T.C.
PAMUKKALE ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ
BİLGİSAYAR MÜHENDİSLİĞİ ANABİLİM DALI

MEME KANSERİ TAHMİNİNDE MAKİNE ÖĞRENMESİ ALGORİTMALARI VE AUTOML

YÜKSEK LİSANS TEZİ

ARSLAN KARAKAYA

DENİZLİ, ŞUBAT - 2024

T.C.
PAMUKKALE ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ
BİLGİSAYAR MÜHENDİSLİĞİ ANABİLİM DALI

MEME KANSERİ TAHMİNİNDE MAKİNE ÖĞRENMESİ ALGORİTMALARI VE AUTOML

YÜKSEK LİSANS TEZİ

ARSLAN KARAKAYA

DENİZLİ, ŞUBAT - 2024

Bu tezin tasarımı, hazırlanması, yürütülmesi, araştırmalarının yapılması ve bulgularının analizlerinde bilimsel etiğe ve akademik kurallara özenle riayet edildiğini; bu çalışmanın doğrudan birincil ürünü olmayan bulguların, verilerin ve materyallerin bilimsel etiğe uygun olarak kaynak gösterildiğini ve alıntı yapılan çalışmalara atfedildiğine beyan ederim.

Arslan KARAKAYA

ÖZET

MEME KANSERİ TAHMİNİNDE MAKİNE ÖĞRENMESİ ALGORİTMALARI VE AUTOML
YÜKSEK LİSANS TEZİ
ARSLAN KARAKAYA
PAMUKKALE ÜNİVERSİTESİ FEN BİLİMLERİ ENSTİTÜSÜ
BİLGİSAYAR MÜHENDİSLİĞİ ANABİLİM DALI

(TEZ DANIŞMANI:PROF. DR. SEZAİ TOKAT)

DENİZLİ, ŞUBAT - 2024

Makine öğrenmesi tıp dahil olmak üzere birçok alanda önemli rol oynamaktadır. Özellikle sağlık alanında makine öğrenmesi çeşitli hastalıkların teşhis ve tedavisine yardımcı olan önemli bir araç olarak hizmet vermektedir. Makine öğrenmesinin önemli uygulamalarından biri özellikle kadınlar arasında dünya genelinde kanserden ölüm nedenleri arasında önde gelen meme kanseri gibi hastalıkların erken teşhisidir. Meme kanserinin teşhisinde yüksek doğruluk elde etmek tedavinin etkinliğini ve hastanın genel durumunu doğrudan etkilediği için çok önemlidir. Veri analizi yapılarak farklı makine öğrenmesi modelleri ile tıbbi veriler doğru şekilde analiz edilebilir. Bu çalışmada çeşitli makine öğrenmesi algoritmalarının (Lojistik Regresyon, Karar Ağacı, KNN, Naive Bayes, Destek Vektör Makinesi, Rassal Orman, Stokastik Gradyan İniş, Adaboost, XGBoost, LightGBM, Yapay Sinir Ağları) önce ön işleme,veri artırma, hiperparametre optimizasyonlu olarak sonra AutoML yöntemlerinin (TPOT, H2O, MLJAR) veri artırma ve ön işlemeler ile Wisconsin Meme Kanseri (Teşhis) veriseti üzerinde makine öğrenmesi modeli geliştirmeye yönelik iki farklı yaklaşımın karşılaştırmalı performansı araştırılmıştır. Standardizasyon, SMOTE ve rassal az örnekleme, hiperparametre optimizasyonu model sonuçlarını genel olarak iyileştirmiştir. AutoML yöntemleri makine öğrenmesi algoritmalarının hepsinden daha yüksek doğruluk ve F1-skor değerleri elde etmiştir. AutoML yöntemleri model oluşturma sürecini otomatikleştirmek ve modeli optimize etmek için etkili bir araç olma potansiyeli göstermektedir.

ANAHTAR KELİMELER: Makine Öğrenmesi, Otomatik Makine Öğrenmesi, Meme Kanseri Tahmini

ABSTRACT

BREAST CANCER DIAGNOSIS WITH MACHINE LEARNING ALGORITHMS AND AUTOML
MSC THESIS
ARSLAN KARAKAYA
PAMUKKALE UNIVERSITY INSTITUTE OF SCIENCE
COMPUTER ENGİNEERİNG

(SUPERVISOR:PROF. DR. SEZAİ TOKAT)

DENİZLİ, FEBRUARY 2024

Machine learning plays an important role in many fields including medicine. Especially in the field of healthcare, machine learning serves as an important tool to help diagnose and treat various diseases. One of the important applications of machine learning is the early detection of diseases such as breast cancer which is the leading cause of cancer death worldwide especially among women. Achieving high accuracy in the diagnosis of breast cancer is very important as it directly affects the effectiveness of treatment and general condition of patient. Medical data can be accurately analyzed with data analysis and different machine learning models. In this study the comparative performance of two different approaches to machine learning model development on the Wisconsin Breast Cancer (Diagnosis) dataset is investigated using various machine learning algorithms (Logistic Regression, Decision Tree, KNN, Naive Bayes, Support Vector Machine, Random Forest, Stochastic Gradient Descent, Adaboost, XGBoost, LightGBM, Artificial Neural Networks) with preprocessing, data augmentation, hyperparameter optimization and AutoML methods (TPOT, H2O, MLJAR) with data augmentation and preprocessing. Standardization, SMOTE and random undersampling and hyperparameter optimization generally improved model results. AutoML methods achieved higher accuracy and F1-score values than all machine learning algorithms. AutoML methods show the potential to be an effective tool for automating the model building process and optimizing the model.

KEYWORDS: Machine Learning, AutoML, Breast Cancer Prediction

İÇİNDEKİLER

Sayfa

ŞEKİL LİSTESİ

Sayfa

TABLO LİSTESİ

Sayfa

ÖNSÖZ

Tez sürecinin tamamında her türlü desteği verip yardımcı olan Sayın Prof. Dr. Sezai TOKAT'a teşekkür etmeyi bir borç bilirim.

1. GİRİŞ

1.1 Tezin Amacı

Bu çalışmanın amacı meme kanseri veriseti üzerinde çeşitli algoritmalar kullanarak karşılaştırmalı bir analiz üzerinden kanser tespitinde hangisi algoritmanın hangi hiperparametreler ile daha performanslı olduğunu bulmak, meme kanseri verilerinde otomatik makine öğrenmesi metotlarının performansını değerlendirmek ve klasik yöntemlere göre verimliliğini araştırmaktır. Yapılan çalışmaların çoğunda makine öğrenmesi sınıflandırma algoritmalarının varsayılan parametreler ile hiperparametre optimizasyonu ya da ön işleme yapılmadan kullanıldığı görülmüştür. Bu bağlamda veri analiziyle makine öğrenmesi algoritmalarını farklı hiperparametre değerleriyle kullanarak, derin öğrenme kullanarak ve ek olarak otomatik makine öğrenmesi kullanarak iyi huylu veya kötü huylu kanser tahmini yapmada daha iyi performans veren yöntemin hangisinin hangi ayarlarda olduğunu belirleme amaçlanmıştır. Bununla birlikte farklı sınıflandırma algoritmaları ile AutoML performans karşılaştırması da yapılmıştır dolayısıyla ön işleme ve hiperparametre optimizasyonu gibi ara işlemlerin etkisi, otomatik makine öğrenmesi metotlarının kanser verileri için performansı araştırılmıştır.

1.2 Tezin Önemi

Kanserin erken teşhis edilmesi hayat kurtarabilir, daha etkili bir tedaviye ve hayatta kalma şansının önemli ölçüde artmasına imkan sağlar (John ve Broggio 2019). Fakat kanser vakalarının neredeyse yarısı sadece ileri aşamada fark edilmektedir. Geliştirilmiş erken teşhis ile kanserde hayatta kalma oranları önemli ölçüde artırılabilir (Crosby ve diğ. 2022).

Makine öğrenmesi tıp alanında önemlidir çünkü hasta sonuçlarını iyileştirme ve maliyetleri azaltma potansiyeline sahiptir. Tıbbi verilerin giderek daha karmaşık ve hacimli hale gelmesi tıp uzmanlarının bu verilerden anlamlı bilgiler çıkarmasını

zorlaştırmaktadır. Makine öğrenmesi algoritmaları çok miktarda veriyi analiz etmek ve insan analistlerin kolayca göremediği örüntüleri belirlemek için tasarlanmıştır. Tıp alanında makine öğrenmesi kişiselleştirilmiş tedavi planlarının geliştirilmesine, hastalıkların erken teşhisine ve hasta sonuçlarının tahmin edilmesine yardımcı olabilir. Ayrıca tıp uzmanlarının bazı durumlar açısından daha yüksek risk altında olan hastaları belirlemelerine yardımcı olarak daha odaklı önleyici tedbirler almalarını sağlayabilir. Bununla birlikte makine öğrenmesi teknikleri tıbbi analizlerin doğruluğunu ve verimliliğini artırmak için kullanılabilir bu da kanser gibi hastalıkların daha erken tespit edilmesini sağlayabilir.

Kanser tahmininde doğruluk erken tanı için çok önemlidir. Bu çalışmada varsayılan ayarlarla yapılan çalışmaların aksine farklı hiperparametre değerleriyle ve AutoML yöntemleriyle kanser tahmini yapılacak ve sonuçlar karşılaştırılmıştır. Böylece kanser tahmininde daha verimli sonuçlar elde edilip edilemeyeceği, AutoML araçlarının tahmin performansı gözlenecektir. Makine öğrenmesi yöntemlerinin en verimli şekilde kullanılması için veri ön işleme ve veri temizleme, uygun özellikleri seçme ve oluşturma, uygun bir model ailesi seçme, model hiperparametrelerini optimize etme, son işleme gibi ara işlemlerin gerçekleştirilmesi gerekir. Otomatik makine öğrenmesi yaklaşımları, makine öğrenmesinin deneme-yanılma yönlerini otomatikleştirmeye odaklanan bir araştırma dalıdır. Otomatik makine öğrenmesi, makine öğrenmesi verimliliğini artırmak, makine öğrenmesi üzerine araştırmaları hızlandırmak için ve makine öğrenmesini makine öğrenmesi bilmeyen uzmanlara kullanılabilir hale getiren yöntemler ve süreçler sağlar. Kanser tahmininde AutoML performansı diğer yöntemler ile karşılaştırılarak bu alandaki araştırmaların, uygulamaların AutoML ile daha verimli şekilde yapılıp yapılamayacağı, en iyi sonuç hangi yöntemde elde edileceği gözlenecektir. Araştırmacıları, geliştiricileri, kullanıcıları desteklemek için makine öğrenmesinin aşamalı otomasyonunun klasik yöntemlere göre daha verimli şekilde sonuçlar üretip üretemeyeceği araştırılmıştır.

1.3 Tezin Akışı

Tezin ilk bölümünde bu çalışmanın öneminden, amacından ve akışından bahsedilmiştir. İkinci bölümde kanser ve meme kanseri hakkında bilgiler verilmiştir. Üçüncü bölümde problem tanımı ile birlikte literatür çalışmaları belirtilerek literatürde yapılmış olan çalışmalar incelenmiş sonuçlarıyla birlikte yer verilmiştir. Dördüncü bölüm bu çalışmada kullanılan araçları ve yöntemleri içermektedir. Yapay zeka, makine öğrenmesi ve algoritmalardan bahsedilip uygulamada kullanılan araç ve yöntemler hakkında bilgiler verilmiştir. Beşinci bölümde gerçekleştirilen uygulama sonuçlarından bahsedilmiştir, veri analizi ve sonuçlar grafikler ile desteklenmiştir. Altıncı bölümde yapılan çalışmayla elde edilen sonuçlar değerlendirilmiştir.

2. KANSER

2.1 Kanser Nedir?

Kanser, vücuttaki hücrelerin kontrolsüz bir şekilde büyümesi ve diğer bölgelere yayılmasıyla ortaya çıkan bir hastalıktır. Normal hücreler vücudun ihtiyaç duyduğu şekilde çoğalır ve hasarlı hücreler ölür ancak anormal hücreler büyür ve çoğalarak tümör adı verilen doku kitlelerini oluşturur. Tümörler iyi huylu veya kötü huylu olabilir, kötü huylu tümörler çevre dokuları istila eder ve metastaz adı verilen bir süreçle vücudun her yerinde yeni tümörler oluşturur. Katı tümörler birçok kanserde yaygındır ancak lösemi gibi kan kanserleri katı tümör üretmez. İyi huylu tümörler yaşamı tehdit etmese de büyük boyutlara ulaşabilir ve özellikle beyin gibi bölgelerde bulunduklarında ciddi semptomlara neden olabilirler. Bununla birlikte iyi huylu tümörler çıkarıldıklarında kanserli tümörlerin aksine genellikle tekrarlamazlar.

Kanser hücresi genotiplerinin büyük kısmı kötü huylu büyümeyi teşvik eden hücresel fizyolojideki altı temel değişiklikten kaynaklanmaktadır.

- Büyüme sinyallerinde kendi kendine yeterlilik
- Büyüme karşıtı sinyallere duyarsızlık
- Programlı hücre ölümünden kaçınma
- Sınırsız çoğalma potansiyeli
- Sürekli anjiyogenez
- Doku istilası ve metastaz

Bu değişiklikler yeni yeteneklerdir, bir tümörün gelişimi sırasında yeni kazanılır ve hücre ile dokulardaki doğuştan gelen antikanser savunma

mekanizmasının etkili bir şekilde atlatılmasını temsil eder. Bu tür altı yetenek insan tümörlerinin hepsinde olmasa da çoğunda mevcuttur (Hanahan ve Weinberg 2000).

Kanserler tipik olarak kaynaklandığı organ veya vücut bölümüne göre adlandırılır, ancak buna neden olan belirli hücre türüne göre de sınıflandırılabilir. Başlıca kanser türleri arasında karsinom, sarkom, miyeloma, lösemi, lenfoma, germ hücreli tümör, blastom bulunmaktadır.

- Karsinomlar epitelyal hücrelerden kaynaklanır ve genellikle meme, prostat ve kolon gibi organlarda görülür.
- Sarkomlar kemik, kıkırdak ve yağ gibi bağ dokularından kaynaklanır.
- Lenfomalar ve lösemiler bağışıklık sistemini ve kan hücrelerini etkiler ve kemik iliğindeki olgunlaşmamış hücrelerden kaynaklanır.
- Germ hücreli tümörler pluripotent hücrelerden ortaya çıkar ve genellikle testis veya yumurtalıkta görülür.
- Blastomlar olgunlaşmamış öncül hücrelerden veya embriyonik dokudan gelişir ve çocuklarda daha yaygındır.

2.2 Risk Faktörleri

Kanser çeşitli faktörlere bağlı olarak gelişebilen karmaşık bir hastalıktır. Bazı insanlar genetik olarak belirli kanser türlerine daha yatkın olabilir (Lynch ve diğ. 2004). Yaşam tarzı faktörleri, çevresel etkenler ve enfeksiyonlar da bu hastalığın gelişimine katkıda bulunabilir. Sigara içmek, aşırı alkol kullanımı, kötü beslenme, hareketsiz yaşam tarzı ve aşırı kilolu veya obez olmak kanser riskini artırdığı gösterilen yaşam tarzı faktörleridir (Anand ve diğ. 2008). İşyerinde zararlı kimyasallara maruz kalma, hava kirliliği ve güneşten gelen UV radyasyonu, kanser gelişimiyle bağlantılı olan diğer çevresel faktörlerdir. HPV, hepatit B ve HIV gibi bazı enfeksiyonlar da kanser riskini artırmaktadır (Williams 2018).

2.3 Tedavi

Hastalara kanser teşhisi konduğunda onkologlar hastaların karmaşık tedavi kararlarını yönlendirmelerine yardımcı olur. Her birinin faydaları, komplikasyonları ve belirsiz etkileri olan birçok tedavi alternatifi vardır. Bilimsel bulgular veya hastaya özgü faktörler gibi çeşitli faktörler nihai tedavi seçimini etkiler. Tek bir hastaya birden fazla tedavi seçeneği sunulabilir bu da karar verme sürecini karmaşıklaştırabilir (Panje ve diğ. 2018). Abbas ve Rehman (2018) cerrahi, kemoterapi ve radyoterapi gibi geleneksel tedavi seçenekleri yaygın olarak kullanılırken dünya çapındaki deneysel çalışmalarla yeni tedavi seçeneklerinin araştırıldığından, hormon tabanlı tedavi, anti-angiyojenik tedaviler, kök hücre tedavileri ve dendritik hücre tabanlı immünoterapi de dahil olmak üzere geleneksel ve modern tedavi yöntemlerinden bahsetmişlerdir ve tedavinin başarısının kanserin türüne, tümörün bulunduğu bölgeye ve ilerleme evresine bağlı olduğunu belirtmişlerdir.

2.3.1 Cerrahi Yöntemler

Cerrahi müdahale kanser hastaları için önemli bir tedavi seçeneğidir. Kanser hücrelerinin tamamı cerrahi olarak çıkarılabilirse hasta o kanserden kurtulmuş olacaktır (Nguyen ve Tsien 2013). Cerrahi prosedürler kanserin yeri ve evresine bağlı olarak yapılır. Kanser erken bir aşamada tespit edilirse ve vücudun bir bölümünde lokalize ise genellikle ilk tedavi seçeneğidir. Bazı durumlarda tekrarlama olasılığını azaltmak için ameliyattan sonra ek tedaviler önerilebilir. Cerrahi müdahale tümörün tamamının çıkarılması amacıyla genellikle kanser tek bir bölgede olduğunda ve vücudun diğer bölgelerine yayılmadığı durumlarda tekrarlamayı önlemek için, tümör yoğunluğunu azaltma amacıyla tümörün boyutunu küçültmek ve kemoterapi veya radyoterapi gibi diğer tedavilerin etkisini artırmak için özellikle de tümörün tamamen çıkarılması bir organa veya vücuda zarar verme riski taşıdığında, kanser semptomlarını hafifletmek için ağrı veya baskıya neden olan tümörleri çıkarmak ve nefes alma, yutma güçlüğü gibi ilişkili semptomları hafifleterek hastanın yaşam kalitesini artırmak için uygulanabilir.

2.3.2 Kemoterapi

Kemoterapi hücre bölünme süreçlerine müdahale ederek kanser hücrelerini öldürmek için ilaçların kullanıldığı bir tedavidir. İlaçlar ağızdan veya enjeksiyon yoluyla verilebilir ve nerede olurlarsa olsunlar kanser hücrelerini hedef almak için kan dolaşımında hareket ederler. İlaçlar kanser hücrelerinin DNA'sına zarar vererek bölünmelerini ve büyümelerini önler (Woods and Turchi 2013). Tedavi genellikle vücudun iyileşmesine izin vermek için aralarında dinlenme dönemleri olan döngüler halinde verilir. Kemoterapi etkili bir tedavi yöntemi olmakla birlikte bulantı, yorgunluk ve saç dökülmesi gibi yan etkilere neden olabilir, bu etkiler ilaç tedavisi ve yaşam tarzı değişiklikleri ile yönetilebilir (Kayl ve Meyers 2006). Kemoterapi metastazları tedavi etmek için kullanılan sistemik bir yaklaşımken cerrahi müdahale ile radyoterapi birincil tümörleri ele almak için kullanılan lokal yöntemlerdir. Başarılı bir iyileştirici tedavi için dozaj ve zamanlamada tutarlılık çok önemlidir ancak hafifletici tedavi almak isteyen hastalar için doz düzenlemeleri gerekli olabilir. Kemoterapi sağlıklı dokuya zararı en aza indirirken kanserli hücreleri ortadan kaldırmak için tasarlanmıştır ancak ilaçların etkinliği tümörün histolojisine ve sınıflandırmasına bağlı olarak değişebilir (Greenhalgh and Symonds 2014).

2.3.3 Radyoterapi

Radyoterapi (radyasyon/ışın tedavisi) kanser hücrelerini öldürmek ve tümörlerin boyutunu küçültmek için yüksek dozda iyonlaştırıcı radyasyon kullanılan bir tedavi yöntemidir. Radyoterapi iyonlaştırıcı radyasyon kullanarak kanserli hücreleri öldürmek için kullanılan bir tedavi yöntemidir. Farklı tedavi türlerinin kendine özgü avantajları ve dezavantajları olduğu gibi radyoterapinin yorgunluk, saç dökülmesi, ağrı, bulantı, kusma, ağız kuruluğu, duygularda değişiklikler gibi yan etkileri olabilir. Yan etkiler genellikle geçicidir fakat uzun vadeli yan etkiler haftalarca veya aylarca sürebilir. Işınlamanın neden olduğu komplikasyonlar hasta için rahatsız edici olabilir. Radyoterapi kanserin türüne ve evresine bağlı olarak iyileştirici veya hafifletici bir tedavi olarak kullanılabilir (Baykara 2015).

2.3.4 Kök Hücre

Kök hücre nakli yüksek doz kemoterapi veya radyasyon tedavisi nedeniyle hücreleri tahrip olmuş kişilerde kan oluşturan kök hücreleri geri kazandırarak kanseri tedavi etme yöntemidir. Bu kök hücreler olgunlaşarak beyaz kan hücreleri, kırmızı kan hücreleri ve trombositler gibi çeşitli kan hücresi türlerine dönüşür. Kök hücre nakilleri otolog, allojenik ve sinjeneik olmak üzere farklı şekillerde gerçekleştirilir. Kök hücre nakilleri doğrudan kanserle mücadele etme şekli olmasa da tedaviden sonra vücudun kök hücre üretme yeteneğinin geri kazanılmasına yardımcı olur. Kök hücre nakli süreci birkaç ay sürebilir ve doktorlar düzenli kan sayımı yaparak yeni kan hücrelerinin ortaya çıkışını sıklıkla değerlendirir (Spencer 2021). Kök hücre temelli tedaviler kanser tedavisi için umut verici bir yöntemdir. Kök hücrelerin doğal olarak tümörleri hedef aldığı ve terapötik ajanları doğrudan kanser hücrelerine iletmek üzere tasarlanabildiği gösterilmiştir. Bu yaklaşım geleneksel tedavilere kıyasla daha etkili olma ve sağlıklı dokulara daha az zarar verme potansiyeline sahiptir (Stuckey and Shah 2014).

2.3.5 Gen Terapisi

Geleneksel kanser tedavilerinin kemoterapi ve radyoterapiye yanıt vermede düşük etkinlik ve zorluk gibi dezavantajları vardır. Ancak gen terapisi sağlıklı hücrelere zarar vermeden yalnızca tümör hücrelerini hedefleyerek umut verici bir alternatif sunmaktadır. Viral ve viral olmayan gen transferi gibi farklı teknikler çeşitli kanserler üzerinde test edilmiş ve milyonlarca kanser hastasının tedavisi için büyük potansiyel göstermiştir (Yahya ve Alqadhi 2021).

2.4 Kanseri Önlemek

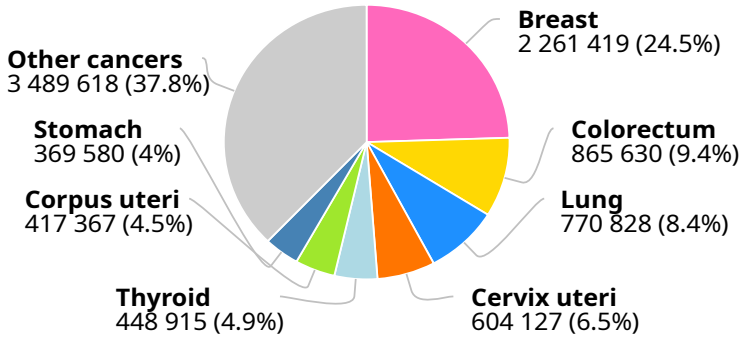
Hastalıkların önlenmesi iki yöntemle gerçekleşebilir bunlar birincil korunma yani risk faktörlerinden kaçınarak koruyucu faktörleri artırmak ile ikincil korunma yani erken teşhis ve müdahaledir (Song ve diğ. 2018). Bayo ve diğ. (2019) kanser vakalarının birincil korunma tedbirleri ile yarı yarıya azaltılabileceğini belirtmişler ve sigara, alkol, yeme alışkanlığı, obezite, fiziksel aktivite, mesleki ve çevresel

faktörler, ultraviyole radyasyon, enfeksiyonlar ve sosyoekonomik faktörler gibi başlıca değiştirilebilir risk faktörlerini inceleyerek birincil korunmanın kanser vakalarını azaltmak için basit ve etkili bir araç olduğu sonucuna varmışlardır.

