna göre FN: %7,8, FP: %27,3'tür.

ALICI İŞLEM KARAKTERİSTİKLERİ EĞRİSİ SONUÇLARINA İLİŞKİN BULGULAR

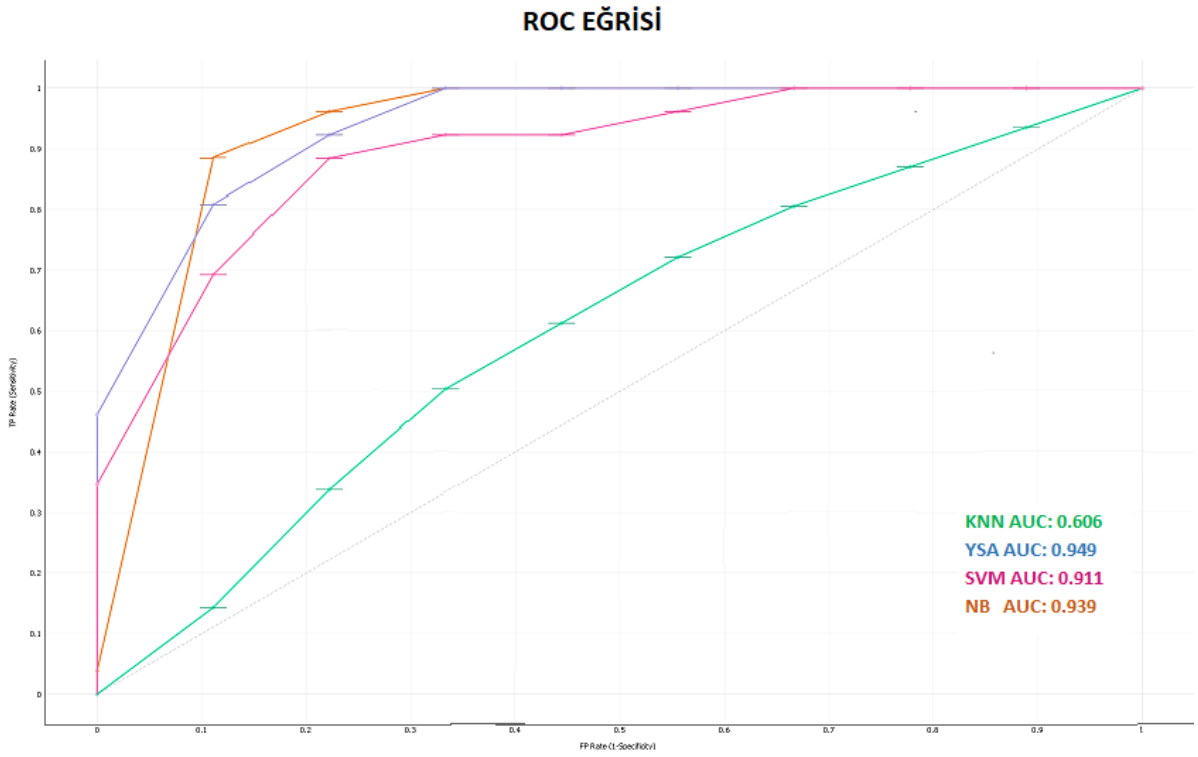
Tablo 6'da, performans değerlendirme sonuçlarına göre sınıflandırma karşılaştırmaları verilmiştir. Sınıflandırma karşılaştırma sonucuna göre en yüksek sınıflandırma sonucu YSA algoritmaları ile elde edilmiştir. Doğruluk yüzdesi baz alındığında en yüksek performanstan en düşüğe doğru sırasıyla sınıflandırma sonuçları; YSA, NB, SVM ve k-NN şeklindedir.

TABLO 6: Sınıflandırma karşılaştırmaları.

Model	Doğruluk yüzdesi (accuracy)	p değeri	F1	Duyarlılık (sensitivity)	Özgüllük (specificity)	ROC eğrisi (AUC)
k-NN	0,753	0,253	0,738	0,724	0,280	0,606
YSA	**0,880**	**p<0,001**	**0,883**	**0,888**	**0,762**	**0,949**
SVM	0,893	p<0,001	0,888	0,887	0,643	0,911
NB	0,893	p<0,001	0,890	0,888	0,674	0,939

ROC: Alıcı işlem karakteristikleri; AUC: Eğri altında kalan alan; k-NN: k en yakın komşu; YSA: Yapay sinir ağları; SVM: Destek vektör makinesi; NB: Naive Bayes.