2.5 Meme Kanseri

Meme kanseri meme dokusundaki hücrelerde meydana gelen bir kanser türüdür. Hem erkeklerde hem de kadınlarda görülebilse de kadınlarda daha yaygındır. Meme kanseri dünya çapında kadınlarda en sık görülen kanserdir ancak hayatta kalma oranları iyileşmekte ve meme kanserinden kurtulanların sayısı artmaktadır (Ewertz and Jensen 2011).

Meme kanseri dünya çapında artan vaka ve ölüm oranlarıyla önemli bir küresel sağlık sorunudur. 2020 yılında yaklaşık 2.3 milyon yeni meme kanseri vakasının teşhis edilmesi ve 684,996 meme kanseri ölümünün gerçekleşmesi tahmin edilmiştir. Ölümler açısından bakıldığında meme kanseri dünya genelinde kansere bağlı ölümlerin %6,7'sini oluşturmaktadır. Meme kanseri 159 ülkede kadınlarda en sık teşhis edilen kanserdir ve kanserden ölümlerin önde gelen nedenidir bunu görülme sıklığı açısından bağırsak ve akciğer kanseri takip etmektedir (Sung ve diğ. 2021). Şekil 2.1'de dünya, kadın, tüm yaşlar kategorisinde 2020 yeni vaka sayısı bulunmaktadır.



Şekil 2.1: 2020 Dünya yeni vaka sayısı, kadın, tüm yaşlar.

2.5.1 Tanı

Meme kanseri teşhisi tipik olarak fiziksel muayeneler, görüntüleme testleri ve doku biyopsilerinin bir kombinasyonunu kapsamaktadır. Meme kanseri için başlıca tanı testleri mamografi, ultrason, manyetik rezonans görüntüleme (MRI) ve biyopsidir.

Fiziksel muayene meme kanseri tanısının önemli bir parçasıdır çünkü meme kanserine işaret edebilecek herhangi bir fiziksel belirti veya semptom gözlenmesine olanak tanır. Meme kanseri genellikle bir yumru olarak ortaya çıkar ancak memenin görünümünde deri çekilmesi, meme ucu değişikliği ve renk değişikliği gibi değişikliklerin yanı sıra memede ağrı, şişlik, ciltte kızarıklık ve lenf bezi büyümesi de yaygın belirtilerdir (Zgajnar 2017).

Mamografi amacı meme kanserinin erken tedavisini sağlamak, hayatta kalma oranını yükseltmek olan meme kanserinin tespiti için en önemli görüntüleme yöntemidir. Tarama ve tanı için iki tür mamografi vardır. Tarama mamografisi küçük kanserleri kendi kendine muayene veya klinik meme muayenesi yoluyla bulunmadan önce erken yakalamak için periyodik olarak yapılırken, tanısal mamografi hastalar şişlik veya meme ucu akıntısı gibi semptomlarla başvurduğunda kullanılır. (Sardanelli ve diğ. 2016). Mamografi genellikle etkili bir tanı aracı olsa da yanlış pozitiflik olasılığı vardır. Yanlış pozitiflik gerçekte kanser olmadığı halde mamografinin kanser varlığını göstermesi durumunda ortaya çıkar. Yanlış pozitif sonuçlar hasta için endişe ve stresin yanı sıra rahatsız edici ve invaziv olabilen biyopsiler gibi gereksiz takip testlerine de neden olabileceği gibi gereksiz ameliyatları veya yaşam kalitesi üzerinde olumsuz etkileri olabilecek diğer tıbbi prosedürleri içerebilir (Seely snd Alhassan 2018). Mamografi 60'lı yaşlardaki kadınlar için 40'lı yaşlardakilere göre daha faydalıdır çünkü 40-50 yaşlarındaki kadınların on yıl içinde yanlış pozitif sonuç alma olasılığı %61'dir (McDonald ve diğ. 2016).

Ultrasonografi iç organların ve dokuların görüntülerini oluşturmak için yüksek frekanslı ses dalgalarını kullanan tıbbi bir görüntüleme tekniğidir. Ultrason klinik uygulamalarda genel olarak 2 megahertz ile 10 megahertz arasında değişen, insanların duyabileceğinden daha yüksek frekanslı ses dalgaları kullanır (Aldrich

2007). Ultrasonun taşınabilirlik, gerçek zamanlı görüntüleme gibi avantajları vardır ve bununla birlikte sınırlı bir görüş alanı, bazı yapıları görüntülemede zorluk ve uzman bir operatör ihtiyacı gibi sınırlamaları da vardır (Hooley ve diğ. 2013). Mamografi ile ultrasonografi birlikte kullanılırsa meme kanserinin erken teşhis oranını etkili bir şekilde artırabilir (Zhang ve diğ. 2019).

Manyetik rezonans görüntüleme vücudun hemen hemen tüm iç yapısının ayrıntılı görüntülerini oluşturmak için manyetik alan ve radyo dalgalarını kullanan tıbbi bir görüntüleme tekniğidir. Manyetik rezonans görüntüleme tümörler, yaralanmalar ve nörolojik bozukluklar gibi çeşitli durumları tespit ve teşhis etmek için kullanılmaktadır. Lehman ve Schnall (2005) manyetik rezonans görüntülemenin diğer görüntüleme türlerinin tespit edemediği kanserleri tespit etmede çok başarılı olduğunu belirtmişlerdir.

Meme biyopsisi mikroskop altında incelenmek üzere küçük bir meme dokusu örneği almak için uygulanan bir prosedürdür. Genellikle meme muayenesi, mamografi veya ultrason incelemesinde bir şişlik veya anormallik tespit edildiğinde uygulanır (Jain ve diğ. 2017). Biyopsi yöntemlerinden bazıları ince iğne aspirasyonu, çekirdek iğne biyopsisi ve vakum destekli biyopsidir. İnce iğne aspirasyonu memedeki şişlikten küçük bir hücre örneği elde etmek için ince bir iğne kullanılan minimal invaziv bir yöntemdir. İnce iğne aspirasyonu basit ve düşük maliyetli bir prosedürdür ancak sonuçları yorumlamak için bir uzmana ihtiyaç vardır ve doku örneklemesinde hata yapılma ihtimali daha yüksektir (Calhoun ve Anderson 2014). Çekirdek iğne biyopsisinde ince iğne aspirasyonuna kıyasla daha büyük bir iğne kullanılır. Doku örneği görüntüsünün büyütülmesine olanak sağladığı için ve ince iğne aspirasyonundan daha yüksek doğruluğa sahip olduğundan çekirdek iğne biyopsisi meme hücrelerinde hastalık olup olmadığını kontrol etmek için ince iğne aspirasyonundan daha iyi bir yöntemdir (Park ve Hong 2014). Vakum destekli meme biyopsisi 1995 yılında çekirdek biyopsisini daha iyi hale getirmek için geliştirilmiş, otomatik biyopsi tabancası kullanılan bir yöntemdir (Park ve Hong 2014). Vakum destekli meme biyopsisi daha doğru bir tanıya ve lezyonun tamamen çıkarılmasına olanak sağladığından çekirdek iğne biyopsisine göre daha iyi bir alternatiftir, hata ve yeniden biyopsi olasılığını azaltır (Park ve Kim 2011).

2.5.2 Risk Faktörleri

Yaş, meme kanseri oluşumuyla ilişkili önemli bir risk faktörüdür. Meme kanserlerinin çoğunluğu 40 yaş ve üzeri kadınlarda teşhis edilmekte olup vakaların %50'si 50-69 yaş arası kadınlarda görülmektedir (Kamińska ve diğ. 2015, Rojas and Stuckey 2016).

Ailesinde meme kanseri öyküsü olan bir kadının hastalığa yakalanma riski yüksektir. Kadının meme kanserinden etkilenmiş bir veya daha fazla anne veya kız kardeş gibi birinci derece akrabası varsa risk daha da yüksektir (Sun ve diğ. 2017). Meme kanserine yakalanma olasılığının artması özellikle BRCA1 ve BRCA2 gibi genlerde önceki nesillerden aktarılan ve bu hastalık riskinin artmasıyla bağlantılı olan genetik mutasyonlara dayandırılmaktadır (Majeed ve diğ. 2014).

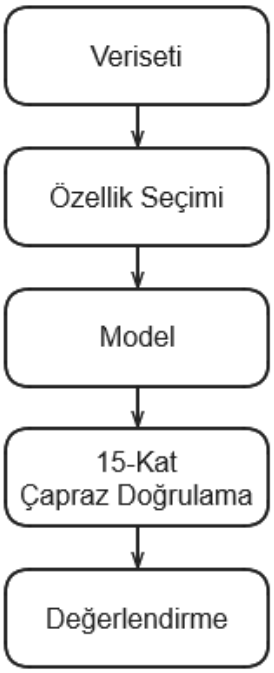
Östrojen ve progesteron hormonlarına maruz kalmak meme kanseri gelişme riskini artırabilir (Fournier ve diğ. 2005). Erken yaşta adet görmeye başlayan veya yaşamının ilerleyen dönemlerinde menopoza giren kadınların meme kanserine yakalanma riski daha yüksektir (Collaborative Group on Hormonal Factors in Breast Cancer 2012).

Yaşam tarzındaki değişiklikler kansere yakalanma riskini artırabilir. Alkol tüketimi kandaki östrojenle ilişkili hormonların seviyesini yükseltebilir (Sun ve diğ. 2017). Menopoz sonrası meme kanseri riski yüksek kilo nedeniyle artabilir (Arnold ve diğ. 2015). Sigara ile ilişkili meme kanseri riskinde artış sadece eski veya mevcut aktif sigara içicilerinde değil aynı zamanda pasif içicilerde de görülmektedir (Dossus ve diğ. 2014).

3. PROBLEM TANIMI

Bilgisayar destekli tıbbi uygulamalar son dönemlerde oldukça popüler hale gelmiştir. Çünkü bilgisayar destekli uygulamalar veriyi yorumlamada ve bunlar arasındaki ilişkiyi tespit etmede son derece etkilidir. Makine öğrenmesi kullanarak bir tahmin elde etmek için veri temizleme, özellik mühendisliği, model eğitme, son işleme işlemlerine gerek vardır (Vieira ve diğ. 2020). Her bir işlem için farklı algoritmalar kullanılabilir ve bütün algoritmalar için hiperparametre optimizasyonu yapılabilir. Aralarından en iyi kombinasyon seçilip kullanıma hazır hale gelir. AutoML, makine öğrenmesi ardışık düzenini özellik mühendisliği, algoritma/model seçimi ve hiperparametre ayarlamasıyla oluşturulan bir uzayda otomatik olarak aradığı için akademik alanda ve endüstri alanında önemli bir ilgi görmüştür (He ve diğ. 2021). Yapılan çalışmaların çoğunda makine öğrenmesi sınıflandırma algoritmalarının varsayılan ayarlar ile hiperparametre optimizasyonu ya da ön işleme yapılmadan kullanıldığı görülmüştür. AutoML kullanımı ile hiperparametre optimizasyonu, özellik mühendisliği, veri temizleme, veri ön işleme gibi ara işlemlerin otomatik yapılması yani makine öğrenmesinin aşamalı otomasyonuyla elde edilen sonuçlar klasik yöntemlerin sonuçlarıyla karşılaştırılmıştır. Eğer hiperparametre optimizasyonu yapılıp en uygun kombinasyon bulunursa daha iyi performans elde edilmesi beklenir.

Obaid ve diğerleri çalışmalarında verisetini Destek Vektör Makineleri için üç tip çekirdek fonksiyonu (doğrusal, kuadratik ve kübik), Karar Ağaçları için üç tip ağaç (karmaşık, orta ve basit) ve K-En Yakın Komşu için üç tip sınıflandırıcı (ince, orta ve kaba) kullanarak eğitmişlerdir, bu üç modelin performansını test etmek için 15 kat çapraz doğrulama kullanmışlardır ve modellerin sınıflandırma doğruluğunu karşılaştırmışlardır. En iyi performansı Destek Vektör Makineleri, K-En Yakın Komşu, Karar Ağaçları arasından ikinci dereceden destek vektör makineleri ile %98.1 doğruluk elde etmişlerdir ve kullandıkları yöntem Şekil 3.1'de bulunmaktadır (Obaid ve diğ. 2018).

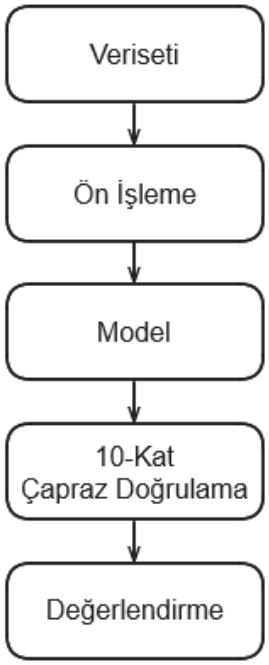


Şekil 3.1: Obaid ve diğ. önerilen yöntem (Obaid ve diğ. 2018).

Agarap çalışmasında verisetine standardizasyon uygulayarak en iyi performansı Doğrusal Regresyon, Çok Katmanlı Algılayıcı (her biri 500 düğümden oluşan üç gizli katman, aktivasyon fonksiyonu ReLU), En Yakın Komşu, Softmax Regresyon, Destek Vektör Makineleri, Kapılı Tekrarlayan Birim-Destek Vektör Makineleri arasından Çok Katmanlı Algılayıcı ile %99.04 doğruluk elde etmiştir (Agarap 2018). Bayrak ve diğerleri çalışmalarında WEKA makine öğrenmesi aracı (ARFF, SMO algoritması, LibSVM, Çok Katmanlı Algılayıcı, Oylanan Algılayıcı) kullanmışlardır. En iyi performansı Destek Vektör Makineleri ve Yapay Sinir Ağı arasından Destek Vektör Makineleri ile %96,9957 doğruluk elde etmişlerdir (Bayrak ve diğ. 2019). Darapureddy ve diğerleri çalışmalarında hiperparametre optimizasyonu ile aşırı öğrenmenin engellenebileceğini belirtmişlerdir (Darapureddy ve diğ. 2019). Mushtaq ve diğerleri çalışmalarının sonuçları Ki-Kare tabanlı özellik seçimini içeren tekniğin Canberra veya Manhattan mesafe fonksiyonları ile en yüksek doğruluk elde ettiğini ve K-En Yakın Komşu için uygun k değerinin, mesafe fonksiyonunun, Ki-Kare tabanlı özellik seçimi ile en yüksek doğruluğu verdiğini gösterdi (Mushtaq ve diğ. 2019). Bharat ve diğerleri çalışmalarında verisetini %80 eğitim %20 test için bölüp K-En Yakın Komşu, Destek Vektör Makineleri, Naive Bayes, Karar Ağacı algoritmalarını kullanmışlardır. Diğerlerine göre Destek Vektör

Makineleri kötü performans göstermiştir ancak giriş standardizasyonu ile performansın artacağını belirtmişlerdir. K-En Yakın Komşu, Destek Vektör Makineleri, Naive Bayes, Karar Ağacı arasından genel yöntem için K-En Yakın Komşu en iyi sonucu vermiştir (Bharat ve diğ. 2018). Shamrat ve diğerleri çalışmalarında ön işleme olarak verisetinden ID sütununu ve sayısal olmayan değerleri silmişlerdir. Naive Bayes, Rassal Orman, Destek Vektör Makinesi, Karar Ağacı, K-En Yakın Komşu, Lojistik Regresyon arasından doğruluk metriği en yüksek olan %97.07 Destek Vektör Makinesi olmuştur (Shamrat ve diğ. 2020). Naji ve diğerleri çalışmalarında verisetini %75 eğitim %25 test için bölmüşlerdir. En iyi performansı Destek Vektör Makineleri, Rassal Orman, Lojistik Regresyon, Karar Ağaçları, K-En Yakın Komşu arasından en yüksek doğruluk değerini %97.2 Destek Vektör Makineleri ile elde etmişlerdir (Naji ve diğ. 2021). Gupta ve Sharma çalışmalarında verisetini %80 eğitim %20 test için bölüp 11 makine öğrenmesi algoritmasına (Lojistik Regresyon, Destek Vektör Makinesi, Ekstra Ağaç, Ada Boost, LGBM, K-En Yakın Komşu , Ridge, Rassal Orman, Naive Bayes, Gradient Boosting sınıflandırıcı ve Karar Ağacı Sınıflandırıcı) ek olarak aykırı değer kaldırma ve normalleştirme teknikleri kullanmıştır. En iyi performansı Lojistik Regresyon ile %97,89 doğruluk elde etmişlerdir (Gupta ve Sharma 2022). Mangukiya çalışmasında verisetine standardizasyon uygulayarak ve uygulamadan karşılaştırma yapmıştır. En iyi performansı XGboost ile %98.24 doğruluk ile elde etmiştir ve standardizasyon XGboost için aynı doğruluk değeri vermiştir sonucu değiştirmemiştir. Standardizasyon diğer modellerin bazılarında doğruluk oranını artırırken bazılarında düşürmüştür (Mangukiya 2022). Nasser ve Behadili çalışmalarında verisetine ön işleme olarak kayıp verilerle başa çıkmak için en sık rastlanan değer ile doldurma yöntemini kullanmışlardır. Verisetini %80 eğitim %20 test için ayırıp Karar Ağacı ve K-En Yakın Komşu algoritmalarını kullanmışlardır. Karar Ağacı'nın Gini endeksi ile budanmasının daha iyi sonuçlar verdiği ve %87.30 doğruluk elde edildiği belirtilmiştir. K-En Yakın Komşu için %86.24 doğruluk elde edilmiştir ve eğitim ve test arasında uzaklığı ölçmek için Öklidyen yaklaşım diğer yaklaşımlara göre daha iyi sonuç verdiğini belirtmişlerdir (Nasser ve Behadili 2022). Divyavani ve Kalpana çalışmalarında verisetine Destek Vektör Makineleri kullanırken ön işleme olarak kayıp verilerle başa çıkmak için ortalama değer ile doldurma yöntemini kullanmışlardır. Verisetini karıştırıp %70 eğitim %30 test için ayırıp Destek Vektör Makineleri ve Yapay Sinir Ağı kullanmışlardır. 10 kat çapraz doğrulama ile Destek

Vektör Makineleri %98, Yapay Sinir Ağı %99 doğruluk elde etmiştir (Divyavani ve Kalpana 2020). Iliyas ve diğerleri çalışmalarında verisetine ön işleme olarak ID sütununu silip kayıp veriler için makine öğrenmesi tabanlı impütasyon yöntemi olan MissForest kullandıklarını belirtmişlerdir. Destek Vektör Makineleri, K-En Yakın Komşu (Öklidyen yaklaşım), Naive Bayes, Lojistik Regresyon, Everişimli Sinir Ağları kullanıp en iyi doğruluğu %98 ile CNN ile elde ettiklerini belirtmişlerdir (Iliyas ve diğ. 2022). Sahu ve diğerleri çalışmalarında meme kanseri sınıflandırması ve tespiti için hibrit özellik seçimi yöntemi ile Yapay Sinir Ağı kullanarak diğer bazı hibrit yöntemlerle ve Rassal Orman, Yapay Sinir Ağı, Naive Bayes ile karşılaştırma yapmışlardır. Önerdikleri yöntem ile verisetini %80 eğitim %20 test bölerek %97 doğruluk elde ettiklerini belirtmişlerdir ve kullandıkları yöntem Şekil 3.2'de bulunmaktadır (Sahu ve diğ. 2018).

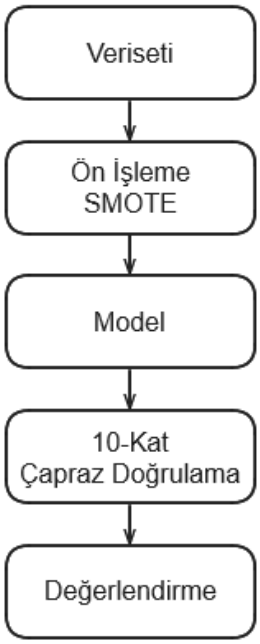


Şekil 3.2: Sahu ve diğ. önerilen yöntem (Sahu ve diğ. 2018).

Kwon ve diğerleri çalışmalarında meme kanseri sınıflandırması için verisetine normalizasyon uygulayıp verisetini %80 eğitim %20 test için bölmüşlerdir ve 5 kat çapraz doğrulama uygulamışlardır. H2O platformunda varsayılan parametre değerleri ile Gradyan Artırma, Dağıtık Rassal Orman, Genelleştirilmiş Doğrusal Model ve Derin Sinir Ağı ile yığılmış topluluk kullanmışlardır. Temel öğreniciyi oluştururken bu modelleri kullanıp her birini tekrar meta öğrenici olarak

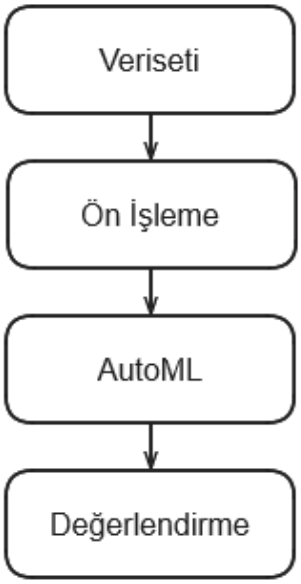
kullanmışlardır. Yığılmış topluluktaki her bir meta öğrenici modelin performansını karşılaştırıp meta öğrenici olarak Gradyan Artırma modeli ile en yüksek doğruluğu (tekil %96.32 topluluk %96.84), meta öğrenici olarak Genelleştirilmiş Doğrusal Model ile düşük kök ortalama kare hatası elde ettiklerini belirtmişlerdir (Kwon ve diğ. 2019). Kumar ve diğerleri çalışmalarında meme kanseri tahmini için on iki sınıflandırıcı algoritma kullanmışlardır. Yığın kalınlığı özelliğini değerlendirme sınıfı olarak aldıklarını ve deneyler sırasında on özellik kullandıklarını belirtmişlerdir. Sonuç olarak ağaç ve tembel sınıflandırıcı algoritmalarının diğerlerine göre daha başarılı olarak %99'a yakın doğruluk elde ettiğini belirtmişlerdir (Kumar ve diğ. 2020). Ghiasi ve Zendehboudi çalışmalarında meme kanseri sınıflandırması için Rassal Orman ve Ekstra Ağaç kullanmışlardır. Verisetinden eksik verileri sildiklerini ve verisetini %85 eğitim %15 test ayırarak 10-kat çapraz doğrulama kullandıklarını belirtmişlerdir. Geliştirdikleri dört ile on sınıflandırma ve regresyon ağacına sahip Rassal Orman modellerinin ve üç ile dokuz sınıflandırma ve regresyon ağacına sahip Ekstra Ağaç modellerinin tüm durumlarda verisetini %100 doğruluk, duyarlılık ve özgüllükle tahmin edebildiğini, kullandıkları Rassal Orman ve Ekstra Ağaç modellerinin kullanımı kolay sonuçlar ve yüksek teşhis performansı sunduğunu belirtmişlerdir (Ghiasi ve Zendehboudi 2021). Dora ve diğerleri çalışmalarında meme kanseri sınıflandırması için Gauss-Newton temsil tabanlı yeni bir algoritma önermişlerdir. Çalışmada verisetini dört farklı şekilde 50-50, 60-40, 70-30, çapraz doğrulama ile eğitim ve test için ayırıp değerlendirmişlerdir. Önerdikleri yöntemin sınıflandırmada eğitim örnekleri için en uygun ağırlıkları belirlemek üzere Gauss-Newton'a dayalı yeni bir yaklaşım sunduğunu ve çeşitli değerlendirme metriklerine göre klasik yöntemlere kıyasla daha başarılı olduğunu belirtmişlerdir (Dora ve diğ. 2017). Kumari ve Singh çalışmalarında meme kanseri tahmini için Lojistik Regresyon, Destek Vektör Makinesi, K-En Yakın Komşu sınıflandırıcıları kullanmışlardır. Verisetine özellik seçimi uygulayıp 10-kat çapraz doğrulama kullanarak en yüksek K-En Yakın Komşu ile %99.28 doğruluk elde ettiklerini belirtmişlerdir (Kumari ve Singh 2018). Jin ve diğerleri çalışmalarında meme kanseri sınıflandırması için çapraz doğrulama için optimal kat ve modifiye edilmiş tek çıkışlı Chebyshev-polinom sinir ağının başlangıç yapısını elde etmek için alt küme yöntemi önermişlerdir. Önerdikleri yöntemin hesaplama karmaşıklığının çok katmanlı algılayıcıdan daha düşük olduğunu ve diğer bazı makine öğrenmesi modellerine kıyasla %100 doğruluk elde ettiklerini belirtmişlerdir (Jin ve diğ. 2019).

Hambali ve diğerleri çalışmalarında meme kanseri sınıflandırması için beş farklı algoritma ve bunların topluluk algoritmasını kullanmışlardır. Veri dengesizliği etkisini azaltmak için sınıflandırma algoritmaları uygulamadan önce verisetine SMOTE algoritmasıyla ön işleme yaptıklarını ve performans değerlendirmesi için 10-kat çapraz doğrulama kullandıklarını belirtmişlerdir. Sonuç olarak ADABOOST-Rassal Orman %82.52 doğruluk ile diğer sınıflandırma algoritmalarına göre başarılı olduğunu belirtmişlerdir ve kullandıkları yöntem Şekil 3.3'te bulunmaktadır (Hambali ve diğ. 2019).



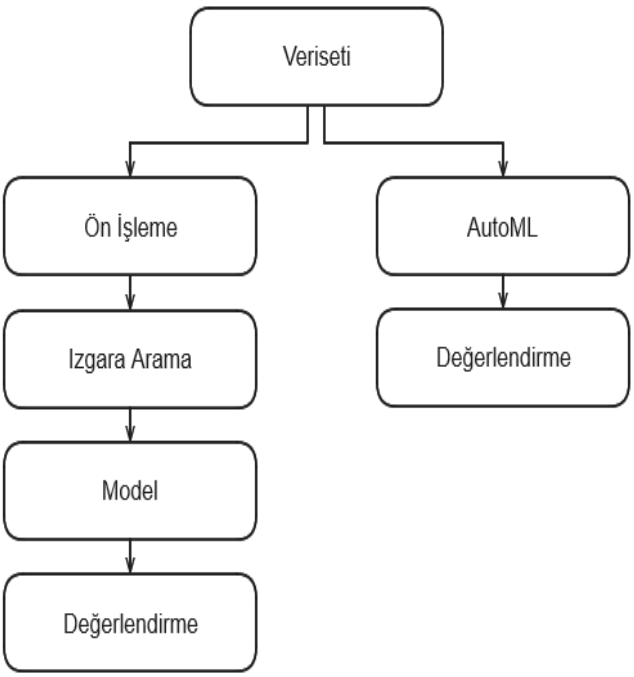
Şekil 3.3: Hambali ve diğ. önerilen yöntem (Hambali ve diğ. 2019).

Rashed ve diğerleri çalışmalarında meme kanseri tespiti için yedi AutoML yöntemini sekiz farklı verisetine uygulamışlardır. Verisetine göre ön işleme uyguladıklarını, verisetini %80 eğitim %20 test için ayırdıklarını ve kullandıkları bazı yöntemlerin çapraz doğrulama desteklemediğini belirtmişlerdir. Lazy Predict ve MATLAB Sınıflandırma Öğrenicisi'nin ikili sınıflandırma için diğer makine öğrenmesi yöntemlerinden verisetlerinin çoğunda daha başarılı olduğunu belirtmişlerdir ve kullandıkları yöntem Şekil 3.4'te bulunmaktadır (Rashed ve diğ. 2023).



Şekil 3.4: Rashed ve diğ., Yunus ve diğ., Madni ve diğ. önerilen yöntem (Rashed ve diğ. 2023, Yunus ve diğ. 2022, Madni ve diğ. 2023).

A. J. B ve Palaniswamy çalışmalarında meme kanseri tespiti için AutoML yöntemi olarak Auto Sklearn, klasik yöntem olarak Ekstra Ağaç sınıflandırıcı ve TPOT kullanmışlardır. Verisetini %75 eğitim %25 test bölerek Ekstra Ağaç sınıflandırıcı için parametreleri otomatik ardışık düzenden elde ettikleri değerler ile kullandıklarını, özellik azaltma, normalizasyon, veri noktalarının dönüştürülmesi uygulandığını, k-kat çapraz doğrulama kullanıldığını belirtmişlerdir. AutoML yöntemi ile %97.5, Ekstra Ağaç ile %99.7, TPOT ile %98.6 doğruluk elde etmişlerdir (A. J. B ve Palaniswamy 2021). Radzi ve diğerleri çalışmalarında TPOT tabanlı model seçimi ile Naive Bayes, Destek Vektör Makinesi ve Yapay Sinir Ağı için ızgara arama parametre ayarı arasında karşılaştırma yapmışlardır. Mamografi görüntülerinden elde ettikleri radyomik özellikleri girdi olarak kullanmışlardır. Verisetini %80 eğitim %20 test için ayırıp 5-kat çapraz doğrulama kullanmışlardır. TPOT için jenerasyon sayısı 100, popülasyon büyüklüğü 50 seçtiklerini belirtmişlerdir. Varsayılan ayarlarda TPOT modelinin, kontrollü TPOT yapılandırma tabanlı modelden ve ızgara arama optimizasyonlu modelden daha performanslı sınıflandırma ardışık düzeni ürettiğini belirtmişlerdir ve kullandıkları yöntem Şekil 3.5'te bulunmaktadır (Radzi ve diğ. 2021).



Şekil 3.5: Radzi ve diğ. önerilen yöntem (Radzi ve diğ. 2021).

Aghalarova ve Bozkurt Keser çalışmalarında öğrencilerin akademik performanslarını tahmin etmek için H2O AutoML yöntemi kullandıklarını ve ön işleme için kullanıcı müdahalesi olmadığını belirtmişlerdir. AutoML sonucunda elde ettikleri Dağıtık Rassal Orman modelinin kullandıkları AutoML sürümünde hiperparametre optimizasyonunun olmadığını belirterek ayrıca ızgara arama uygulamışlardır ve modelin varsayılan değerleri ile %77.5 doğruluk hiperparametre optimizasyonu yapılmış hali ile %82.3 doğruluk elde ettiklerini belirtmişlerdir (Aghalarova ve Bozkurt Keser 2021). Ahmed ve diğerleri kardiyovasküler ve kronik solunum yolu hastalıklarının çok değişkenli zaman serisi sensör yaşamsal belirti tahmini için elde ettikleri ham veriyi tek dosya haline getirip, kayıp verileri temizleyip iki hastalık için iki dosya oluşturduklarını ve normalizasyon uyguladıklarını belirtmişlerdir. Kullandıkları hibrit modelde zaman serisi tahmin modeline ek olarak sınıflandırıcı için TPOT kullanmışlardır. TPOT'un uygunluk fonksiyonunu iyileştirdiği ve en etkili topluluk öğrenme stratejisini belirlediğini belirtmişlerdir (Ahmed ve diğ. 2023). Pais ve diğerleri meme, akciğer ve böbrek tümörü transkriptomik verileri toplayıp aralarından 58 gen seçmişlerdir. Veri oluştururken yalnızca tanıdan sonra 5 yıldan fazla hayatta kaldığı bildirilen veya ilk 2 yıl içinde öldüğü bildirilen hastalarla ilişkili transkriptomları seçmişlerdir. Model

seçimi için TPOT kullandıklarını belirtmişlerdir. Verisetini %50 eğitim %50 test bölerek 100 jenerasyon sayısı 100, popülasyon büyüklüğü 50, optimize edilecek metrik için AUC seçtiklerini belirtmişlerdir. Meme, akciğer, böbrek için sırasıyla %84 %52 %71 doğruluk, %53 %48 %70 AUC elde etmişlerdir (Pais ve diğ. 2023). Yunus ve diğerleri radyomik özelliklere dayalı aterosklerotik plakların sınıflandırılması için TPOT kullanmışlardır. 202 görüntüden 606 ilgi hacmi elde edip 4 sınıfa ayırmışlardır ve farklı türde girdi değişkenleri kullanılarak 4 çeşit sınıflandırma modeli oluşturmuşlardır. Verisetini %80 eğitim %20 test için bölerek varsayılan ayarlarda TPOT kullandıklarını ve TPOT'un girdi değişkeni olarak farklı radyomik özelliklerden oluşan verisetine sahip her model için farklı sınıflandırıcılar önerdiğini belirtmişlerdir ve kullandıkları yöntem Şekil 3.4'te bulunmaktadır (Yunus ve diğ. 2022). Orlenko ve diğerleri TPOT ile elde edilen model ile seçilen makine öğrenmesi sınıflandırıcılarına (Lojistik Regresyon, Karar Ağacı, Rassal Orman) hiperparametre optimizasyonu yaparak karşılaştırma yapmışlardır. Koroner arter hastalığının anjiyografik tanılarını tahmin etmek için TPOT kullanmışlardır. Verisetini %75 eğitim %25 test için ayırıp 10-kat çapraz doğrulama kullanmışlardır. TPOT tarafından oluşturulan makine öğrenmesi ardışık düzenleri birden fazla performans ölçütünde ızgara aramasıyla optimize edilmiş modellerden daha iyi performans gösterdiğini belirtmişlerdir (Orlenko ve diğ. 2019). Angarita-Zapata ve diğerleri kaza şiddet tahmini için AutoML yönteminin klasik yöntemlere karşı ne ölçüde rekabetçi olabileceğini incelemişlerdir. AutoML yöntemi olarak 15, 60, 150 dakika çalıştırma süreleriyle Auto Sklearn (Feurer ve diğ. 2015), TPOT, AutoGluon (Erickson ve diğ. 2020), taban çizgisi yöntemi olarak Gradyan Artırma, Gaussian Naive Bayes, K-En Yakın Komşu, Çok Katmanlı Algılayıcı, Rassal Orman kullanmışlardır. Hem AutoML hem de klasik yöntemlerin hiperparametre değerlerinde optimizasyon veya ayarlama yapmadıklarını, 10-kat çapraz doğrulama kullandıklarını ve optimize edilecek metrik için ROC_AUC kullandıklarını ek olarak iki istatiksel test kullandıklarını belirtmişlerdir. Sonuç olarak kullandıkları AutoML yöntemleri arasından 150 dakika ile Auto Sklearn, taban çizgisi yöntemleri arasından Gradyan Artırma yönteminin daha başarılı olduğunu belirtmişlerdir (Angarita-Zapata ve diğ. 2021). Agrapetidou ve diğerleri banka iflas tahmini için AutoML yöntemi kullanmışlardır. Verisetini 1100 eğitim 343 test için ayırdıklarını ve AutoML yönteminin tekrarlı katmanlı k-Kat çapraz doğrulama kullanarak eğitim setinde %94.6 test setinde %97.4 doğruluk elde ettiklerini belirtmişlerdir (Agrapetidou ve

diğ. 2020). Madni ve diğerleri çalışmalarında su kalite tahmini için H2O AutoML ve altı farklı makine öğrenmesi algoritması kullanmışlardır. Birinci aşamada verisetinden eksik verileri silerek, ikinci aşamada eksik verileri KNN doldurma yöntemi ile 10-kat çapraz doğrulama kullanarak model karşılaştırması yapmışlardır. Sonuç olarak sunulan yığılmış H2O AutoML çerçevesinin KNN doldurma yöntemi ile veri silme yöntemine göre daha başarılı olduğunu ve %97 doğruluk elde ettiklerini belirtmişlerdir ve kullandıkları yöntem Şekil 3.4'te bulunmaktadır (Madni ve diğ. 2023). Putrada ve diğerleri çalışmalarında kredi kartı başvuru onay sınıflandırması için AutoML aracı TPOT kullanıp önceki çalışmalarda kullanılan üç algoritma Naive Bayes, Destek Vektör Makinesi, Karar Ağacı kullanarak karşılaştırmışlardır ve TPOT'un %89 ile daha yüksek doğruluk, %94 ile daha yüksek AUC elde ettiğini belirtmişlerdir (Putrada ve diğ. 2022). Kulkarni ve diğerleri çalışmalarında diyabetik tahmin için AutoML araçlarından TPOT ve H2O kullanıp manuel olarak oluşturdukları Adaboost modelini farklı ön işleme yöntemleriyle kullanarak karşılaştırmışlardır. Aykırı değer kaldırma ve ölçekleme uyguladıkları birinci model %77 doğruluk, ikinci modelde ek olarak veri dengeleme ile %99 doğruluk ve üçüncü modelde hiperparametre optimizasyonu ile %100 doğruluk elde ederken TPOT ve H2O ile sırasıyla %81 ve %82 doğruluk elde ettiklerini belirtmişlerdir (Kulkarni ve diğ. 2021). Mohsen ve diğerleri çalışmalarında tip 2 diyabet sınıflandırması için AutoML aracı H2O kullanıp manuel olarak oluşturdukları Rassal Orman, Destek Vektör Makinesi, Adaboost, Topluluk modelleriyle karşılaştırmışlardır. Verisetinin dengesiz olması nedeniyle çoğunluk sınıfı azınlık sınıfın sayısına indirgediklerini, değerlendirme için ise 5-kat çapraz doğrulama kullandıklarını belirtmişlerdir. AutoML modeli ile sırasıyla %90, %95,9, %83, %89 ile üstün doğruluk, hassasiyet, duyarlılık ve F1-skor elde etmişlerdir (Mohsen ve diğ. 2022).

4. MATERYAL VE YÖNTEM

4.1 Yapay Zeka

"Yapay zeka" terimini 1955 yılında ortaya atan John McCarthy, bu terimi "akıllı makineler yapma bilimi ve mühendisliği" olarak tanımlamıştır (McCarthy 2007). McCarthy ve diğ. (1955) yapay zeka problemini "bir makinenin davranışını eğer bir insan o şekilde davransaydı zeki olarak adlandırılabilecek şekilde sağlamak" olarak tanımlanmaktadır.

Yapay zeka, makinelerin genellikle insan benzeri zeka veya deneyimlerden öğrenme yeteneği gerektiren şeyleri nasıl yapabildiğinin incelenmesidir. Yapay zeka tanımı kişinin zekayı nasıl tanımladığına ve hangi görevlerin zeki olarak kabul edildiğine bağlı olabilir. Bazı insanlar dört işlem yapan bir bilgisayarı bir yapay zeka örneği olarak görebilirken bazıları insan zekasını taklit eden daha karmaşık görevleri yerine getirebilen bir sistemi gerçek yapay zeka olarak değerlendirebilir (Artasanchez ve Joshi 2020). Yapay zeka, makinelerin içinde bulunduğu ortamı kavramasını ve durumlara insanlara benzer bir şekilde yanıt vermesini sağlayan yöntemlerin incelenmesini içerir.

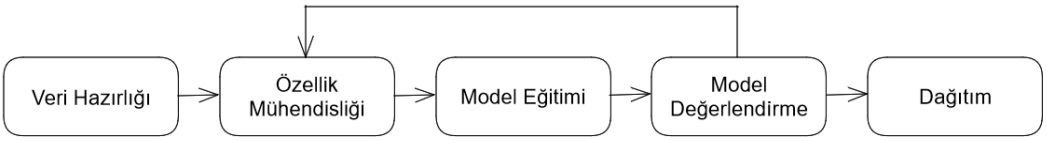
Yapay zeka, insan düşünce süreçlerini ve davranışlarını taklit ederek ya da rasyonalitenin soyut, biçimsel tanımlarını takip ederek akıllı davranışlar sergileyebilen makineler veya sistemler oluşturmayı amaçlayan bir çalışma ve geliştirme alanıdır. İnsan davranış ve düşünce süreçlerinin deneysel gözlemi, matematiksel ve mühendislik teknikleri ile istatistik, kontrol teorisi ve ekonomi dahil olmak üzere bir dizi yaklaşım ve yöntemi kapsar (Russel ve Norvig 2020).

4.2 Makine Öğrenmesi

Yapay zekanın bir alt dalı olan makine öğrenmesi, deneyim yoluyla performansı artırma yöntemlerine odaklanır. Bazı yapay zeka sistemleri yetkinliğe

ulaşmak için makine öğrenmesi yöntemlerinden yararlanırken, bazıları bu yöntemleri kullanmazlar (Russel ve Norvig 2020). Makine öğrenimi algoritmaları birçok algısal görev için girdi verileri ile çıktı verileri arasındaki ilişkilerin nasıl bulunacağını öğrenme kapasitesine sahiptir (Dietterich 1996). Makine öğrenmesinin arkasındaki temel fikir verilere dayalı olarak doğru ve tutarlı tahminler yapabilen veya kararlar verebilen tahmine dayalı modeller oluşturmaktır. Bu modeller görüntü tanıma, doğal dil işleme, dolandırıcılık tespiti ve tavsiye sistemleri gibi çok çeşitli uygulamalarda kullanılabilir. Son yıllarda büyük verinin kullanılabilirliği ve işlem gücündeki gelişmeler ile birlikte popülerlik kazanmıştır.

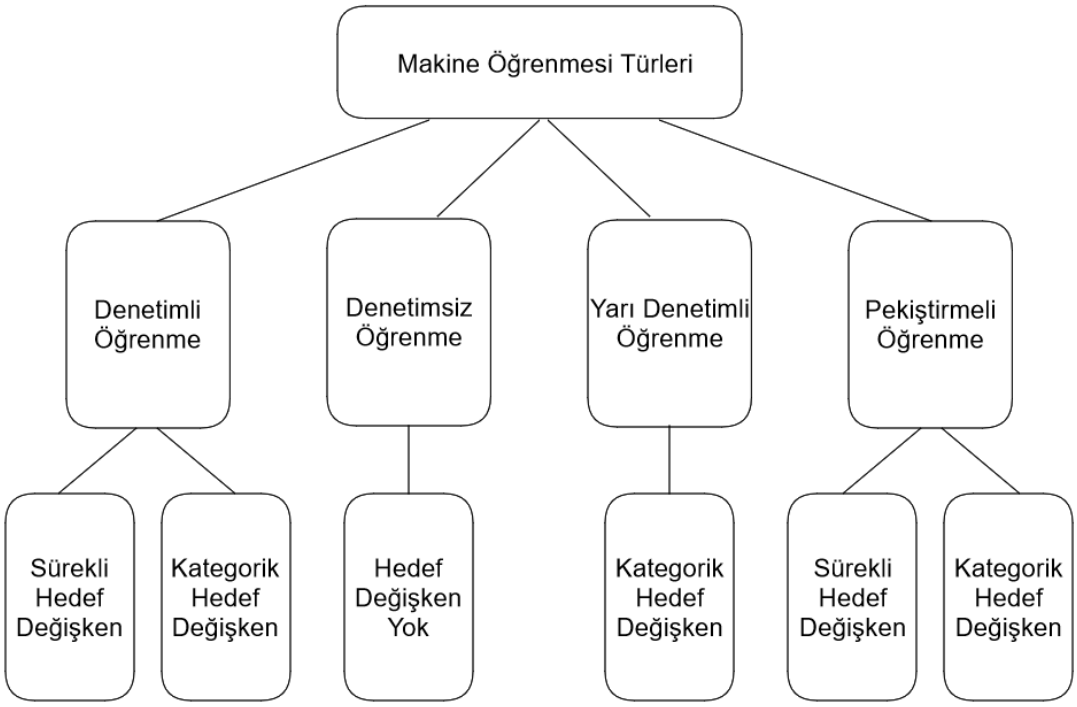
Makine öğrenmesi kurumsal iş akışlarının ve günlük görevlerin ayrılmaz bir parçası haline gelmiştir. Bu entegrasyon daha verimli ve doğru süreçlere duyulan ihtiyaçtan kaynaklanmaktadır. Makine öğrenmesi sağlık, finans, perakende ve ulaşım dahil olmak üzere çok çeşitli alanların iş akışlarını iyileştirilebilecek potansiyele sahip olduğundan önem kazanmıştır ve günümüzde giderek daha önemli hale gelmiştir. Ayrıca makine öğrenmesi otonom araçlar ve robotik gibi en son teknolojilerin geliştirilmesinde önemli bir rol oynamaktadır. Büyük verisetlerini analiz etme ve bu analize dayalı tahminlerde bulunma yeteneği işletmelerin ve kuruluşların çalışma biçiminde pozitif etki yaratmıştır. Şekil 4.1'de makine öğrenmesi iş akışı bulunmaktadır.



Şekil 4.1: Makine Öğrenmesi İş Akışı (Gad 2023).

Makine öğrenmesi sistemleri eğitim sırasında nasıl denetlendiklerine, ilerledikçe yeni şeyler öğrenip öğrenmediklerine ve tahminlerde bulunmak için verileri nasıl analiz ettiklerine göre farklı kategorilere ayrılabilir (Géron 2022). Makine öğrenmesi algoritmaları dört temel kategoride incelenebilir. Denetimli öğrenme, denetimsiz öğrenme, yarı denetimli öğrenme ve pekiştirmeli öğrenme. Her

tekniğin kendine özgü güçlü ve zayıf yönleri vardır ve belirli uygulamalar için kullanılır. Şekil 4.2'de makine öğrenmesi türleri bulunmaktadır.

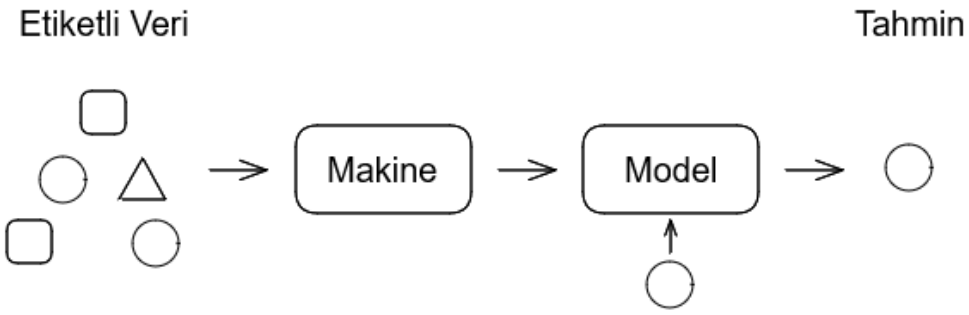


Şekil 4.2: Makine Öğrenmesi Türleri (Gavrilova 2023).

4.2.1 Denetimli Öğrenme

Bu tür makine öğrenmesi bir hedefi doğrulamak ve öğrenmek için girdi olarak etiketli bir veriseti kullanır. Amaç önceden etiketlenmiş verilere dayanarak tahminlerde bulunmak veya yeni verileri sınıflandırmaktır. Denetimli öğrenmede ajan girdiyi çıktıya eşleyen bir fonksiyon öğrenir ve daha sonra bu fonksiyon yeni girdiler üzerinde tahminler yapmak için kullanır (Russel ve Norvig 2020). Bu tür makine öğrenmesi girdi verilerinin farklı sınıflara kategorize edilebildiği durumlarda kullanılır. Denetimli öğrenme algoritmaları regresyon ve sınıflandırma problemleri için kullanılabilir. Tahmin edilecek hedef değeri nümerik ise problem regresyon, kategorik ise sınıflandırmadır. Regresyon problemlerinde algoritma sürekli bir çıktı tahmin ederken sınıflandırma problemlerinde algoritma kategorik bir çıktı tahmin eder. Yatak odası sayısı, banyo, metrekare, konum gibi özelliklerine göre bir evin fiyatını tahmin etme regresyon problemiyken bir e-postanın spam olup olmadığını

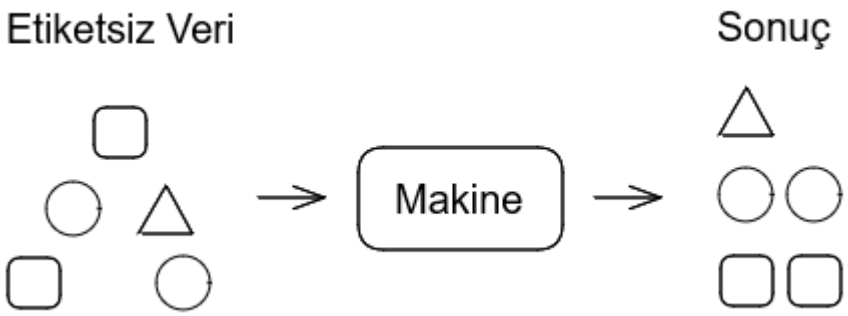
belirleme sınıflandırma problemidir. Şekil 4.3'te denetimli öğrenme örneği bulunmaktadır.