Şekil 7'de alıcı işlem karakteristikleri [receiver operating characteristic (ROC)] eğrisi gösterilmiştir. Şekle göre YSA yöntemine ait ROC eğrisi altında kalan alan 0,949, NB yöntemine ait ROC eğrisi altında kalan alan [area under the curve (AUC)] 0,939, SVM yöntemine ait ROC AUC 0,911 ve k-NN yöntemine ait ROC AUC 0,606'dır. Sınıflandırma sonuçlarına göre $0,90<AUC_{YSA}<1,0$ olduğundan YSA yönteminin sınıflandırması mükemmeldir. Kullanılan yöntemler arasında en düşük performanslı sınıflandırma yönteminin k-NN (AUC=0,606) olduğu görülmüştür.



ŞEKİL 7: ROC eğrisi.

TARTIŞMA

Stark ve ark.nın çalışmalarında, 5 yıllık meme kanseri riski BCRAT, YSA, lojistik regresyon ve doğrusal diskriminant analizi yöntemleriyle karşılaştırılmıştır. Karşılaştırma sonucunda BCRAT doğruluk yüzdesinin %56,3, lojistik regresyon ve doğrusal diskriminant analizi doğruluk yüzdelerinin %61,3 ve YSA doğruluk yüzdesinin %60,8 olduğu gözlenmiştir.[32]

Tseng ve ark.nın çalışmasında, meme kanseri metastazını tahmin etmek için makine öğrenmesi algoritmalarından rastgele orman, NB, SVM ve lojistik regresyon yöntemleri karşılaştırılmıştır. Karşılaştırma sonucuna göre rastgele orman AUC=0,746, NB AUC=0,648, SVM AUC=0,645 ve lojistik regresyon AUC=0,581 olduğu gözlenmiştir.[33]

Ganggayah ve ark.nın çalışmasında, 1993-2016 yılları arasında Malezya'daki Malaya Üniversitesi Tıp Merkezinden elde edilen meme kanseri veri seti ile meme kanseri hayatta kalma oranını belirlemede; karar ağacı, rastgele orman, YSA, "extreme boost", lojistik regresyon ve SVM yöntemleri karşılaştırılmıştır. Karşılaştırma sonucunda karar ağacı doğruluk yüzdesinin %72, YSA doğruluk yüzdesinin %84, lojistik regresyon ve SVM'nin doğruluk yüzdelerinin %85, rastgele orman doğruluk yüzdesinin %86, "extreme boost" doğruluk yüzdesinin %87 olduğu gözlenmiştir.[34]

Ming ve ark.nın çalışmasında, Markov Zinciri Monte Carlo Genelleştirilmiş Doğrusal Karma Model [Markov Chain Monte Carlo Generalized Linear Mixed Models (MCMCGLMM)], AdaBoost ve rastgele orman ve BOADICEA modeli kullanılmıştır. BOADICEA %63,9, AdaBoost %88,9, MCMCGLMM %85,1, rastgele orman %84,3 tahmin doğrulu ile sınıflandırılmıştır.[35]

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-NN, YSA, SVM ve NB makine öğrenme algoritmalarının performansları karşılaştırılmıştır. Çalışma-

nın amacı, verileri her algoritmanın verimlilik ve etkinlik açısından sınıflandırmadaki doğruluğunu; doğruluk, Kappa, duyarlılık, özgüllük ve ROC eğrisi açısından değerlendirmektir.

SONUÇ

Bu çalışmada, R Studio Version 3.6.3 (RStudio Delaware Public Benefit Corporation (PBC) 250 Northern Ave, Boston, MA 02210) programı kullanılarak 500 adet veri normal dağılımdan türetilmiştir. Türetilen veri senaryosunda Gail modeli ile meme kanseri riski hesaplanmış, daha sonra riski hesaplanan verilerin makine öğrenmesi yöntemlerine göre sınıflandırılması amaçlanmıştır. Sınıflandırma sonuçları %80 eğitim, %20 test olmak üzere eğitim ve test veri setinde karşılaştırılmıştır.

Sınıflandırma performansını ölçmek için doğruluk yüzdesi, F1, duyarlılık, özgüllük ve ROC eğrisi ölçütleri kullanılmıştır.

Karşılaştırma sonuçları Tablo 6'da gösterilmiştir. Bu karşılaştırmaya göre en yüksek sınıflandırma sonucu YSA algoritmaları ile elde edilmiştir. Doğruluk yüzdesi baz alındığında, en yüksek performanstan en düşüğe doğru sırasıyla sınıflandırma sonuçları; YSA, NB, SVM ve k-NN şeklindedir.

Değerlendirme sonuçları sırasıyla k-NN (AUC=0,606), SVM (AUC=0,911), NB (AUC=0,939) ve YSA'dır (AUC=0,949). YSA'nın en düşük hata oranı ile en yüksek doğruluğu verdiği görülmüştür.