Şekil 4.3: Denetimli Ögrenme (Kozan 2021).

4.2.2 Denetimsiz Öğrenme

Bu tür makine öğrenmesi etiketli veri içermez. Ajan kategoriler veya etiketler hakkında önceden bilgi sahibi olmadan girdi verilerindeki kalıpları veya ilişkileri bulmak için eğitilir. Denetimsiz öğrenme girdi verilerinin yapısının bilinmediği veya kategorize edilemeyecek kadar karmaşık olduğu senaryolarda kullanılır. Denetimsiz öğrenme algoritmaları kümeleme ve boyut azaltma gibi görevler için kullanılabilir. Kümeleme algoritmaları veri noktalarının benzerliğine dayalı olarak giriş verilerini gruplara ayırmak için kullanılır. Kümeleme problemine bir örnek olarak bir perakende şirketi için müşteri segmentasyonu verilebilir. Amaç müşterileri satın alma davranışlarına, demografik özelliklerine ve diğer ilgili faktörlere göre gruplandırmaktır. Bu, şirketin kişiselleştirilmiş pazarlama kampanyaları ve promosyonlarla belirli grupları hedeflemesine yardımcı olabilir. Kümeleme algoritması verilerdeki örüntüleri belirleyerek müşterileri davranışlarındaki benzerlik ve karakteristiklere göre kümeler halinde gruplandıracaktır. Şekil 4.4'te denetimsiz öğrenme örneği bulunmaktadır.



Şekil 4.4: Denetimsiz Öğrenme (Kozan 2021).

4.2.3 Yarı Denetimli Öğrenme

Bu tür makine öğrenmesi denetimli ve denetimsiz öğrenmenin bir kombinasyonudur. Ajan az sayıda etiketli verisetinin yanı sıra çok sayıda etiketsiz veriseti üzerinde eğitilir. Bu yaklaşımın amacı modelin sınıflandırma doğruluğunu artırmak için kolayca elde edilen çok miktarda etiketsiz veriyi kullanmaktır. Yarı denetimli öğrenme etiketli verilerin elde edilmesinin çok maliyetli veya zaman alıcı olduğu durumlarda kullanılabilir (Géron 2022).

4.2.4 Pekiştirmeli Öğrenme

Bu tür makine öğrenmesi eylemlerinin sonucunda geri bildirim alarak bir hedefe ulaşmak için çevre ile etkileşime giren bir ajan veya makine öğrenmesi sistemidir. Geri bildirim pozitif veya negatif olabilir ve amaç negatif geri bildirimi minimuma indirirken pozitif geri bildirimi maksimuma çıkarmaktır. Pekiştirmeli öğrenme algoritmaları oyun oynama, robotik ve karar verme gibi problemler için kullanılabilir.

Pekiştirmeli öğrenme sayısal bir ödül sinyalini en üst düzeye çıkarmak için ne yapılacağını, durumların eylemlerle nasıl eşleştirileceğini öğrenmektir. Öğrenene hangi eylemleri yapması gerektiği bildirilmez bunun yerine hangi eylemlerin en fazla

ödülü getireceğini öğrenenin deneyerek keşfetmesi gerekir. En ilginç ve zorlu durumlarda eylemler yalnızca anlık ödülü değil aynı zamanda bir sonraki durumu ve bunun aracılığıyla sonraki tüm ödülleri de etkileyebilir. Bu iki özellik yani deneme-yanılma araştırması ve gecikmeli ödül pekiştirmeli öğrenmenin en önemli iki ayırt edici özelliğidir (Sutton ve Barto 2018). Deneme-yanılma araştırma sürecinde bir ajan gelecekte daha yüksek ödüller getirebilecek yeni eylemleri keşfetme ile geçmişte ödül getirdiği bilinen etkili eylemleri kullanma arasında denge kurmalıdır. Ajan çeşitli eylemleri dener, uzun vadede en fazla kümülatif ödülü kazandıranları kademeli olarak tercih eder bu değer fonksiyonu olarak tanımlanır. Ajan sadece anlık ödülleri dikkate almaz aynı zamanda zaman içindeki kümülatif ödüllerin genel tablosuna da bakar (Yu ve He 2019).

4.3 Python

Python popüler ve çok yönlü programlama dillerinden biridir. Ücretsiz ve açık kaynak kodludur. Basit ve minimalist bir sözdizimi ile okunması, öğrenilmesi ve yazılması kolay bir dil olarak kabul edilir. Python nesne tabanlı, zorunlu, fonksiyonel ve yordamsal programlama dahil olmak üzere çoklu programlama stillerini destekler. Python yorumlanan bir dildir yani derleme gerekmez ve kod doğrudan çalıştırılabilir. Bu, geliştirme döngüsünü çok hızlı hale getirir. Python yorumlayıcısı programları çalışma zamanında yürütür. Python, kodun yeniden kullanılabilirliğini teşvik eden modülleri ve paketleri destekler. Python'un standart kütüphanesi çeşitli işlemleri yerine getirmek için önceden oluşturulmuş fonksiyonlar ve modüllerden oluşan geniş bir koleksiyon içerir. Hem komut dosyası oluşturma hem de uygulama geliştirme için kullanılan genel amaçlı bir dildir. Web geliştirme, yazılım mühendisliği, eğitim, bilimsel hesaplama, görüntü işleme vb. dahil olmak üzere birçok alanda kullanılmaktadır. Python'da oluşturulan popüler çerçevelerden ve kütüphanelerden bazıları; Django, Flask, NumPy, Pandas, Matplotlib... Python kodu taşınabilir ve platformdan bağımsızdır. Derlenen Python kodu Windows, Unix tabanlı (Linux, macOS vb.) dahil olmak üzere birçok işletim sisteminde çalışabilir. Şekil 4.5'te Python sözdizimine bir örnek bulunmaktadır.

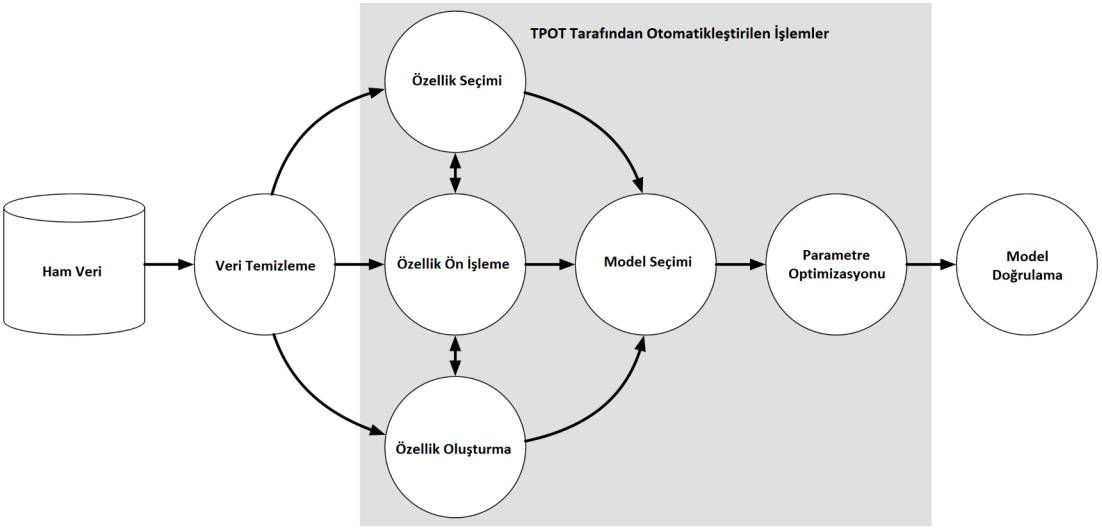
```python
# Variables and data types
name = "Jane"
age = 19
height = 5.8
is_student = True

# Basic arithmetic
sum_result = 10 + 5
difference_result = 10 - 5
product_result = 10 * 5
division_result = 10 / 5

# Conditional statement
if age >= 18:
    print("Adult")
else:
    print("Minor")

# Loop
for i in range(5):
    print("Iteration:", i)

# Lists
fruits = ["apple", "banana", "orange"]
print("Second fruit:", fruits[1])

# Function
def greet(name):
    print("Hello,", name)

greet("John")
```

Şekil 4.5: Örnek Python Sözdizimi.

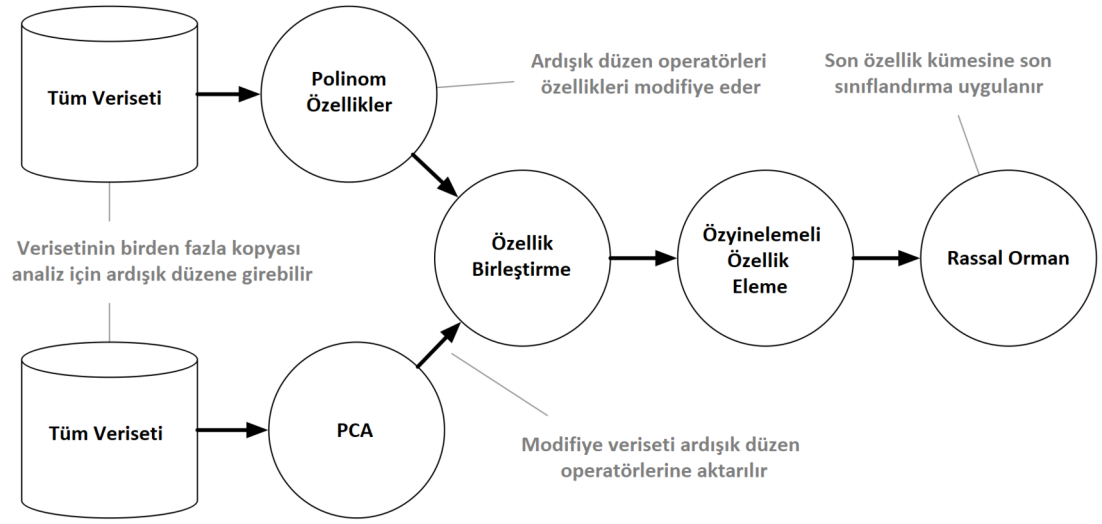
4.4 TPOT

TPOT veya Ağaç Tabanlı Ardışık Düzen Optimizasyon Aracı (Tree-based Pipeline Optimization Tool), Pennsylvania Üniversitesi'ndeki Hesaplamalı Genetik Laboratuvarı'nda (Computational Genetics Lab) geliştirilen açık kaynak kodlu bir AutoML aracıdır. TPOT'un amacı genetik programlama gibi stokastik arama algoritmaları ile ardışık düzenlerin esnek bir ifade ağacı temsilini birleştirerek makine öğrenmesi ardışık düzenlerinin oluşturulmasını otomatikleştirmektir. TPOT,

makine öğrenmesi araç seti olarak Python tabanlı bir makine öğrenmesi kütüphanesi olan scikit-learn (Pedregosa ve diğ. 2011) kullanmaktadır (TPOT 2023). TPOT, belirli bir veriseti ve problem için makine öğrenmesi ardışık düzeni otomatik olarak aramak ve optimize etmek için genetik algoritmalar kullanır (Olson ve Moore 2019). Yüksek performanslı modeller oluşturmak için çok az makine öğrenmesi uzmanlığına sahip kullanıcılar için basit ve anlaşılır bir arayüz sağlar. Sınıflandırma, regresyon, kümeleme, boyutsallık azaltma ve daha fazlası dahil olmak üzere çeşitli makine öğrenmesi görevlerini destekler. Farklı şekillerde birleştirilebilen dönüştürücüler, özellik seçiciler, ölçekleyiciler, ardışık düzen ve tahmin edicilerden oluşan geniş bir kütüphaneye sahiptir. Kullanıcıların yalnızca verisetlerini temin etmeleri ve problem türünü belirtmeleri gerekir ve TPOT, doğruluk, F1 puanı, R2 vb. gibi tercih edilen performans ölçütünü optimize eden tahmin ediciler ve hiperparametrelerin kombinasyonunu bulmak için ardışık düzen uzayını arar. Hız ve verimlilik TPOT'un bazı temel avantajları arasında yer almaktadır. Binlerce ardışık düzen seçeneği arasında arama yapabilir ve hiperparametreleri optimize edebilir. Keşfettiği ardışık düzenler genellikle insan uzmanlar tarafından manuel olarak oluşturulanlardan daha iyi performans gösterir. TPOT ayrıca pratik kurallara veya kişisel tercihlere dayanmak yerine tarafsız, veri odaklı bir yaklaşım sunar. Optimize edilmiş ardışık düzenler tekrarlanabilirdir ve her bir bileşen ve parametre seçiminin arkasındaki mantığın bir açıklaması ile birlikte gelir. TPOT, sağlık, finans, eğitim gibi birçok alanda çok çeşitli gerçek dünya problemlerine uygulanabilir bununla birlikte çeşitli kıyaslama verisetlerinde ve zorluklarda en son teknolojiye sahip sonuçlar elde etmiştir. Araştırma topluluğu tarafından önerilen yeni özellikler ve geliştirmelerle sürekli olarak gelişmektedir. Otomatik makine öğrenmesini daha erişilebilir, pratik ve etkili hale getirmeye yardımcı olur. TPOT, hız ve verimlilik ile yüksek performanslı makine öğrenmesi modelleri sunan güçlü ancak kullanımı kolay bir AutoML aracıdır. Makine öğrenmesi ardışık düzenleri oluşturmak ve model performansını optimize etmek için basit, tarafsız ve veri odaklı bir yaklaşım sağlar. TPOT, makine öğrenmesini yaygınlaştırmaya ve bu alandaki ilerlemeyi hızlandırmaya yardımcı olarak birçok soruna uygulanabilir. Şekil 4.6'da makine öğrenmesi ardışık düzen örneği, Şekil 4.7'de ise TPOT ardışık düzen örneği bulunmaktadır.



Şekil 4.6: Makine Öğrenmesi Ardışık Düzen Örneği (TPOT 2022).



Şekil 4.7: TPOT Ardışık Düzen Örneği (TPOT 2022).

4.5 H2O

H2O açık kaynak kodlu, bellek içi, dağıtık, hızlı ve ölçeklenebilir bir makine öğrenmesi ve tahmine dayalı analitik platformudur. Bu platform büyük ölçekli verisetleri üzerinde makine öğrenmesi modellerinin oluşturulmasını kolaylaştırır ve bu modellerin kurumsal ortamlarda dağıtımını basitleştirir. H2O temel olarak Java dilinde yazılmıştır ve H2O ekosistemindeki tüm düğümler ve makineler arasında

verilere, modellere ve nesnelere sorunsuz erişim sağlayan bir Dağıtık Anahtar/Değer deposu kullanır. H2O performansı artırmak için makine öğrenmesi algoritmalarını uygulamak için dağıtılmış bir Map/Reduce çerçevesi kullanır ve verimli çoklu iş parçacığı için Java Fork/Join çerçevesinden yararlanır. Platform, verileri paralel olarak okur, kümeye dağıtarak ve sıkıştırılmış, sütunlu bir format ile bellekte depolayarak verimli bir şekilde işler. H2O'nun veri ayrıştırıcısı gelen veri kümesinin yapısını akıllıca tahmin eder ve çeşitli kaynaklardan çok sayıda formatta veri alımını kolaylaştırır. Derin öğrenme, ağaç toplulukları ve genelleştirilmiş düşük sıralı modeller gibi çeşitli gelişmiş denetimli ve denetimsiz algoritmaların hızı, kalitesi, kullanım kolaylığı ve model dağıtım yetenekleri, H2O'yu büyük veri veri bilimi için oldukça talep gören bir uygulama programlama arayüzü haline getirmektedir (H2O 2023[a]).

H2O AutoML, belirli bir veriseti için en iyi makine öğrenmesi modellerini oluşturma ve seçme sürecini otomatikleştiren bir makine öğrenmesi çerçevesidir. Algoritmaların seçilmesi, özelliklerin oluşturulması, hiperparametrelerin ince ayarının yapılması, modellerin iyileştirilmesi ve performansın değerlendirilmesi gibi makine öğrenmesi modellerinin oluşturulmasındaki karmaşıklıkları basitleştirir. Tekrarlanan görevleri otomatikleştirerek bireylerin çözmeleri gereken iş sorunlarına ve veriyi anlamaya konsantre olmalarını sağlar. Kullanıcıların makine öğrenmesi algoritmaları ve tekniklerinde derinlemesine uzmanlığa ihtiyaç duymadan hızlı bir şekilde yüksek performanslı tahmin modelleri geliştirmelerini kolaylaştırmak için tasarlanmıştır. Amaçlarından biri verimliliktir; en uygun modelleri hızla eğiterek değerli mesai saatlerinden tasarruf etmeyi amaçlamaktadır. Bu, Python, R veya bir web grafiksel kullanıcı arayüzü aracılığıyla erişilebilen bir arayüz ile gerçekleştirilebilir. Diğer bir amaç yoğun emek gerektiren kodlama görevlerini otomatikleştirerek derin makine öğrenmesi uzmanlığına olan bağımlılığı azaltmaktır. Bu basitleştirme kapsamlı teknik bilgi olmadan makine öğrenmesinin gücünden yararlanmak isteyen profesyoneller için önemli bir avantajdır. Ayrıca tekrarlanabilirlik açısından bilimsel araştırma ve pratik uygulamalar için kriterlerin oluşturulmasında önemli bir rol oynar. Hadoop, Spark ve Kubernetes gibi dağıtık bilgi işlem ortamlarına sorunsuz bir şekilde uyum sağladığından platformun ölçeklenebilirliği de dikkat çekicidir.

Temel özellikleri otomatik veri ön işleme de dahil olmak üzere çok çeşitli işlevleri kapsar. H2O AutoML, imputasyon, One Hot Encode ve standardizasyon gibi işlemleri uygulayarak verilerin modelleme için uygun şekilde hazırlanmasını sağlar. Hiperparametre optimizasyonu da öne çıkan bir diğer özelliktir. Platform, iyi yapılandırılmış hiperparametre uzaylarını kullanarak çeşitli H2O modellerini araştırır ve çapraz doğrulama yoluyla bireysel modellere ince ayar yapar. Model performansını en üst düzeye çıkarmak için topluluk öğrenme tekniklerinden özellikle de yığılmış topluluklardan yararlanır. Birden fazla modelin güçlü yönlerini birleştirerek üstün sonuçlar elde edilebilir. Performans değerlendirmesi açısından oluşturulan tüm modellere erişilebilir ve bir lider tablosunda çeşitli metriklere göre sıralanır. Bu, kullanıcıların kendi özel ihtiyaçları için en uygun modeli seçmelerine yardımcı olur. Ayrıca modeller H2O açıklanabilirlik modülü kullanılarak otomatik olarak açıklanabilir. Böylece makine öğrenmesi modellerinin davranışına ilişkin otomatik içgörüler sağlayarak karmaşık modelleri bile daha anlaşılır hale getirebilir. Oluşturulan modellerin prodüksiyonda kullanılabilmesini sağlamak için dışa aktarılabilir (H2O 2023[b]). H2O eğitim koduna örnek Şekil 4.8'de verilmiştir.

```
import h2o
from h2o.automl import H2OAutoML

# Start the H2O cluster (locally)
h2o.init()

# Import a sample binary outcome train/test set into H2O
train = h2o.import_file("https://s3.amazonaws.com/erin-data/higgs/higgs_train_10k.csv")
test = h2o.import_file("https://s3.amazonaws.com/erin-data/higgs/higgs_test_5k.csv")

# Identify predictors and response
x = train.columns
y = "response"
x.remove(y)

# For binary classification, response should be a factor
train[y] = train[y].asfactor()
test[y] = test[y].asfactor()

# Run AutoML for 20 base models
aml = H2OAutoML(max_models=20, seed=1)
aml.train(x=x, y=y, training_frame=train)

# View the AutoML Leaderboard
lb = aml.leaderboard
lb.head(rows=lb.nrows)  # Print all rows instead of default (10 rows)

# The leader model is stored here
aml.leader
```

Şekil 4.8: H2O Eğitim Örnek Kod (H2O 2023[b]).

4.6 MLJAR

mljar-supervised, tablo verilerini yüksek verimlilikle işlemek için tasarlanmış yenilikçi bir AutoML Python paketidir. Bu araç veri bilimcilerin iş akışını optimize etmek, zaman ve emek tasarrufu sağlamak için geliştirilmiştir. Temel bir özelliği veri ön işleme, makine öğrenmesi modeli oluşturma ve hiperparametre ayarlama ile bağlantılı tekrarlayan süreçleri otomatikleştirme yeteneğidir. Bu süreçleri otomatikleştirerek veri bilimcilerin çalışmalarının temel unsurlarına odaklanmalarını sağlar. mljar-supervised bir kara kutu sistemi değildir. Kullanıcıların her bileşenin nasıl yapılandırıldığına dair kapsamlı bir kavrayış edinmelerine olanak sağlayarak makine öğrenmesi ardışık düzenine ilişkin bilgiler sunar. Ayrıca her makine öğrenmesi modeli için markdown formatında kapsamlı raporlar oluşturarak ayrıntılı analizlere olanak tanır. Bu paket, verisetinin anlaşılmasına ve yorumlanmasına yardımcı olmak, çeşitli makine öğrenmesi modellerini analiz etmek, analiz sırasında değerlendirilen tüm modeller hakkında ayrıntılı markdown raporları oluşturmak ve analizleri ve makine öğrenmesi modellerini kaydetme, yeniden yürütme ve yeniden yüklemeyi kolaylaştırmak gibi çok sayıda avantaj sağlar. Bu özellikler makine öğrenmesi ve veri bilimi alanındaki uzmanlar için değerli bir araç haline getirmekte ve her aşamada şeffaflığı ve kontrolü koruyup tablo halindeki veri ödevlerini etkili bir şekilde ele almalarını sağlar.

mljar-supervised ile verilerin taban çizgisi hesaplanarak makine öğrenmesinin gerekli olup olmadığı belirlenebilir, bir makine öğrenmesi modeline gerek olmayabilir. Taban çizgisi, makine öğrenmesinde bir modelin etkinliğini belirlemeye yardımcı olan bir kavramdır. Esasen bir modelin performansını mümkün olan en basit yaklaşımla karşılaştırmak için kullanılan bir ölçüttür. Ayrıca taban çizgisi makine öğrenmesi modellerinin etkinliğini ölçmek için bir referans noktası görevi görür. Örneğin modelin doğruluğu taban çizgisinden önemli ölçüde yüksekse bu modelin iyi çalıştığını ve faydalı sonuçlar ürettiğini gösterir. Tersine modelin doğruluğu taban çizgisinden düşükse modelin etkili olmadığı ve daha fazla iyileştirme gerektirdiği anlamına gelir. Taban çizgisi, sınıflandırma için önceki sınıf dağılımı ve regresyon için basit ortalama kullanılarak oluşturulur. mljar-supervised, Taban Çizgisi, Lineer Regresyon, Rassal Orman, Ekstra Ağaçlar, LightGBM, Xgboost, CatBoost, Sinir Ağları ve En Yakın Komşu algoritmalarını kullanır. Eksik

değerlerin işlenmesi ve kategorik değişkenlerin dönüştürülmesini içeren güçlü veri ön işleme desteğine sahiptir. Ayrıca çok sayıda makine öğrenmesi projesinde genellikle göz ardı edilen ancak önemli bir işlem olan hedef değerlerin ön işlemesinde de yardımcı olur. Model performansını daha da arttırmak için çok da rassal olmayan arama algoritmasının (tanımlanmış değer kümesi üzerinde rassal arama) ve tepe tırmanışı yöntemleriyle etkili bir hiperparametre ayar stratejisi kullanır. Böylece son modelleri en iyi sonuçlar için ince ayarlar. Ayrıca açgözlü algoritmaya dayalı topluluk yöntemi hesaplayabilir. mljar-supervised, modelleri üst üste koyarak seviye 2 toplulukları oluşturabilir. Her algoritma için permütasyona dayalı özellik önemi içgörüleri sağlayarak model açıklanabilirliğine öncelik verir. Özellik önemi, bağımlılık grafikleri ve karar grafikleri dahil olmak üzere SHAP açıklamaları oluşturarak kullanıcıların modellerini etkili bir şekilde anlamalarını ve yorumlamalarını sağlar. Açıklama seviyesi explain_level parametresi kullanılarak özelleştirilebilir (MLJAR 2023). MLJAR ikili sınıflandırma için kullanılabilecek örnek kod Şekil 4.9'da verilmiştir.

```
import pandas as pd
from sklearn.model_selection import train_test_split
from supervised.automl import AutoML

df = pd.read_csv(
    "https://raw.githubusercontent.com/pplonski/datasets-for-start/master/adult/data.csv",
    skipinitialspace=True,
)
X_train, X_test, y_train, y_test = train_test_split(
    df[df.columns[:-1]], df["income"], test_size=0.25
)

automl = AutoML()
automl.fit(X_train, y_train)

predictions = automl.predict(X_test)
```

Şekil 4.9: MLJAR İkili Sınıflandırma Örnek Kod (MLJAR 2023).

4.7 Makine Öğrenmesi Algoritmaları

4.7.1 Lojistik Regresyon

Lojistik regresyon, girdi değişkenlerini kullanarak sonucu belirleyen bir veya daha fazla bağımsız değişkenin bulunduğu verisetlerinde ikili çıktıların olasılığını

tahmin etmek için kullanılan istatistiksel bir modeldir. İki çıktı sınıfını ayıran bir karar sınırı öğrenen ayırt edici bir sınıflandırıcıdır. Birçok alanda önemli başarılar elde etmiş ve sınıflandırıcının tek tek veri noktalarından ziyade bir dizi örneğe dayalı tahminler yapması gereken çok örnekli öğrenme gibi daha karmaşık senaryoları kapsayacak şekilde genişletilmiştir (Wang ve diğ. 2019). Lojistik Regresyon, hedef değişkenin belirli bir kategoriye ait olma olasılığını tahmin ederek bağımlı değişken ile bir veya daha fazla bağımsız değişken arasındaki ilişkiyi modeller. Belirli bir değeri tahmin etmez ancak belirli bir sınıfa ait olma olasılığını tahmin eder. Bir dizi semptom göz önüne alındığında bir kişinin belirli bir hastalığa veya duruma sahip olup olmadığı gibi sağlıkla ilgili konular lojistik regresyonun uygun olduğu durumlara örnek olarak yaygın şekilde kullanılmaktadır (Seufert 2014).

Lojistik regresyon modeli, herhangi bir reel değerli girdiyi 0 ile 1 arasında bir değere eşleyen matematiksel bir fonksiyon olan lojistik yada sigmoid fonksiyona dayanmaktadır.

Sigmoid fonksiyon;

$$f(z)=\frac{1}{1+e^{-z}}=\frac{e^{z}}{e^{z}+1}=1-f(-z) \tag{4.1}$$

z'nin *x*'e bağlı lineer fonksiyon olduğu varsayılırsa *z* aşağıdaki gibi ifade edilebilir.

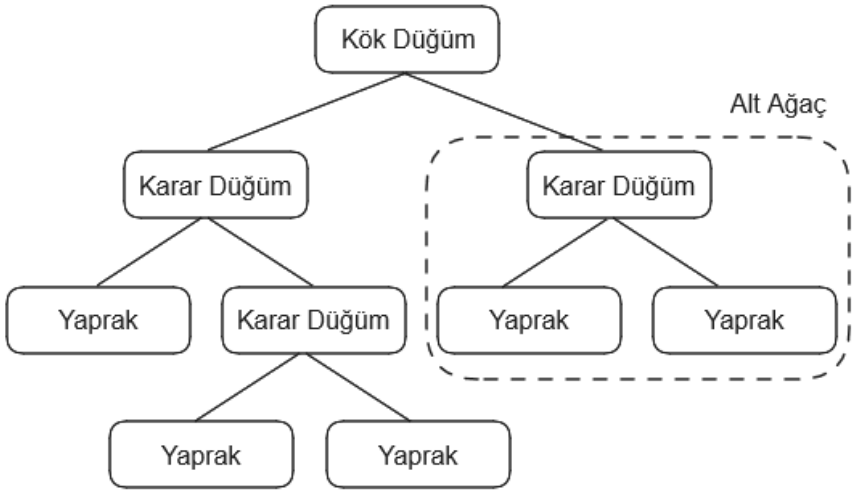
$$z=\beta_0+\beta_1 x \tag{4.2}$$

Genel lojistik fonksiyon ise aşağıdaki gibi yazılabilir.

$$f(z)=\frac{1}{1+e^{-(\beta_0+\beta_1 x)}} \tag{4.3}$$

4.7.2 Karar Ağacı

Karar Ağacı, hem sınıflandırma hem de regresyon problemleri için kullanılabilen denetimli makine öğrenmesi algoritmasıdır. Baş aşağı bir ağaca benzeyen bir karar verme sürecinin grafiksel ifadesidir. Kök düğüm bu sürecin başlangıç noktasını temsil eder.



Şekil 4.10: Örnek Karar Ağacı (Sneha ve Gangil 2019).

Karar ağacı her bir özelliği değerlendirirken bir yaprak düğüme ulaşana kadar bir dizi dalı takip eder. Ağaçtaki her iç düğüm bir girdi özelliğinin değerlendirilmesine karşılık gelir ve dallar potansiyel özellik değerleriyle etiketlenir. Fonksiyon tarafından belirlenen çıktı yaprak düğümleri ile gösterilir (Russel ve Norvig 2020). Eldeki verilerin nasıl bölüneceğine karar vermek için kullanılan çeşitli teknikler vardır. Karar ağaçlarının temel amacı verileri en iyi şekilde doğru kategorilere ayıracak düğümler arasında en iyi bölünmeleri yapmaktır. Bunu yapmak için doğru karar kuralları kullanmak gerekir. Kurallar, algoritmanın performansını doğrudan etkileyen faktörlerdir. Karar ağaçları oluşturulurken yapılan bazı varsayımlar bulunur. Kök değerlendirmesi için başlangıçta tüm veriseti kök düğüm olarak ele alınır. Daha sonra bölümler oluşturmak ya da kökü daha küçük alt ağaçlara ayırmak için algoritmalar devreye girer. Özellik değerlerinin kategorik olduğu varsayılır. Değerlerin sürekli olduğu durumlarda ağaç modelini oluşturmadan önce değerler ayrılırlar. Algoritma, verileri farklı özelliklerin değerlerine göre

özyinelemeli olarak bölümlere ayırır. Verileri farklı gruplara en iyi şekilde ayıran bir özellik seçilir ve ardından bu özelliğe dayalı olarak ağaçta bir düğüm oluşturur. Veriler daha sonra bu özelliğin değerlerine göre alt kümelere ayrılır ve aynı işlem tüm alt kümeler saf olana veya ağaç önceden tanımlanmış bir durdurma kriterine ulaşana kadar her alt küme için özyinelemeli olarak tekrarlanır. İstatistiksel özellik sıralaması yapılır; hangi özelliklerin ağacın kök veya iç düğümleri olarak kullanılacağının belirlenmesi en etkili karar verme sürecini sağlamak için istatistiksel bir yaklaşım izler. Gini, Ki-kare ve bilgi kazanımı özellik seçimi ve değerlendirmesi için kullanılır. Gini indeksi, bir karar ağacındaki bir düğümün safsızlığını ölçen bir bölme kriteridir (Zaman ve diğ. 2020). Genellikle sınıflandırma problemlerinde bir bölünmenin iyiliğini değerlendirmek için kullanılır. Ki-kare testi, iki kategorik değişken arasındaki bağımsızlığı belirlemek için kullanılan istatistiksel bir testtir. Bir özellik ile hedef değişken arasındaki ilişkiyi ölçerek özellik seçimi için kullanılabilir. Bilgi kazancı, bir ayrımın kalitesini değerlendirmek için karar ağacı algoritmalarında kullanılan başka bir ölçüdür. Bir bölme yapıldıktan sonra entropi veya kirlilikteki azalmayı hesaplar. Sınıflandırma için en bilgilendirici özelliklerin seçilmesine yardımcı olur.

4.7.3 K-En Yakın Komşu

K-En Yakın Komşu (KNN), modelin yeni örneklerin sonucunu tahmin etmek için tüm eğitim setini kullandığı örnek tabanlı bir denetimli makine öğrenmesi algoritmasıdır ve regresyon ve sınıflandırma problemleri için kullanılabilir. Parametrik olmayan bir algoritmadır yani verilerin dağılımı hakkında herhangi bir varsayımda bulunmaz (Niwas ve diğ. 2013). Sınıflandırma problemlerinde bir veri noktasının hangi sınıftan olduğunu belirlemek için algoritma yakınındaki en yakın k veri noktalarının durumlarına bakar. Veri noktalarının çoğunluğu hangi sınıftansa söz konusu veri noktasının o sınıftan olması daha olasıdır. Regresyon problemlerinde ise k en yakın komşusunun değerlerine dayanarak sonuç tahmin eder. Eğitim setinde girdi veri noktasına en yakın k veri noktasını bularak ve ardından çıktıyı tahmin etmek için değerlerini kullanarak çalışır. Tahmin, k en yakın komşunun çıktı değerlerinin ortalaması veya ağırlıklı ortalaması alınarak yapılır.

Yakın komşu seçimi yaparken kullanılabilecek metriklerden bazıları Öklidyen, Minkowski, Hamming, Manhattan uzaklıklarıdır. Öklidyen uzaklık KNN'de en yaygın kullanılan uzaklık ölçütlerinden biridir. Çok boyutlu bir özellik uzayında iki veri noktası arasındaki düz çizgi mesafesini ölçer. Manhattan uzaklık veri noktaları arasındaki mesafeleri, karşılık gelen özellikleri arasındaki mutlak farkların toplamına göre hesaplamak istendiğinde kullanılır. Minkowski uzaklık hem Öklidyen hem de Manhattan uzaklıklarının genelleştirilmiş halidir ve KNN'de farklı boyutlara duyarlılığı kontrol etmeyi sağlayan bir parametre ile kullanılır. Hamming uzaklık ikili veri ile çalışırken veya ikili vektörleri karşılaştırırken kullanılır.

Öklidyen uzaklık;

$$\sqrt{\sum_{i=1}^{k} (x_i - y_i)^2} \tag{4.4}$$

Manhattan uzaklık;

$$\sum_{i=1}^{k} |x_i - y_i| \tag{4.5}$$

Minkowski uzaklık;

$$\left(\sum_{i=1}^{k} |x_i - y_i|^p \right)^{\frac{1}{p}} \tag{4.6}$$

4.7.4 Naive Bayes

Naive Bayes, makine öğrenmesi alanında özellikle denetimli sınıflandırma ve olasılıksal modellemede popüler bir algoritmadır. Naive Bayes yöntemleri sınıf değişkeninin değeri göz önüne alındığında her bir özellik çifti arasında koşullu bağımsızlığın varsayımı ile olasılık teorisinde temel bir kavram olan Bayes teoreminin uygulanmasına dayanır. Naive Bayes, spam e-posta algılama, duygu belirleme, kötü amaçlı yazılım analizi, belgeleri kategorize etme ve diğer sınıflandırma problemlerini çözme gibi görevler için yaygın olarak kullanılmaktadır.

Bayes teoreminin matematiksel ifadesi;

$$P(A|B)=\frac{P(B|A)P(A)}{P(B)} \tag{4.7}$$

$P(A|B)$: B olayı gerçekleştiğinde A olayının gerçekleşme olasılığı, sonsal olasılık, koşullu olasılık.

$P(B|A)$: A olayı gerçekleştiğinde B olayının gerçekleşme olasılığı, olabilirlik, koşullu olasılık.

$P(A)$: A olayının gerçekleşme olasılığı, öncül olasılık.

$P(B)$: B olayının gerçekleşme olasılığı, marjinal olasılık.

Naive Bayes matematiksel ifadesi;

$$\hat{y}=\underset{y}{argmax}\, P(y)\prod_{i=1}^{n} P(x_i|y) \tag{4.8}$$

Naive Bayes sınıflandırıcılarının üç ana türü vardır:

- Gaussian Naive Bayes, sürekli verilerle çalışırken yapılan tipik bir varsayım her bir sınıfla ilişkili sürekli değerlerin Gauss dağılımı göstermesidir. Özelliklerin normal bir dağılım izlediğini varsayarak tahmin edicilerin ayrık yerine sürekli değerler alması durumunda modelin bu değerlerin Gauss dağılımından örneklendiğini varsayması anlamına gelir.

- Multinomial Naive Bayes, veriler çok terimli dağıtılım gösterdiğinde tercih edilir. Genellikle dökümana kategori atamak gibi metin sınıflandırma problemlerinde kullanılır. Belgelerdeki kelimelerin veya özelliklerin görülme sıklığından yararlanarak çalışır. Farklı kategorilerle ilişkili ayırt edici kelime kalıplarını ve dağılımlarını yakalamada başarılıdır.

- Bernoulli Naive Bayes, ayrık veriler için kullanılır ve verisetideki özelliklerin çok değişkenli Bernoulli dağılımı gösterdiği varsayımı ile çalışır. Özellikleri

yalnızca doğru veya yanlış, 1 veya 0 gibi ikili değerler olarak kabul eder. Özellik değerleri ikili olduğunda tercih edilir.

4.7.5 Destek Vektör Makinesi

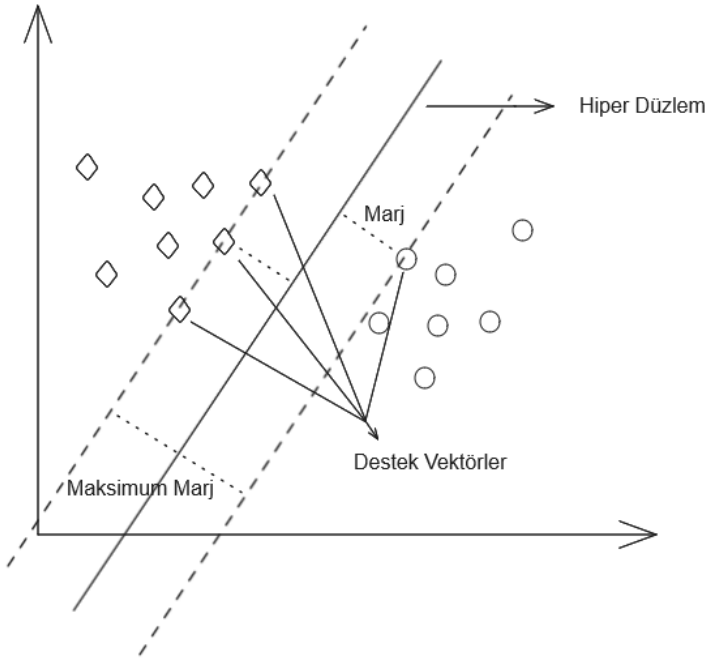
Destek Vektör Makinesi (SVM) (Cortes ve Vapnik 1995), sınıflandırma, regresyon ve aykırı değer belirleme problemleri için kullanılabilen bir denetimli makine öğrenmesi algoritmasıdır. Bir özellik uzayında farklı sınıflara ait veri noktalarını en iyi şekilde ayıran optimum hiper düzlemi bularak çalışır.

Hiper düzlem, iki boyutlu uzayda bir çizgidir. Daha yüksek boyutlu uzaylarda ise hiper düzlem girdi uzayından bir boyut daha küçük bir alt uzaydır. Destek Vektör Makineleri'nin amacı farklı sınıfların en yakın iki veri noktası arasındaki mesafeyi maksimize eden hiper düzlemi bulmaktır. Bu en yakın veri noktalarına destek vektörleri denir. Marj, hiper düzlem ile destek vektörleri arasındaki mesafedir. Destek Vektör Makineleri marjı maksimize etmeyi amaçlar çünkü daha büyük marj genel olarak daha iyi genelleme ve sınıflandırma performansı sağlar. Çekirdek Yöntemi ile Destek Vektör Makineleri bir çekirdek fonksiyonu kullanarak özellik uzayını dönüştürerek doğrusal olmayan ayrılabilir verileri işleyebilir. Yaygın çekirdek fonksiyonları arasında doğrusal, polinom, radyal bazlı fonksiyon (RBF) ve sigmoid bulunur. Bu dönüşüm, Destek Vektör Makineleri'nin dönüştürülmüş uzayda verileri etkili bir şekilde ayıran bir hiper düzlem bulmasını sağlar (Saini 2021).

Destek Vektör Makineleri'nin avantajlarından biri yüksek boyutlu uzaylarda iyi performans gösterebilmesidir. Boyut sayısı örnek sayısından fazla olduğunda da etkili sonuçlar verebilir. Karar fonksiyonunda destek vektörleri olarak bilinen eğitim noktalarının bir alt kümesi kullanıldığından bellek açısından verimlidir. Ayrıca karar fonksiyonu için farklı çekirdek fonksiyonları belirlenebildiğinden Destek Vektör Makineleri çok yönlüdür. Yaygın çekirdekler doğrusal, polinom ve radyal baz fonksiyonudur, özel çekirdekler belirlemek de mümkündür. Bu Destek Vektör Makineleri'nin çeşitli sınıflandırma problemlerine uyarlanmasına olanak tanır.

Nokta çarpımı lineer çekirdek için kullanılan benzerlik ölçüsüdür çünkü mesafe girdilerin lineer bir kombinasyonudur. Yeni veriler ile destek vektörleri

arasındaki benzerliği veya mesafe ölçüsünü tanımlar. Polinom çekirdek, lineer çekirdeğin daha genel halidir, verileri daha yüksek boyutlu uzaya eşlemek için bir polinom fonksiyonu kullanır. Bu, polinom çekirdeğinin lineer olmayan girdi uzayını ayırt etmesini sağlar. Radyal tabanlı fonksiyon çekirdek, Destek Vektör Makinesi sınıflandırmasında yaygın olarak kullanılan ve bir girdi uzayını sonsuz boyutlu uzaya eşleyebilen bir çekirdek fonksiyonudur. Şekil 4.11'de Destek Vektör Makinesi örneği bulunmaktadır.

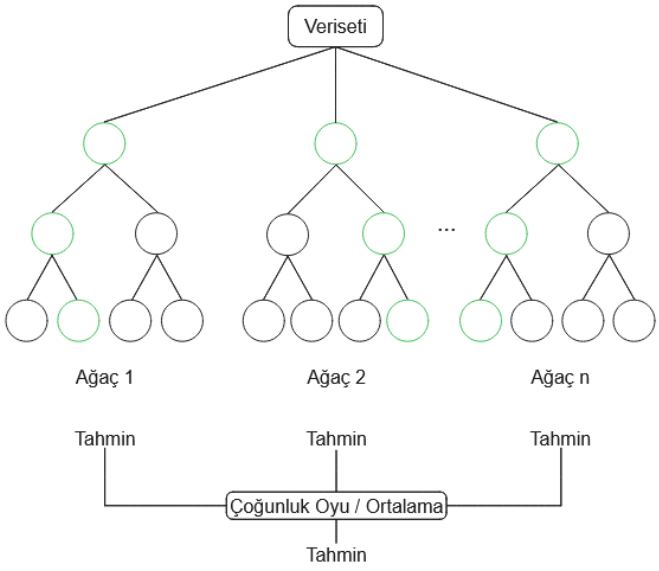


Şekil 4.11: Örnek Destek Vektör Makinesi (Mammone ve diğ. 2009).

4.7.6 Rassal Orman

Rassal Orman (Breiman 2001), değerlendirme için farklı rassal özellik alt uzayları kullanan dolayısıyla daha iyi genelleme performansı sağlayan bir karar ağaçları topluluğudur (Ganaie ve diğ. 2022). Torbalama (Breiman 1996), bir eğitim verisetinin farklı alt kümelerine birden fazla model uyduran ve ardından tüm modellerden gelen tahminleri birleştiren bir topluluk algoritmasıdır. Torbalama yönteminin genişletilmiş hali olan Rassal Orman da her bir veri örneğinde kullanılan özelliklerin alt kümelerini rassal olarak seçer. Rassal Orman'da topluluktaki her ağaç eğitim setinden çekilen bir örnekten oluşturulur. Her ağacın tahmini bağımlı

değişkenin tahmin edilen değerini vermek için regresyon problemlerinde ortalaması alınır veya sınıflandırma problemlerinde bir sınıf üzerinde oylanır. Ancak çarpık sınıf dağılımına sahip sınıflandırma problemleri için uygun olmayabilirler (Chen ve diğ. 2004). Şekil 4.12'de Rassal Orman örneği bulunmaktadır.



Şekil 4.12: Örnek Rassal Orman (Kibria ve Matin 2022).

4.7.7 Stokastik Gradyan İniş

Gradyan İniş, makine öğrenmesinde ve diğer alanlarda yaygın olarak kullanılan bir optimizasyon algoritmasıdır ve amacı modelin ne kadar iyi performans gösterdiğinin bir ölçüsü olan maliyet fonksiyonunu en aza indirmek için modelin parametrelerini iteratif olarak ayarlamaya çalışır (Wang ve diğ. 2021). Gradyan inişinin arkasındaki temel fikir bir başlangıç parametre seti ile başlamak ve ardından bunları maliyet fonksiyonunun en dik gradyan yönünde iteratif olarak ayarlamaktır. Bu süreç maliyet fonksiyonunun her bir parametreye göre kısmi türevlerinin hesaplanması ve ardından parametreleri maliyet fonksiyonunu azaltacak şekilde güncellemek için bu türevlerin kullanılmasıyla gerçekleşir (Géron 2022). Gradyan İniş doğrusal regresyon, lojistik regresyon, sinir ağları ve diğerleri de dahil olmak üzere çok çeşitli problemleri çözmek için kullanılabilen güçlü ve esnek bir optimizasyon algoritmasıdır. Özellikle maliyet fonksiyonunun türevlenebilir ve

sürekli olduğu ve optimize edilmesi gereken çok sayıda parametrenin bulunduğu problemler için uygundur.

Çok sayıda eğriliğe veya gürültülü gradyanlara sahip optimizasyon problemlerinde arama uzayında sıçrayabilmesi ve arama uzayında gradyanı olmayan düz noktalarda sıkışıp kalabilmesi Gradyan İniş yönteminde karşılaşılabilen bir sorundur. Gradyan İniş performansını artırmak için kullanılabilecek Stokastik Gradyan İniş, Mini Toplu Gradyan İniş ve Momentum Tabanlı Gradyan İniş gibi farklı çeşitleri vardır. Bunlar algoritmanın hesaplama karmaşıklığını azaltmaya, yakınsama özelliklerini iyileştirmeye ve yerel minimumlarda takılıp kalmasını önlemeye yardımcı olabilir.

Stokastik Gradyan İniş ile Gradyan İniş'in aksine tek bir parametre güncellemesi yapmak için eğitim setindeki tüm veri noktaları yinelenip modelin parametrelerini yalnızca eğitim setindeki tüm veri noktalarını yineledikten sonra güncellemek gerekmez. Modelin parametreleri her bir veri noktası yinelendikten sonra güncellendiği için optimum parametre daha hızlı öğrenilir dolayısıyla daha hızlı yakınsama sağlanır ve bu da eğitim süresini azaltır.