Finansal Kaynak

Bu çalışma sırasında, yapılan araştırma konusu ile ilgili doğrudan bağlantısı bulunan herhangi bir ilaç firmasından, tıbbi alet, gereç ve malzeme sağlayan ve/veya üreten bir firma veya herhangi bir ticari firmadan, çalışmanın değerlendirme sürecinde, çalışma ile ilgili verilecek kararı olumsuz etkileyebilecek maddi ve/veya manevi herhangi bir destek alınmamıştır.

Çıkar Çatışması

Bu çalışma ile ilgili olarak yazarların ve/veya aile bireylerinin çıkar çatışması potansiyeli olabilecek bilimsel ve tıbbi komite üyeliği veya üyeleri ile ilişkisi, danışmanlık, bilirkişilik, herhangi bir firmada çalışma durumu, hissedarlık ve benzer durumları yoktur.

Yazar Katkıları

Fikir/Kavram: *Fezan Mutlu, Berfu Parçalı;* ***Tasarım:*** *Berfu Parçalı;* ***Denetleme/Danışmanlık:*** *Fezan Mutlu;* ***Veri Toplama ve/veya İşleme:*** *Berfu Parçalı;* ***Analiz ve/veya Yorum:*** *Berfu Parçalı, Fezan Mutlu;* ***Kaynak Taraması:*** *Berfu Parçalı;* ***Makalenin Yazımı:*** *Berfu Parçalı, Fezan Mutlu;* ***Eleştirel İnceleme:*** *Berfu Parçalı, Fezan Mutlu.*

KAYNAKLAR

1. Barton MB. Breast cancer screening: benefits, risks, and current controversies. Postgrad Med. 2005;118(2):27-46. [Crossref] [PubMed]
2. Phillips KA, Glendon G, Knight JA. Putting the risk of breast cancer in perspective. N Engl J Med. 1999;340(2):141-4. [Crossref] [PubMed]
3. Chapman C, Murray A, Chakrabarti J, Thorpe A, Woolston C, Sahin U, et al. Autoantibodies in breast cancer: their use as an aid to early diagnosis. Ann Oncol. 2007;18(5):868-73. [Crossref] [PubMed]
4. Boyle P, Mezzetti M, La Vecchia C, Franceschi S, Decarli A, Robertson C. Contribution of three components to individual cancer risk predicting breast cancer risk in Italy. Eur J Cancer Prev. 2004;13(3):183-91. [Crossref] [PubMed]
5. Dumitrescu R, Cotarla I. Understanding breast cancer risk-where do we stand in 2005? J Cell Mol Med. 2005;9(1):208-21. [Crossref] [PubMed] [PMC]
6. Amir E, Freedman OC, Seruga B, Gareth Evans D. Assessing women at high risk of breast cancer: a review of risk assessment models. J Natl Cancer Inst. 2010;102(10):680-91. [Crossref] [PubMed]
7. Costantino JP, Gail MH, Pee D, Anderson S, Redmond CK, Benichou J, et al. Validation studies for models projecting the risk of invasive and total breast cancer incidence. J Natl Cancer Inst. 1999;91(18):1541-8. [Crossref] [PubMed]
8. Karakayali FY, Ekici Y, Sevmiş Ş, Pehlivan S, Arat Z, Moray G. Meme kanseri için risk belirlenmesinde Gail modeli [Gail model for determination of the risk factors of breast cancer]. Turkish Journal of Surgery. 2007;23(4):129-35. [Link]
9. Gail MH, Brinton LA, Byar DP, Corle DK, Green SB, Schairer C, et al. Projecting individualized probabilities of developing breast cancer for white females who are being examined annually. J Natl Cancer Inst. 1989;81(24):1879-86. [Crossref] [PubMed]