4.7.8 AdaBoost

Artırma, zayıf bir sınıflandırıcıyı güçlü bir sınıflandırıcıya dönüştürme işlemidir. Adaptif Artırma kısaca AdaBoost (Adaptive Boosting) (Freund ve Schapire 1997), hem ikili sınıflandırma hem de regresyon algoritmaları için geliştirilmiş çok başarılı bir Artırma algoritmasıdır. Güçlü bir öğrenici oluşturmak için çoklu yineleme kullanır ve her bir zayıf öğrenicinin tahminini birleştirerek birkaç zayıf öğreniciden güçlü bir öğrenici elde edilir. AdaBoost algoritması eğitimin her turunda topluluğa yinelemeli olarak yeni bir zayıf öğrenici ekleyerek güçlü bir öğrenici oluşturur. Önceki turlarda yanlış sınıflandırılan veri noktalarının sınıflandırmasını düzeltmek için her turda ağırlık vektörü ayarlanır (Thomas ve diğ. 2020). Her yinelemeden sonra yanlış sınıflandırılmış örneklerin ağırlıkları artırılır ve doğru sınıflandırılmış örneklerin ağırlıkları azaltılır.

4.7.9 XGBoost

XGBoost (eXtreme Gradient Boosting), yüksek verimlilik, esneklik ve taşınabilirlik sağlamak için optimize edilmiş, açık kaynak kodlu, dağıtık Gradyan Artırma kütüphanesidir. Gradyan Artırma çerçevesinde makine öğrenmesi algoritmaları uygular. Paralel ağaç artırma yöntemi ile birçok veri bilimi problemini hızlı ve doğru bir şekilde çözmeyi sağlar. Paralel ağaç artırma kullanılması birden fazla karar ağacının aynı anda artırılmasına olanak tanır ve bu özellik algoritmanın büyük verisetlerinde de hızlı, verimli sonuçlar elde etmesini sağlar.

XGBoost'un etkisi makine öğrenmesi ve veri madenciliği yarışmalarında geniş çapta kabul görmüştür ve çok çeşitli problemlerde üstün sonuçlar vermektedir. XGBoost'un başarısının arkasındaki en önemli faktör tüm senaryolarda ölçeklenebilir olmasıdır. Sistem tek bir makinede mevcut popüler çözümlerden on kat daha hızlı çalışır ve dağıtık veya bellek sınırlı ortamlarda milyarlarca örneğe ölçeklenir. XGBoost'un ölçeklenebilirliği, seyrek verileri işlemek için yeni bir ağaç öğrenme algoritması, teorik olarak gerekçelendirilmiş ağırlıklı kantil taslak prosedürü ile yaklaşık ağaç öğrenmede örnek ağırlıklarının işlenmesini sağlaması, paralel ve dağıtık hesaplamayla öğrenmeyi daha hızlı hale getirerek daha hızlı model keşfi sağlaması gibi bazı önemli sistem ve algoritmik optimizasyonlar ile sağlanmaktadır (Chen ve Guestrin 2016).

4.7.10 LightGBM

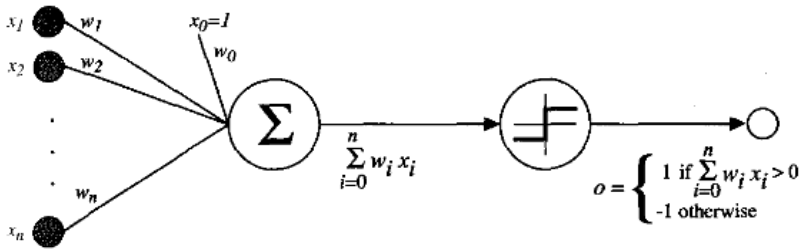
Gradyan Artırma Karar Ağacı (GBDT) verimliliği, doğruluğu ve yorumlanabilirliği ile bilinen, çok sınıflı sınıflandırma ve sıralama gibi çeşitli makine öğrenmesi problemlerinde başarılı olan popüler bir makine öğrenmesi algoritması olmasına rağmen verilerde çok sayıda örnek ve özellik bulunduğunda ölçeklenebilirliği ve verimliliği yeterli olmayabilir. Bunun sebebi her bir özelliğin tüm potansiyel bölünme noktalarından bilgi kazanımı tahmin etmek için tüm veri örneklerinin taranmasıdır. LightGBM bu problemin üstesinden gelmek için geliştirilmiştir. LightGBM, çok sayıda veri örneği ile başa çıkmak için Gradyan Tabanlı Tek Taraflı Örnekleme (GOSS) ve çok sayıda özellik ile başa çıkmak için Özel Özellik Birleştirme (EFB) tekniği içerir (Ke ve diğ. 2017).

LightGBM, ağaç tabanlı öğrenme algoritmaları kullanan bir Gradyan Artırma çerçevesidir. Daha hızlı eğitim hızı ve daha yüksek verimlilik, daha düşük bellek kullanımı, daha iyi doğruluk, paralel, dağıtık olma ve GPU öğrenme desteği, büyük ölçekli verileri işleme yeteneği avantajlarıyla dağıtık ve verimli olacak şekilde tasarlanmıştır (LightGBM 2023[a]). Çoğu karar ağacı algoritması ağaçları büyütmek için seviye (derinlik) bazlı bir yaklaşım izlerken LightGBM maksimum delta maliyetine sahip yaprağı seçerek yaprak bazlı bir yaklaşım izler. Bu yaklaşım seviye bazlı algoritmalara göre daha az maliyet ile sonuçlanma eğilimindedir ancak veri sınırlı olduğunda aşırı öğrenmeye neden olma olasılığı daha yüksektir. Aşırı öğrenmeye karşı ağaç derinliğini sınırlamaya yarayan maksimum derinlik parametresi bulunur. Bu parametre belirtilse bile ağaçları yaprak bazında büyütmeye devam etmektedir (LightGBM 2023[b]).

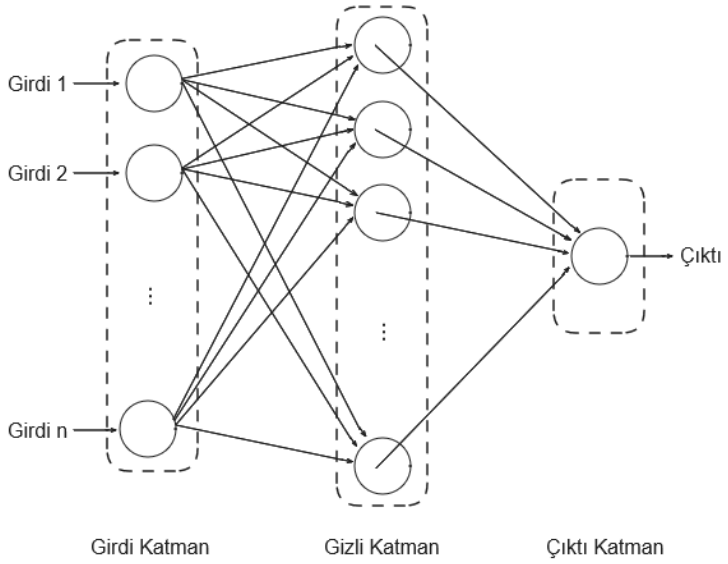
4.7.11 Yapay Sinir Ağları

Yapay Sinir Ağı (ANN), insan beyninin yapısı ve işleyişinden esinlenilerek karmaşık verileri işlemek ve analiz etmek için tasarlanmış bir hesaplamalı modeldir. Nöron veya düğüm, çeşitli araştırma alanlarında yaygın olarak kullanılan yapay sinir ağlarının temel yapı taşıdır, bilgi iletimini sağlar.

Yapay nöron olarak da bilinen algılayıcı (perceptron), girdiler, ağırlıklar ve bias, toplam fonksiyonu, aktivasyon fonksiyonu ve çıktıdan oluşur. Algılayıcı girdi verileri üzerinde hesaplamalar gerçekleştirir. Giriş sinyallerini alır, bunlara ağırlıklar uygular, toplar ve bir çıktı üretmek için sonucu bir aktivasyon fonksiyonundan geçirir. Girdilere atanan ağırlıklar girdilerin önemini belirler ve aktivasyon fonksiyonu modele lineer olmama özelliği katarak yapay sinir ağlarının karmaşık örüntüleri öğrenmesini ve karar vermesini sağlar. Yapay sinir ağları birbirine bağlı katmanlardan oluşur ve temel olarak girdi katmanı, gizli katman ve çıktı katmanı bulunur. Şekil 4.13'te Algılayıcı, Şekil 4.14'te ise Çok Katmanlı Algılayıcı bulunmaktadır.



Şekil 4.13: Perceptron (Algılayıcı) (Mitchell 1997).



Şekil 4.14: Çok Katmanlı Algılayıcı (Fath ve diğ. 2018).

4.8 Değerlendirme Metrikleri

Değerlendirme metrikleri modelin ne kadar iyi performans gösterdiğini, eğitildiği ve test edildiği veriler üzerinde doğru tahminler yapıp yapmadığını anlamaya yardımcı olur ve modele iyileştirmeler yapmak için yol gösterir. Değerlendirme metrikleri farklı modelleri karşılaştırmak ve belirli bir problemde en iyisini seçmek için önemlidir. Tek başına bir metrik, modeli değerlendirmek için yeterli olmayabilir veya doğru olmayan çıkarımlara sebep olabilir bu yüzden çeşitli metriklerin birlikte değerlendirilmesi farklı problemlerde modeli daha doğru yorumlamaya olanak tanır. Sınıflandırma için kullanılan bazı yaygın değerlendirme metrikleri arasında karışıklık matrisi (confusion matrix), doğruluk (accuracy),

duyarlılık (recall), hassasiyet (precision), F1-skor, ROC eğrisi (receiver operating characteristic curve) ve AUC (Area under the ROC Curve) bulunmaktadır.

4.8.1 Karışıklık Matrisi

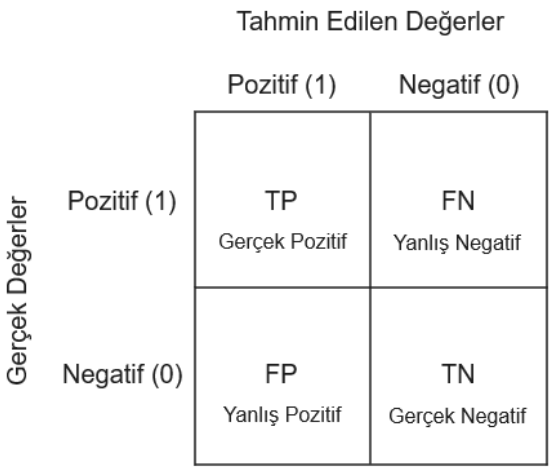
Karışıklık matrisi veya hata matrisi (Congalton ve diğ. 1983) bir sınıflandırma modelinin performansını gerçek değerleri mevcut olan bir test verisi üzerinde karşılaştırılarak görselleştirmeye yarayan tablodur. İkili sınıflandırma problemleri için 2x2 boyutunda matris kullanılır ancak birden fazla sınıf içeren verilerde daha fazla satır ve sütunla çok sınıflı sınıflandırmaya uygun olacak şekilde genişletilebilir.

Gerçek Pozitif (True Positive, TP): Bu durumlar modelin verisetindeki pozitif sınıfı asıl gerçeğe paralel olarak doğru tahmin ettiği durumlardır. Örneğin bir tıbbi teşhis bağlamında TP, bir hastalığı olan doğru tanı koyulmuş bireylerin sayısını temsil eder.

Gerçek Negatif (True Negative, TN): Bu durumlar modelin verisetindeki negatif sınıfı asıl gerçeğe paralel olarak doğru tahmin ettiği durumlardır. Doğru tanı koyulmuş sağlıklı bireylerin sayısını temsil eder.

Yanlış Pozitif (False Positive, FP): Bu durumlar modelin verisetindeki negatif sınıfı tahmin etmesi gerekirken asıl gerçeğe aykırı olarak pozitif sınıfı tahmin ettiği durumlardır. FP aynı zamanda Tip I hata olarak da bilinmektedir. Yanlışlıkla hastalık teşhisi konmuş aslen sağlıklı olan bireylerin sayısını temsil eder.

Yanlış Negatif (False Negative, FN): Bu durumlar modelin verisetindeki pozitif sınıfı tahmin etmesi gerekirken asıl gerçeğe aykırı olarak negatif sınıfı tahmin ettiği durumlardır. FN aynı zamanda Tip II hata olarak da bilinmektedir. Yanlışlıkla sağlıklı olarak teşhis edilen ve aslen bir hastalığı olan bireylerin sayısını temsil eder.



Şekil 4.15: Karışıklık Matrisi (Stazio ve diğ. 2019).

4.8.2 Doğruluk

Doğruluk, eğitilmiş makine öğrenmesi modelinin elde ettiği doğru sınıflandırmaların oranıdır yani doğru yapılan tahminlerin sayısının tüm sınıflar için yapılan toplam tahmin sayısına bölünmesiyle elde edilir. Modelin doğru çıktıyı ne sıklıkla tahmin ettiğini tanımlar.

$$Doğruluk = \frac{TP + TN}{TP + TN + FP + FN} \tag{4.9}$$

4.8.3 Duyarlılık

Duyarlılık veya gerçek pozitif oranı (TPR), modelin doğru şekilde pozitif olarak sınıflandırdığı örnek sayısının verideki toplam pozitif örnek sayısına olan oranıdır. Gerçek pozitiflerin model tarafından ne kadarının doğru tahmin edildiğini tanımlar. Model performansının iyi olarak değerlendirilebilmesi için olabildiğince yüksek olmalıdır.

$$Duyarlılık = \frac{TP}{TP + FN} \tag{4.10}$$

4.8.4 Hassasiyet

Hassasiyet, modelin doğru şekilde tahmin ettiği tüm pozitif sınıflardan kaçının gerçekten doğru olduğunu tanımlar.

$$Hassasiyet = \frac{TP}{TP + FP} \tag{4.11}$$

4.8.5 F1-Skor

F1-skor, hassasiyet ve duyarlılık metriklerini birlikte ele alır ve bunların harmonik ortalaması ile hesaplanır. Model performansının dengeli bir şekilde değerlendirilmesini sağlar. Hassasiyet ve duyarlılık değerlerinin F1-skora katkısı eşittir. F1-skor için en iyi değer 1 ve en kötü değer 0'dır.

$$\text{F1-Skor} = \frac{2}{\frac{1}{Hassasiyet} + \frac{1}{Duyarlılık}} = \frac{2 * Hassasiyet * Duyarlılık}{Hassasiyet + Duyarlılık} \tag{4.12}$$

4.8.6 ROC Eğrisi ve AUC

ROC (Alıcı İşletim Karakteristikleri) eğrisi, gerçek pozitif oranı (duyarlılık, TPR) ile yanlış pozitif oranı (FPR) arasındaki ilişkiyi gösterir. Yanlış pozitif oranı modelin yanlışlıkla pozitif olarak sınıflandırdığı örnek sayısının verideki toplam negatif örnek sayısına olan oranıdır. Yanlış pozitif oranı x ekseninde, gerçek pozitif oranı y ekseninde gösterilir.

$$Yanlış\ Pozitif\ Oranı = \frac{FP}{FP + TN} \tag{4.13}$$

AUC (Area Under the ROC Curve) yani ROC eğrisinin altında kalan alan, ROC eğrisinin 0 ile 1 arasında yer alan bir sayısal değer ile özetidir. Daha yüksek AUC değeri daha iyi model performansına işaret eder. Modelin verisetindeki sınıfları ayırt etme yeteneğinin değerlendirmesidir. Kusursuz model 1 değerini alırken rassal bir sınıflandırıcı 0.5 değeri alır.

4.8.7 Tekrarlı Katmanlı k-Kat Çapraz Doğrulama

Çapraz Doğrulama, bir modelin yeni verilere genellenebilirliğini değerlendirmek için modeli eğitim setine dahil edilmemiş bir veya daha fazla farklı veri alt kümesi üzerinde test etme yöntemidir.

Çapraz doğrulamaya yönelik temel yaklaşım k-kat çapraz doğrulama olarak bilinmektedir ve diğer varyasyonlar için temel niteliğindedir. Bazıları k-kat çapraz doğrulamanın özel durumlarıyken bazıları birden fazla tekrarını kullanır. K-kat çapraz doğrulamada veriseti başlangıçta her biri yaklaşık olarak eşit büyüklükte olan k kata ayrılır. Daha sonra k defa eğitim ve doğrulama işlemi gerçekleştirilir. Her yinelemede verilerin farklı bir katı doğrulama için ayrılırken kalan k-1 kat eğitim için kullanılır (Refaeilzadeh ve diğ. 2009, James ve diğ. 2013).

Güvenilir ve tutarlı değerlendirmeler elde etmek için çok sayıda performans değerlendirmesine sahip olmak daha iyidir. K-kat çapraz doğrulama ile modelin performansına ilişkin verilerin k farklı alt kümesine bağlı olarak yalnızca k tahmin elde edilir. Değerlendirme sayısını artırmak ve veri bölünmesindeki rassallığın etkisini azaltmak için k-kat çapraz doğrulama birden fazla kez çalıştırılabilir. Veriler her seferinde k kata bölünmeden önce karıştırılır. Bu yöntem tekrarlı k-kat çapraz doğrulamadır.

Katmanlı k-kat çapraz doğrulama ise her katın verisetiyle aynı sınıf dağılımında olmasını sağlayan bir varyasyondur. Her kat, her bir hedef sınıftan yaklaşık olarak tüm verisetiyle aynı oranda örnek içerir. Bu özellikle bazı sınıfların diğerlerinden çok daha az örneğe sahip olduğu dengesiz verisetleri ile uğraşırken etkili yaklaşımlardan biridir. Ancak verilerin farklı bölümleri farklı sonuçlara yol açabileceğinden bir kez çalıştırılan katmanlı k-kat çapraz doğrulama modelin

performansının gürültülü şekilde değerlendirmesine neden olabilir. Tekrarlı katmanlı k-kat, her tekrarda farklı rassallaştırma ile katmanlı k-kat çapraz doğrulamanın n kez tekrarlanmasıdır.

4.9 Ölçekleme

Veri normalizasyonu bir verisetinin özelliklerini standart bir aralığa ölçekleme işlemidir (Pandey ve Jain 2017). Bu genellikle tüm özelliklerin yapılan analize eşit şekilde katkıda bulunmasını sağlamak ve herhangi bir özelliğin daha büyük ölçeği nedeniyle baskın olmasını önlemek amacıyla yapılmaktadır. Bu çalışmada normalizasyon aşamasında kullanılan normalleştirici ile örnekler ayrı ayrı birim norma normalleştirilmiştir. En az bir sıfır olmayan bileşene sahip her örnek yani verisetindeki her satır, l2 normu bire eşit olacak şekilde diğer örneklerden bağımsız olarak yeniden ölçeklendirilmiştir.

Standardizasyon, z-skor normalizasyon olarak da bilinen ve verisetindeki özelliklerin aritmetik ortalaması 0 standart sapması 1 olacak şekilde verisetinin dönüştürülmesini sağlayan bir özellik ölçekleme yöntemidir. Bütün özelliklerin modele eşit katkıda bulunması için özelliklerin aynı ölçekte olmasını sağlayarak makine öğrenmesi algoritmalarının veriyi doğru şekilde işlemesini sağlar.

Standardizasyon işleminde verisetindeki her özellik için ortalama değer hesaplanır. Ardından özellik içindeki her veri noktasından aritmetik ortalama çıkarırılır. Böylece veriler sıfır merkezlenir. Bu değerlerin her biri özelliğin standart sapmasına bölünür. Standart sapma, verilerin ne kadar dağınık olduğunu belirtir. Veriler ölçeklendirilerek standart sapmanın 1 olması sağlanır. μ özellik değerlerinin aritmetik ortalaması, σ özellik değerlerinin standart sapması olmak üzere:

$$z = \frac{x - \mu}{\sigma} \tag{4.14}$$

4.10 Temel Bileşen Analizi

Temel bileşen analizi (TBA, Principal Components Analysis, PCA) (Pearson 1901), boyutluluk azaltmak için kullanılan, keşifsel veri analizi, ön işleme, görselleştirme gibi uygulama alanları olan bir yöntemdir. Temel bileşen analizi ile veriler temel bileşenler olarak bilinen yeni bir koordinat sistemine doğrusal olarak dönüştürülür ve bu şekilde verilerdeki en büyük varyasyonu yakalayan yönlerin tanımlanması sağlanır. Verisetindeki değişken sayısını azaltırken olabildiğince fazla varyasyonu korumak amaçlanır ve orijinal değişkenlerin birbirinden bağımsız olan temel bileşenlere dönüştürülmesiyle gerçekleştirilir. Bu temel bileşenler ilk birkaçı başlangıçtaki değişkenlerde mevcut olan varyasyonun çoğunu koruyacak şekilde sıralanır (Mishra ve diğ. 2017). Bu yöntem özelliklerin standardizasyonunu ve özelliklerin kovaryans matrisinin hesaplanmasını içerir. Kovaryans matrisi değerleri n boyutlu bir uzayda yönleri temsil eden vektörler olarak ele alınır. Daha sonra bu vektörlerin ortalaması alınarak özvektörler oluşturulur ve daha büyük özdeğerler daha fazla ortalama kovaryans vektörünü simgeler. Özvektörler özdeğerlerine göre sıralanır ve istenen bir varyans eşiği seçilir. Seçilen özvektörler, standardize edilmiş özellikleri temel bileşenler matrisi ile çarparak giriş verilerini sıkıştırmak için kullanılır (Mikulski 2019). Temel bileşen analizinde en önemli adım değişkenkenlerin standardizasyonudur çünkü değişkenlerin varyanslarının açıklanması amaçladığından bazı değişkenlerin bu varyansa aşırı katkıda bulunmamasını sağlamak önemlidir (Greenacre ve diğ. 2022).

4.11 Veri Örnekleme

4.11.1 SMOTE

Verisetinde azınlık sınıftaki örneklerin çok az sayıda olması makine öğrenmesi algoritmalarının azınlık sınıfı yeteri kadar öğrenmemesine sebep olabilir dolayısıyla model performansını olumsuz etkiler. Dengesiz verisetlerindeki bu çarpık sınıf dağılımı sınıflandırıcıların çoğunlukta olan sınıfı tahmin etmeye eğilimli olmasına neden olur (Maimon ve Rokach 2010, Gupta ve diğ. 2020). SMOTE

(Sentetik Azınlık Örnekleme Tekniği) (Chawla ve diğ. 2002), dengesiz verisetlerinde karşılaşılan bu sorunu çözmek için yaygın şekilde kullanılan bir veri artırma yöntemidir. SMOTE azınlık sınıfın sentetik örneklerini oluşturmak için önce azınlık sınıfı örneklerinin k-en yakın komşularını belirler ve bu komşulardan birini rassal seçer. Seçilen bu komşu benzer örnek oluşturmak için kullanılır. Örnek ile rassal seçilen komşu arasındaki mesafe uzaklık metriği kullanılarak bulunur 0 ile 1 arasındaki rassal bir değer ile çarpılarak yeni sentetik örnek oluşturulur. Bilgi sızmasını önlemek için test veya doğrulama verilerine uygulanmamalıdır. SMOTE ve az örnekleme birlikte kullanıldığında daha başarılı sonuçlar elde edilebilmektedir (Chawla ve diğ. 2002).

4.11.2 Rassal Az Örnekleme

Rassal az örnekleme verisetindeki dengesiz sınıf dağılımı problemini çözmek için nadiren tercih edilen yöntemlerden biridir. Çok tercih edilmemesinin sebebi veri silerken değerli bilgi kaybı riski olmasıdır. Sınıf dağılımını dengelemek için azınlık sınıfın örnek sayısıyla eşit olana kadar veya belirli azınlık çoğunluk oranı sağlanana kadar çoğunluk sınıftan örnekler rassal olarak silinir.

4.12 Hiperparametre Optimizasyonu

Hiperparametre, eğitim sürecinden önce ayarlanabilen, makine öğrenmesi algoritmasının işleyişini belirleyen tüm parametreleri ifade eder (Ippolito 2022). Her makine öğrenmesi modelinin ayarlanabilen farklı hiperparametreleri bulunmaktadır. Makine öğrenmesi algoritmaları için en uygun hiperparametrelerin ayarlanması genellikle modelin verisetine uygulanmasından daha fazla zaman gerektirir (Jin 2022). Hiperparametre optimizasyonu için rassal arama, ızgara arama, Bayes optimizasyonu gibi çeşitli yöntemler bulunmaktadır.

4.12.1 Izgara Arama

Izgara arama, hiperparametrelerin en iyi kombinasyonunu bulmak için deneme-yanılma yöntemi yerine hiperparametre uzayının tüm kombinasyonlarının denendiği bir yöntemdir. Hiperparametrelerin alabilecekleri değerler için aralıklar belirlenip belirlenen aralıklardan ana noktalar seçilerek hiperparametreler için değer listeleri oluşturulur daha sonra tüm değerlerin kombinasyonları için sonuçlar gözlenir.

5. UYGULAMA SONUÇLARI

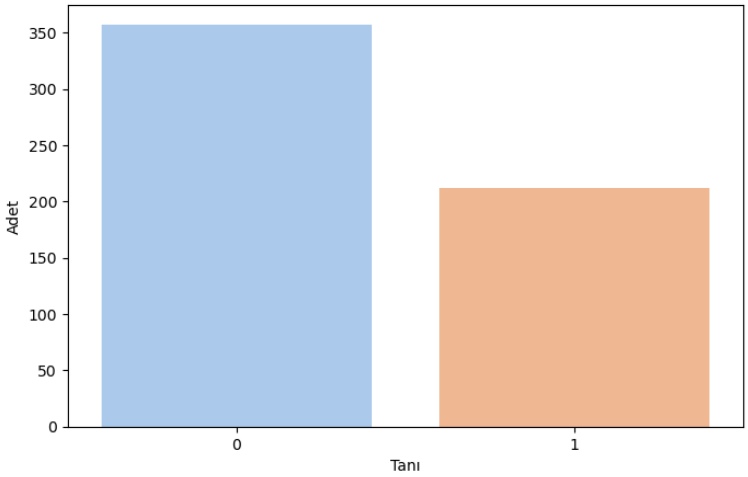
5.1 Veri Analizi

Üzerinde çalışılan Wisconsin Meme Kanseri (Teşhis) veriseti UCI Makine Öğrenmesi Deposu'ndan temin edilmiş olup 569 veri noktası içermektedir. Sınıf dağılımı 357 iyi huylu ve 212 kötü huylu şeklindedir ve özellik için 32 sütun içerir. Bir meme kitlesinin ince iğne aspiratının (FNA) sayısallaştırılmış görüntüsü özellikleri hesaplamak için kullanılmıştır. Bunlar resimde gösterilen hücre çekirdeğinin özelliklerini tanımlarlar. Özellikler tanımlayıcı numarası, tanı (M = kötü huylu, B = iyi huylu), yarıçap (hücre çekirdeğinin merkezinden çevre üzerindeki noktalara olan mesafelerin ortalaması), doku (gri ölçekli değerlerin standart sapma değeri, heterojenlik), çevre, alan, pürüzsüzlük (yarıçap uzunluklarında yerel varyasyon), kompaktlık (çevrenin karesi / alan - 1.0), konkavlık (konturun konkav kısımlarının şiddeti), konkav noktalar (konturun konkav kısımlarının sayısı), simetri, fraktal boyut ("kıyı şeridi yaklaşımı" - 1, şekil düzensizliği) gibi bilgileri içerir. Verisetinde hücre çekirdeğine ait aslen 10 özellik ve bunlara ek olarak her birinin ortalama, standart hata, en kötü değerleriyle birlikte toplam 30 özellik bulunmaktadır. Şekil 5.1'de verisetindeki özelliklere ait bilgiler içeren veriseti açıklaması bulunmaktadır.

	count	mean	std	min	25%	50%	75%	max
radius_mean	569.0	14.127292	3.524049	6.981000	11.700000	13.370000	15.780000	28.11000
texture_mean	569.0	19.289649	4.301036	9.710000	16.170000	18.840000	21.800000	39.28000
perimeter_mean	569.0	91.969033	24.298981	43.790000	75.170000	86.240000	104.100000	188.50000
area_mean	569.0	654.889104	351.914129	143.500000	420.300000	551.100000	782.700000	2501.00000
smoothness_mean	569.0	0.096360	0.014064	0.052630	0.086370	0.095870	0.105300	0.16340
compactness_mean	569.0	0.104341	0.052813	0.019380	0.064920	0.092630	0.130400	0.34540
concavity_mean	569.0	0.088799	0.079720	0.000000	0.029560	0.061540	0.130700	0.42680
concave points_mean	569.0	0.048919	0.038803	0.000000	0.020310	0.033500	0.074000	0.20120
symmetry_mean	569.0	0.181162	0.027414	0.106000	0.161900	0.179200	0.195700	0.30400
fractal_dimension_mean	569.0	0.062798	0.007060	0.049960	0.057700	0.061540	0.066120	0.09744
radius_se	569.0	0.405172	0.277313	0.111500	0.232400	0.324200	0.478900	2.87300
texture_se	569.0	1.216853	0.551648	0.360200	0.833900	1.108000	1.474000	4.88500
perimeter_se	569.0	2.866059	2.021855	0.757000	1.606000	2.287000	3.357000	21.98000
area_se	569.0	40.337079	45.491006	6.802000	17.850000	24.530000	45.190000	542.20000
smoothness_se	569.0	0.007041	0.003003	0.001713	0.005169	0.006380	0.008146	0.03113
compactness_se	569.0	0.025478	0.017908	0.002252	0.013080	0.020450	0.032450	0.13540
concavity_se	569.0	0.031894	0.030186	0.000000	0.015090	0.025890	0.042050	0.39600
concave points_se	569.0	0.011796	0.006170	0.000000	0.007638	0.010930	0.014710	0.05279
symmetry_se	569.0	0.020542	0.008266	0.007882	0.015160	0.018730	0.023480	0.07895
fractal_dimension_se	569.0	0.003795	0.002646	0.000895	0.002248	0.003187	0.004558	0.02984
radius_worst	569.0	16.269190	4.833242	7.930000	13.010000	14.970000	18.790000	36.04000
texture_worst	569.0	25.677223	6.146258	12.020000	21.080000	25.410000	29.720000	49.54000
perimeter_worst	569.0	107.261213	33.602542	50.410000	84.110000	97.660000	125.400000	251.20000
area_worst	569.0	880.583128	569.356993	185.200000	515.300000	686.500000	1084.000000	4254.00000
smoothness_worst	569.0	0.132369	0.022832	0.071170	0.116600	0.131300	0.146000	0.22260
compactness_worst	569.0	0.254265	0.157336	0.027290	0.147200	0.211900	0.339100	1.05800
concavity_worst	569.0	0.272188	0.208624	0.000000	0.114500	0.226700	0.382900	1.25200
concave points_worst	569.0	0.114606	0.065732	0.000000	0.064930	0.099930	0.161400	0.29100
symmetry_worst	569.0	0.290076	0.061867	0.156500	0.250400	0.282200	0.317900	0.66380
fractal_dimension_worst	569.0	0.083946	0.018061	0.055040	0.071460	0.080040	0.092080	0.20750

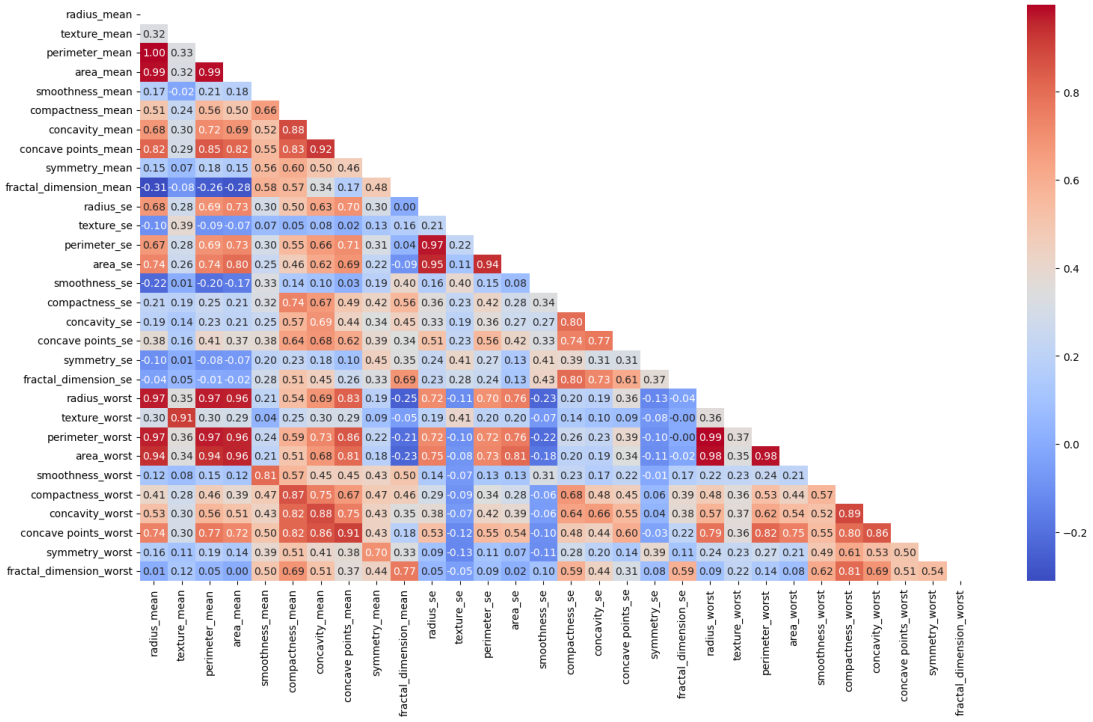
Şekil 5.1: Veriseti Açıklaması.

Tanımlayıcı numara olan id sütunu sınıflandırmada kullanılmayacağından verisetinden çıkarılmıştır. Yinelenen veya eksik veri bulunmamaktadır. Tahmin edilecek hedef sütun yani tanı sütunu M ve B değerlerinden oluşmaktadır. Hedef sütun değerleri M (kötü huylu) 1, B (iyi huylu) 0 olarak eşlenerek kategorik değerden nümerik değere çevrilmiştir. Şekil 5.2'de verisetindeki sınıf dağılımları yer almaktadır.



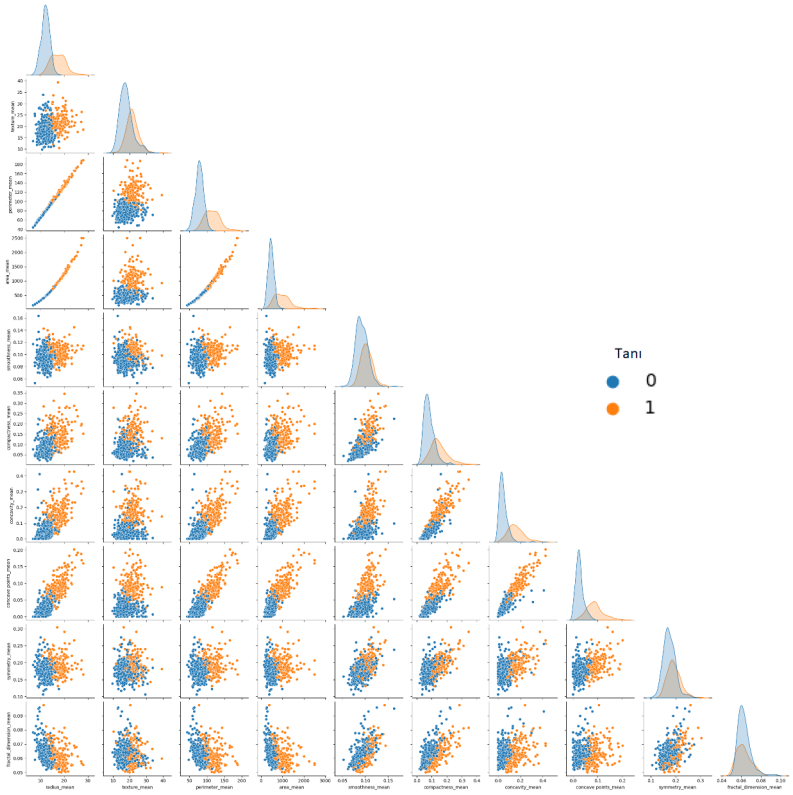
Şekil 5.2: Sınıf Dağılımı.

Özelliklerin birbirleriyle olan ilişkisini gösteren korelasyon ısı haritası Şekil 5.3'te bulunmaktadır. Pozitif değer aynı yönde, negatif değer ters yönde ilişkiyi temsil eder. Koyu renkler yüksek korelasyonu gösterir. Isı haritası özellikler arasındaki ilişkileri hızlı bir şekilde belirlenmesine yardımcı olabilir. Genellikle pozitif korelasyonlar sıcak renklerle, negatif korelasyonlar soğuk renklerle temsil edilir. Bağımsız değişkenler arasındaki yüksek korelasyonun belirlenmesinde, değişkenler arasındaki ilişkilerin anlaşılmasında kullanılır.



Şekil 5.3: Korelasyon Isı Haritası.

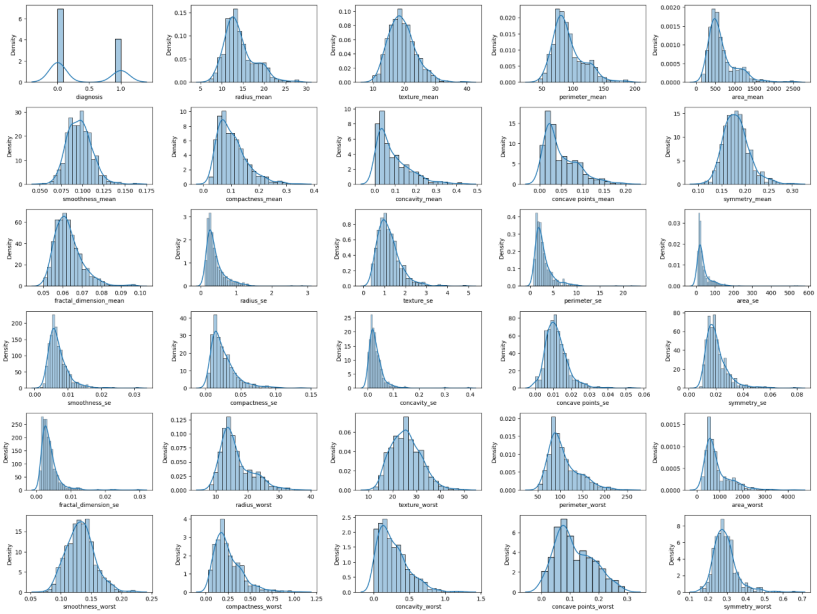
1 tam pozitif korelasyonu, -1 tam negatif korelasyonu, 0 ise iki özellik arasında korelasyon bulunmadığını gösterir.



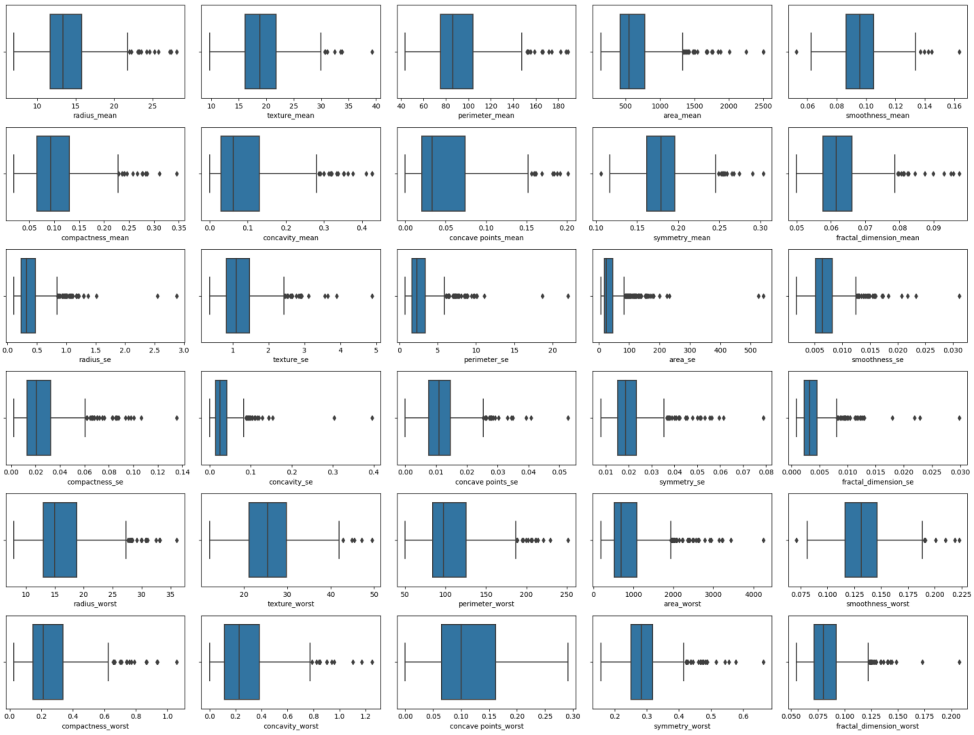
Şekil 5.4: Ortalama Özelliklerin İkili İlişkileri.

Şekil 5.4 verisetindeki ortalama özelliklerin ikili ilişkilerini göstermektedir. Bazı özelliklerin yüksek değerleri kötü huylu olma eğilimi gösterirken bazıları için böyle bir eğilim yoktur.

Verinin merkezi eğilimi, yayılımı, özelliklerin veriseti içinde ne tür bir dağılım sergilediği, değerlerin frekansı gibi bilgiler Şekil 5.5'teki gibi histogram kullanılarak gözlemlenebilir.

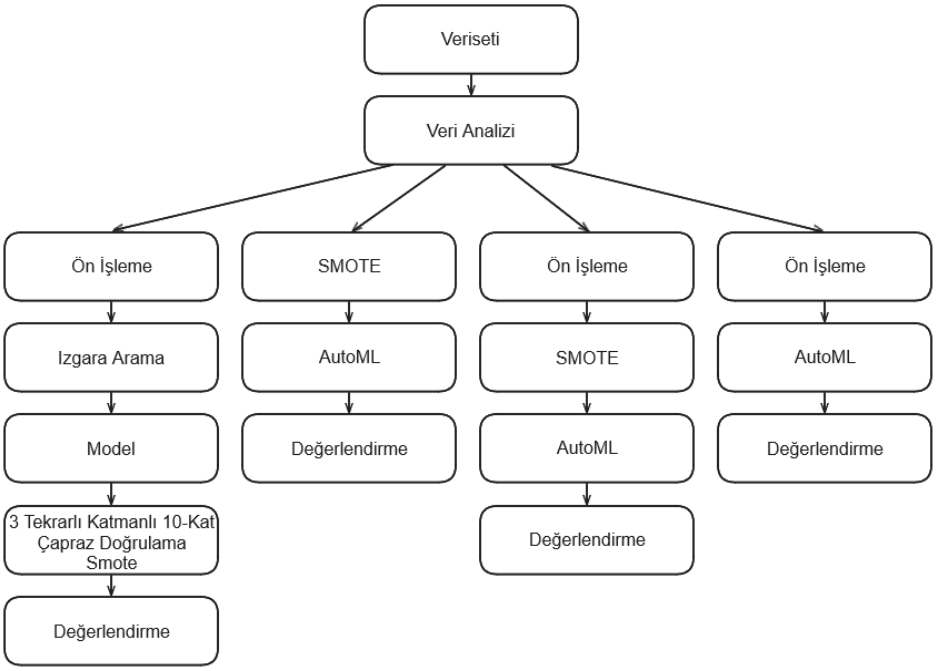


Şekil 5.5: Histogram Grafiği.



Şekil 5.6: Kutu Grafiği.

Kutu grafiği ile özellik değerlerinin minimum, birinci kartil, medyan, üçüncü kartil ve maksimum istatistikleri görülebilir. Aykırı değerleri tespit etmek için kullanışlıdır. Şekil 5.6'da verisetindeki özelliklere ait kutu grafikleri bulunmaktadır. Bu çalışmada kullanılan sınıflandırma iş akışı Şekil 5.7'de bulunmaktadır.



Şekil 5.7: Bu çalışmada kullanılan sınıflandırma iş akışı.

5.2 Sınıflandırıcılar ve Sonuçları

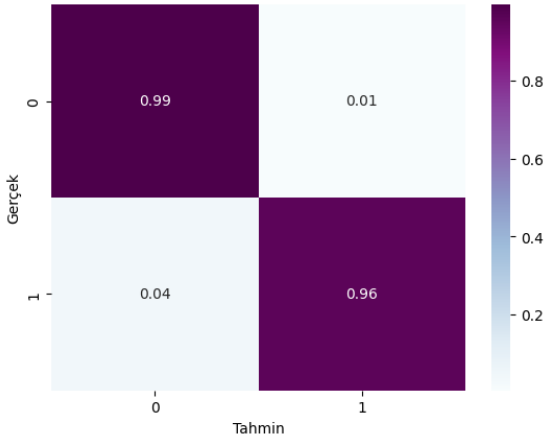
5.2.1 Lojistik Regresyon

Varsayılan Lojistik Regresyon parametreleri: LogisticRegression(penalty='l2', *, dual=False, tol=0.0001, C=1.0, fit_intercept=True, intercept_scaling=1, class_weight=None, random_state=None, solver='lbfgs', max_iter=100, multi_class='auto', verbose=0, warm_start=False, n_jobs=None, l1_ratio=None)

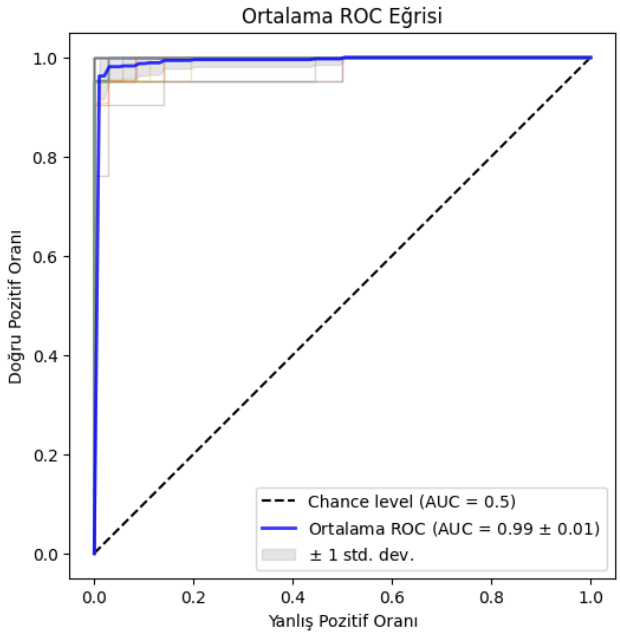
Standardizasyon, SMOTE ve hiperparametre optimizasyonu ile tekrarlı katmanlı k-kat çapraz doğrulama için 3 tekrar 10-kat kullanılarak 0.9801 doğruluk, 0.9559 duyarlılık, 0.9907 hassasiyet, 0.9725 F1-skor elde edilmiştir.

Ayarlanan parametreler: LogisticRegression(C=0.1, penalty='l2', 'solver='liblinear')

Şekil 5.8'de %99 gerçek pozitif, %96 gerçek negatif başarıyla tahmin edilmiştir. Şekil 5.9'da katlardaki ve ortalama ROC eğrisi bulunmaktadır. 0.99 AUC elde edilmiştir. Mavi eğri ortalama ROC eğrisi iken diğer eğriler her kattaki ROC eğrileridir. ROC eğrisi y ekseninde doğru pozitif oran, x ekseninde yanlış pozitif oran olmak üzere iki parametreye bağlı bir olasılık eğrisi grafiğidir. AUC değeri ayrılabilirlik derecesini gösterir. Eğri altında kalan alan yani AUC ne kadar büyük olursa modelin performansı o kadar iyi anlamına gelir. 0.99 değeri, rastgele örneklenen pozitif sınıfın rastgele örneklenen negatif sınıftan daha yüksek tahmin olasılığına sahip olma olasılığını gösterir.



Şekil 5.8: Lojistik Regresyon Karışıklık Matrisi.



Şekil 5.9: Lojistik Regresyon ROC Eğrisi AUC.