10. Akay EÇ. Ekonometride yeni bir ufuk: büyük veri ve makine öğrenmesi [A new horizon in econometrics: big data and machine learning]. Sosyal Bilimler Araştırma Dergisi. 2018;7(2):41-53. [Link]
11. Sharma SK, Wang X. Towards massive machine type communications in ultra-dense cellular IoT networks: current issues and machine learning-assisted solutions. IEEE Communications Surveys & Tutorials. 2019. [Crossref]
12. Clark IA, Niehaus KE, Duff EP, Di Simplicio MC, Clifford GD, Smith SM, et al. First steps in using machine learning on fMRI data to predict intrusive memories of traumatic film footage. Behav Res Ther. 2014;62:37-46. [PubMed] [PMC]
13. Hastie T, Tibshirani R, Wainwright M. Statistical Learning with Sparsity: The Lasso and Generalizations. 1st ed. Boca Raton: Chapman and Hall/CRC; 2015. [Crossref]
14. Coşkun S, Kartal M. Lojistik regresyon analizinin incelenmesi ve diş hekimliğinde bir uygulaması. Cumhuriyet Üniversitesi Diş Hekimliği Fakültesi Dergisi. 2004;7(1):42-50. [Link]
15. Seber GA, Lee AJ. Linear Regression Analysis. Vol. 329. 2nd ed. Hoboken, N.J: John Wiley & Sons; 2012.
16. Keller JM, Gray MR, Givens JA. A fuzzy k-nearest neighbor algorithm. IEEE Transactions on Systems, Man, and Cybernetics. 1985;(4):580-5. [Crossref]
17. Hu LY, Huang MW, Ke SW, Tsai CF. The distance function effect on k-nearest neighbor classification for medical datasets. Springerplus. 2016;5(1):1304. [Crossref] [PubMed] [PMC]
18. Hamzaçebi C, Kutay F. Yapay sinir ağlari ile Türkiye elektrik enerjisi tüketiminin 2010 yılına kadar tahmini [Electric consumption forecasting of Turkey using artificial neural networks up to year 2010]. J Fac Eng Arch Gazi Univ. 2004;19(3):227-33. [Link]
19. Kalogirou SA. Applications of artificial neural networks in energy systems. Energy Conversion and Management. 1999;40(10):1073-87. [Crossref]
20. Koç ML, Balas CE, Arslan A. Taş dolgu dalgakıranların yapay sinir ağları ile ön tasarımı [Preliminary design of ruble mound breakwaters by using artificial neural networks]. İMO Teknik Dergi. 2004;15(74):3351-75. [Link]
21. Bose NK, Garga AK. Neural network design using Voronoi diagrams. IEEE Trans Neural Netw. 1993;4(5):778-87. [Crossref] [PubMed]
22. Harrington P. Machine Learning in Action. 1st ed. Shelter Island, NY: Manning Publications Co; 2012.
23. Lin CF, Wang SD. Fuzzy support vector machines. IEEE Trans Neural Netw. 2002;13(2):464-71. [Crossref] [PubMed]
24. Chang YW, Hsieh CJ, Chang KW, Ringgaard M, Lin CJ. Training and testing low-degree polynomial data mappings via linear SVM. Journal of Machine Learning Research. 2010;11(4):1471-90. [Link]
25. Lewis DD. Naive (Bayes) at forty: the independence assumption in information retrieval. 10th European Conference on Machine Learning Chemnitz, Germany, April 21-23, 1998 Proceedings. Springer; 1998. p.4-15. [Crossref]
26. Zhang Z. Naïve Bayes classification in R. Ann Transl Med. 2016;4(12):241. [Crossref] [PubMed] [PMC]
27. Rish I. An empirical study of the naive Bayes classifier. IJCAI 2001 workshop on empirical methods in artificial intelligence. 2001. p.41-6. [Link]
28. United States Census Bureau. Race and Hispanic Origin. Erişim tarihi: 2020 Temmuz 2020. Erişim linki: [Link]
29. Palmer JR, Rosenberg L, Wise LA, Horton NJ, Adams-Campbell LL. Onset of natural menopause in African American women. Am J Public Health. 2003;93(2):299-306. [Crossref] [PubMed] [PMC]
30. Ahuja M. Age of menopause and determinants of menopause age: a PAN India survey by IMS. J Midlife Health. 2016;7(3):126-31. [Crossref] [PubMed] [PMC]
31. Statista. Age of mothers at first birth in the U.S. by Hispanic origin 2018. Erişim tarihi: 20 Temmuz 2020. Erişim linki: [Link]
32. Stark GF, Hart GR, Nartowt BJ, Deng J. Predicting breast cancer risk using personal health data and machine learning models. Plos One. 2019;14(12):e0226765. [Crossref] [PubMed] [PMC]
33. Tseng YJ, Huang CE, Wen CN, Lai PY, Wu MH, Sun YC, et al. Predicting breast cancer metastasis by using serum biomarkers and clinicopathological data with machine learning technologies. Int J Med Inform. 2019;128:79-86. [Crossref] [PubMed]
34. Ganggayah MD, Taib NA, Har YC, Lio P, Dhillon SK. Predicting factors for survival of breast cancer patients using machine learning techniques. BMC Med Inform Decis Mak. 2019;19(1):48. [Crossref] [PubMed] [PMC]
35. Ming C, Viassolo V, Probst-Hensh N, Dinov ID, Chappuis PO, Katapodi MC. Machine learning-based lifetime breast cancer risk reclassification compared with the BOADICEA model: impact on screening recommendations. Br J Cancer. 2020;123(5):860-7. [Crossref] [PubMed] [PMC]