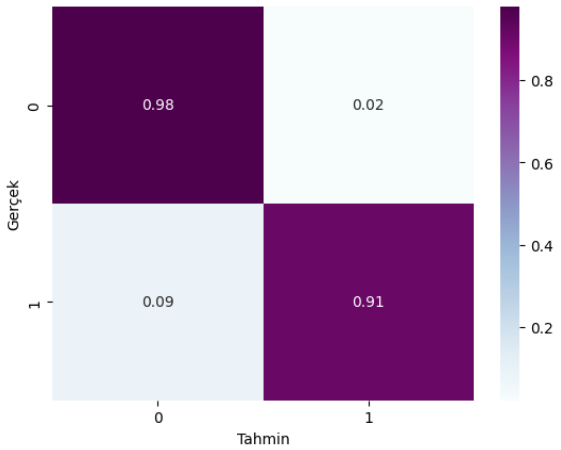
5.2.2 Karar Ağacı

Varsayılan Karar Ağacı parametreleri: DecisionTreeClassifier(*, criterion='gini', splitter='best', max_depth=None, min_samples_split=2, min_samples_leaf=1, min_weight_fraction_leaf=0.0, max_features=None, random_state=None, max_leaf_nodes=None, min_impurity_decrease=0.0, class_weight=None, ccp_alpha=0.0)

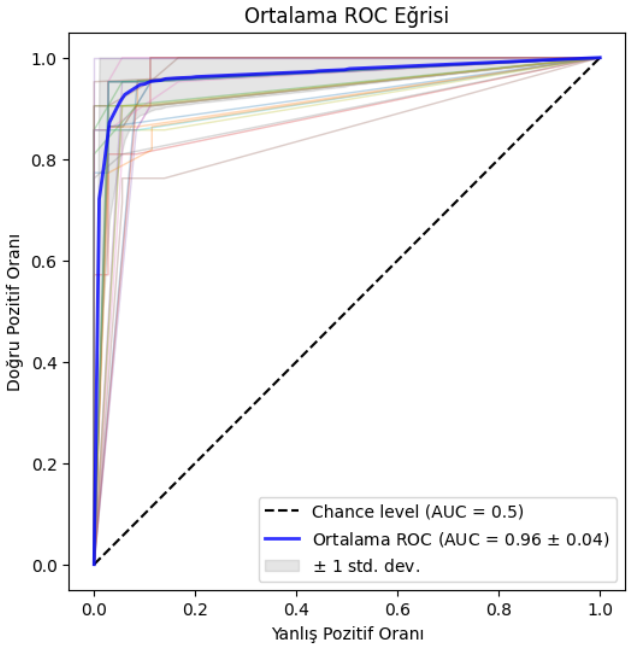
Standardizasyon, SMOTE ve hiperparametre optimizasyonu ile tekrarlı katmanlı k-kat çapraz doğrulama için 3 tekrar 10-kat kullanılarak 0.9256 doğruluk, 0.8680 duyarlılık, 0.9320 hassasiyet, 0.8955 F1-skor elde edilmiştir.

Ayarlanan parametreler: DecisionTreeClassifier(criterion='gini', max_depth=None, min_samples_leaf=4, min_samples_split=5)

Şekil 5.10'da %98 gerçek pozitif, %91 gerçek negatif başarıyla tahmin edilmiştir. Şekil 5.11'de katlardaki ve ortalama ROC eğrisi bulunmaktadır. 0.96 AUC elde edilmiştir.



Şekil 5.10: Karar Ağacı Karışıklık Matrisi.



Şekil 5.11: Karar Ağacı ROC Eğrisi AUC.

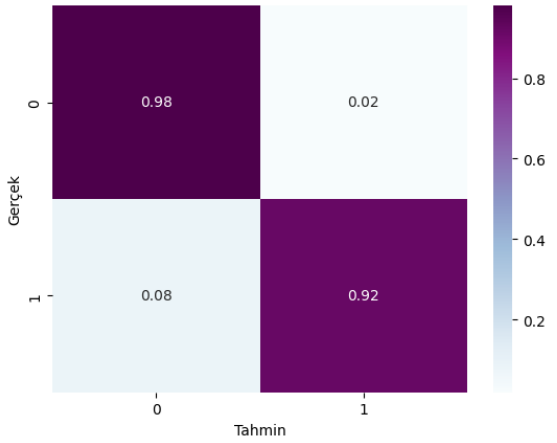
5.2.3 K-En Yakın Komşu

Varsayılan KNN parametreleri: KNeighborsClassifier(n_neighbors=5, *, weights='uniform', algorithm='auto', leaf_size=30, p=2, metric='minkowski', metric_params=None, n_jobs=None)

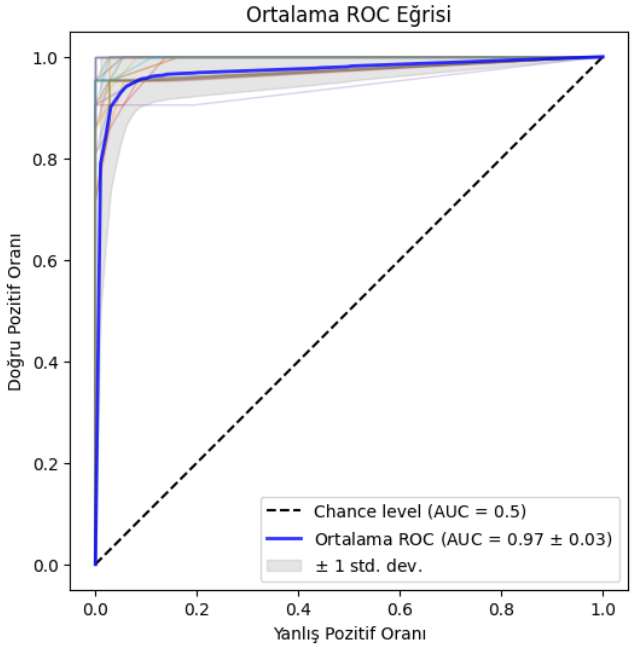
Standardizasyon, SMOTE ve hiperparametre optimizasyonu ile tekrarlı katmanlı k-kat çapraz doğrulama için 3 tekrar 10-kat kullanılarak 0.9672 doğruluk, 0.9339 duyarlılık, 0.9781 hassasiyet, 0.9545 F1-skor elde edilmiştir.

Ayarlanan parametreler: KNeighborsClassifier(leaf_size=10, n_neighbors=5, p=1, weights='uniform')

Şekil 5.12'de %98 gerçek pozitif, %92 gerçek negatif başarıyla tahmin edilmiştir. Şekil 5.13'te katlardaki ve ortalama ROC eğrisi bulunmaktadır. 0.97 AUC elde edilmiştir.



Şekil 5.12: KNN Karışıklık Matrisi.



Şekil 5.13: KNN ROC Eğrisi AUC.

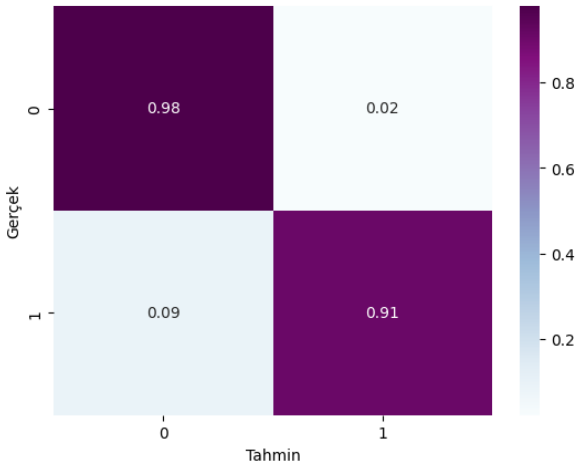
5.2.4 Naive Bayes

Varsayılan Gaussian Naive Bayes parametreleri: GaussianNB(*, priors=None, var_smoothing=1e-09)

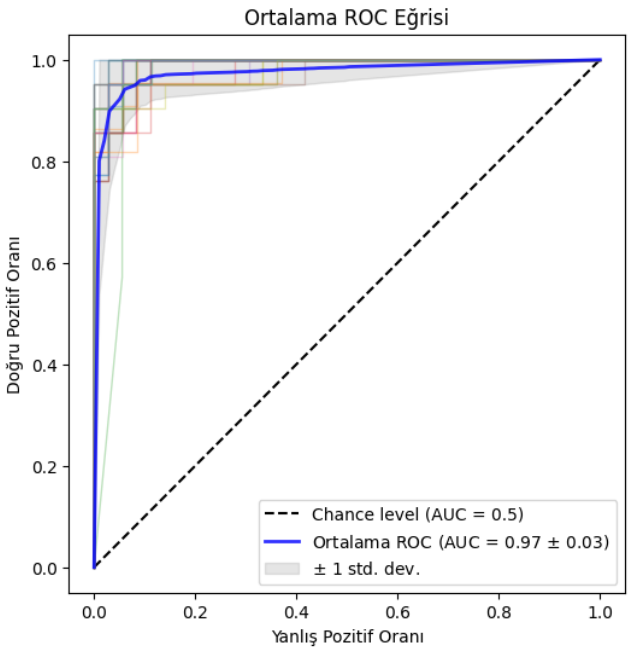
Standardizasyon, SMOTE ve hiperparametre optimizasyonu ile tekrarlı katmanlı k-kat çapraz doğrulama için 3 tekrar 10-kat kullanılarak 0.9379 doğruluk, 0.8791 duyarlılık, 0.9532 hassasiyet, 0.9129 F1-skor elde edilmiştir.

Ayarlanan parametreler: GaussianNB(var_smoothing=0.1)

Şekil 5.14'te %98 gerçek pozitif, %91 gerçek negatif başarıyla tahmin edilmiştir. Şekil 5.15'te katlardaki ve ortalama ROC eğrisi bulunmaktadır. 0.97 AUC elde edilmiştir.



Şekil 5.14: Naive Bayes Karışıklık Matrisi.



Şekil 5.15: Naive Bayes ROC Eğrisi AUC.

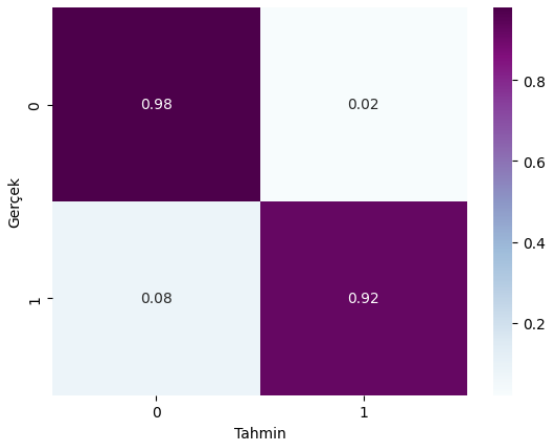
5.2.5 Destek Vektör Makinesi

Varsayılan SVM parametreleri: SVC(*, C=1.0, kernel='rbf', degree=3, gamma='scale', coef0=0.0, shrinking=True, probability=False, tol=0.001, cache_size=200, class_weight=None, verbose=False, max_iter=-1, decision_function_shape='ovr', break_ties=False, random_state=None)

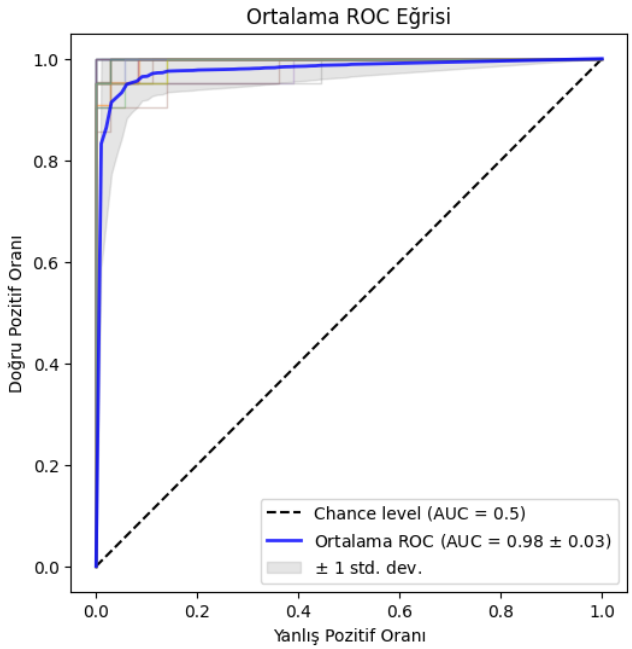
Standardizasyon, SMOTE ve hiperparametre optimizasyonu ile tekrarlı katmanlı k-kat çapraz doğrulama için 3 tekrar 10-kat kullanılarak 0.9754 doğruluk, 0.9608 duyarlılık, 0.9745 hassasiyet, 0.9669 F1-skor elde edilmiştir.

Ayarlanan parametreler: SVC(C=1, gamma='scale', kernel='rbf')

Şekil 5.16'da %98 gerçek pozitif, %92 gerçek negatif başarıyla tahmin edilmiştir. Şekil 5.17'de katlardaki ve ortalama ROC eğrisi bulunmaktadır. 0.98 AUC elde edilmiştir.



Şekil 5.16: SVM Karışıklık Matrisi.



Şekil 5.17: SVM ROC Eğrisi AUC.

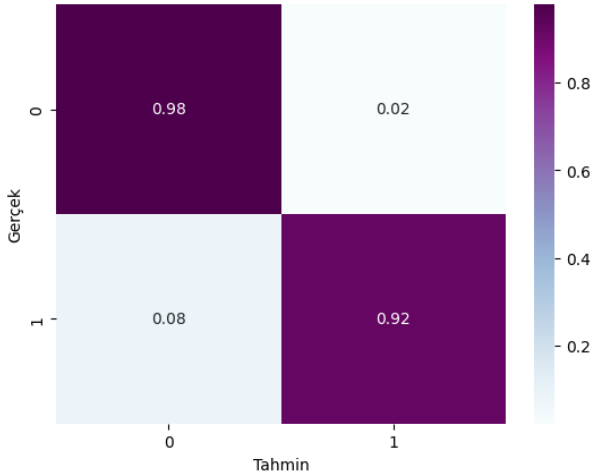
5.2.6 Rassal Orman

Varsayılan Rassal Orman parametreleri: RandomForestClassifier (n_estimators=100, *, criterion='gini', max_depth=None, min_samples_split=2, min_samples_leaf=1, min_weight_fraction_leaf=0.0, max_features='sqrt', max_leaf_nodes=None, min_impurity_decrease=0.0, bootstrap=True, oob_score=False, n_jobs=None, random_state=None, verbose=0, warm_start=False, class_weight=None, ccp_alpha=0.0, max_samples=None

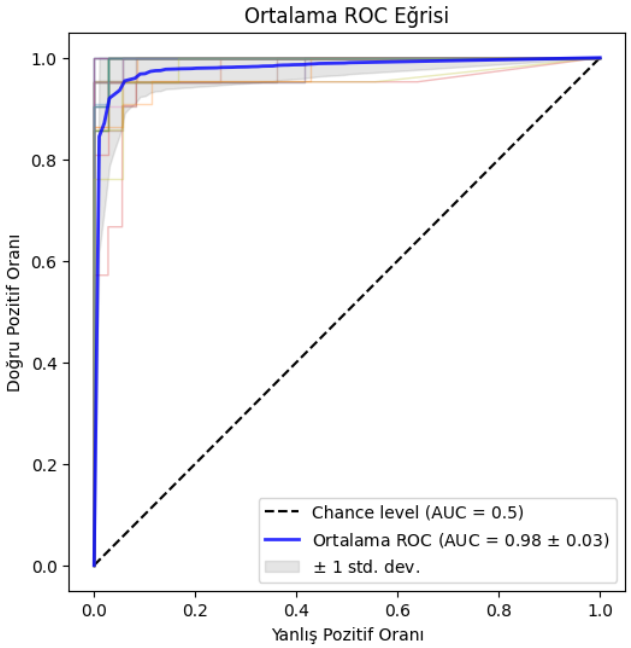
Standardizasyon, SMOTE ve hiperparametre optimizasyonu ile tekrarlı katmanlı k-kat çapraz doğrulama için 3 tekrar 10-kat kullanılarak 0.9614 doğruluk, 0.9339 duyarlılık, 0.9624 hassasiyet, 0.9460 F1-skor elde edilmiştir.

Ayarlanan parametreler: RandomForestClassifier(criterion='entropy', min_samples_leaf=4, min_samples_split=2, n_estimators=50)

Şekil 5.18'de %98 gerçek pozitif, %92 gerçek negatif başarıyla tahmin edilmiştir. Şekil 5.19'da katlardaki ve ortalama ROC eğrisi bulunmaktadır. 0.98 AUC elde edilmiştir.



Şekil 5.18: Rassal Orman Karışıklık Matrisi.



Şekil 5.19: Rassal Orman ROC Eğrisi AUC.

5.2.7 Stokastik Gradyan İniş

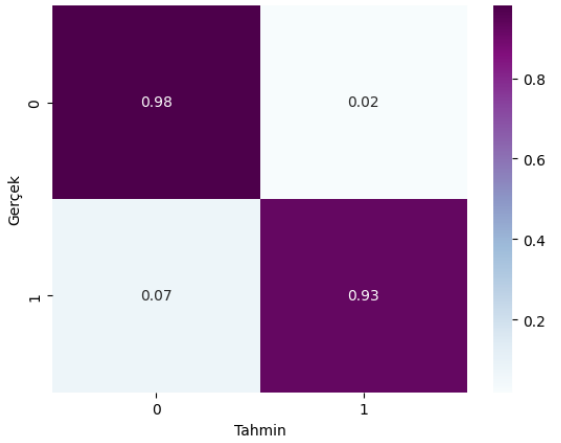
Varsayılan SGD parametreleri: SGDClassifier(loss='hinge', *, penalty='l2', alpha=0.0001, l1_ratio=0.15, fit_intercept=True, max_iter=1000, tol=0.001, shuffle=True, verbose=0, epsilon=0.1, n_jobs=None, random_state=None, learning_rate='optimal', eta0=0.0, power_t=0.5, early_stopping=False,

validation_fraction=0.1, n_iter_no_change=5, class_weight=None, warm_start=False, average=False)

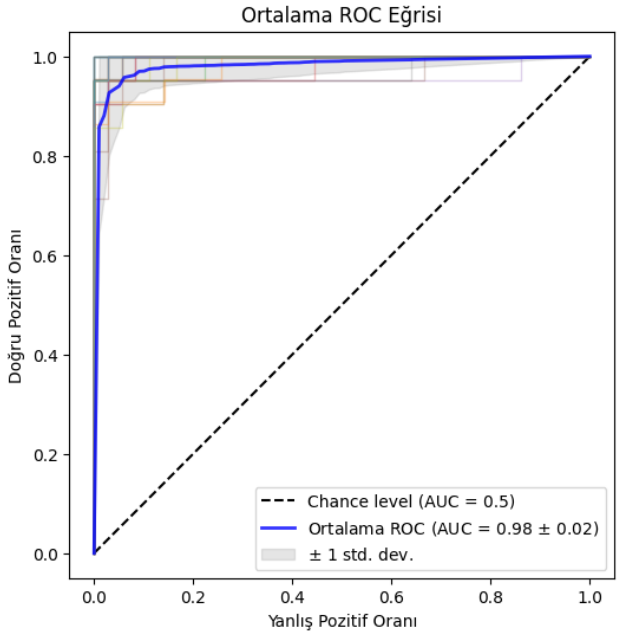
Standardizasyon, SMOTE ve hiperparametre optimizasyonu ile tekrarlı katmanlı k-kat çapraz doğrulama için 3 tekrar 10-kat kullanılarak 0.9666 doğruluk, 0.9354 duyarlılık, 0.9748 hassasiyet, 0.9535 F1-skor elde edilmiştir.

Ayarlanan parametreler: SGDClassifier(alpha=0.01, fit_intercept=True, loss='hinge', penalty='l1', power_t=0.5)

Şekil 5.20'de %98 gerçek pozitif, %93 gerçek negatif başarıyla tahmin edilmiştir. Şekil 5.21'de katlardaki ve ortalama ROC eğrisi bulunmaktadır. 0.99 AUC elde edilmiştir.



Şekil 5.20: SGD Karışıklık Matrisi.



Şekil 5.21: SGD ROC Eğrisi AUC.

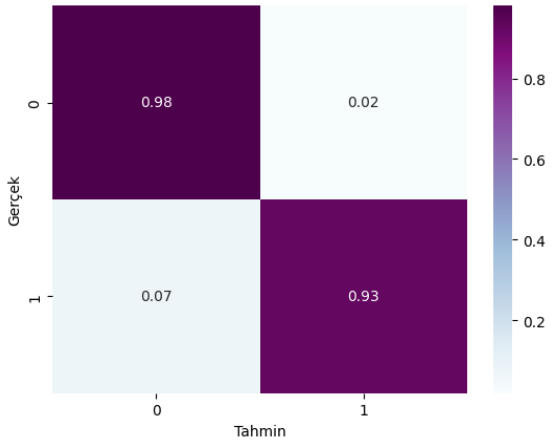
5.2.8 AdaBoost

Varsayılan AdaBoost parametreleri: AdaBoostClassifier(estimator=None, *, n_estimators=50, learning_rate=1.0, algorithm='SAMME.R', random_state=None, base_estimator='deprecated')

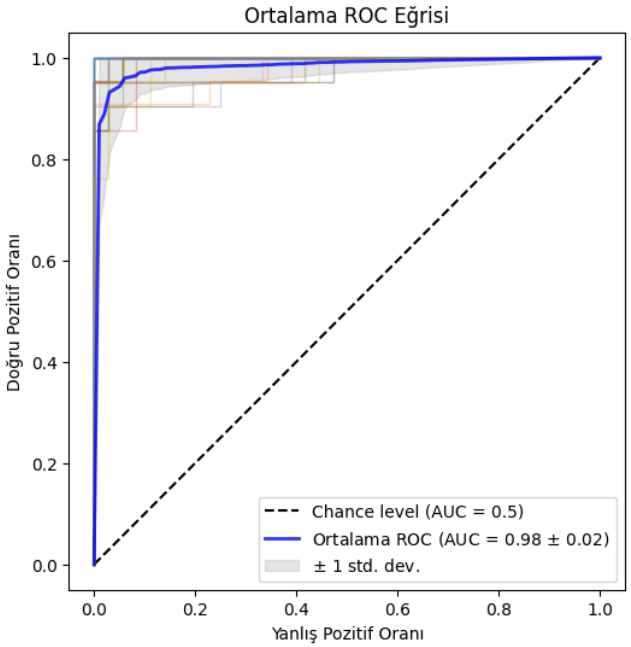
Standardizasyon, SMOTE ve hiperparametre optimizasyonu ile tekrarlı katmanlı k-kat çapraz doğrulama için 3 tekrar 10-kat kullanılarak 0.9655 doğruluk, 0.9370 duyarlılık, 0.9713 hassasiyet, 0.9524 F1-skor elde edilmiştir.

Ayarlanan parametreler: AdaBoostClassifier(learning_rate=1.0, n_estimators=150)

Şekil 5.22'de %98 gerçek pozitif, %93 gerçek negatif başarıyla tahmin edilmiştir. Şekil 5.23'te katlardaki ve ortalama ROC eğrisi bulunmaktadır. 0.98 AUC elde edilmiştir.



Şekil 5.22: AdaBoost Karışıklık Matrisi.



Şekil 5.23: AdaBoost ROC Eğrisi AUC.

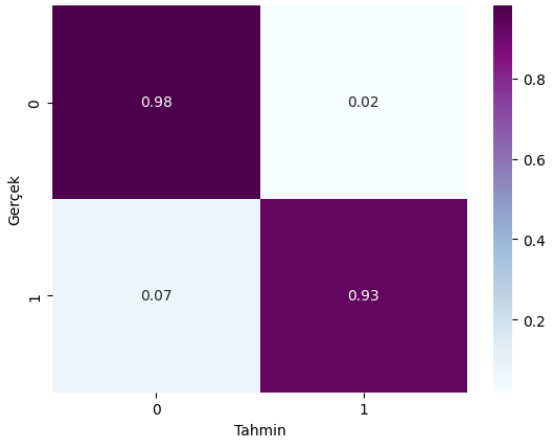
5.2.9 XGBoost

XGBoost çok sayıda parametreye sahiptir. Bu çalışmada değerlendirilen parametrelerin varsayılan değerleri: XGBClassifier(max_depth=6, min_child_weight=1, n_estimators=100)

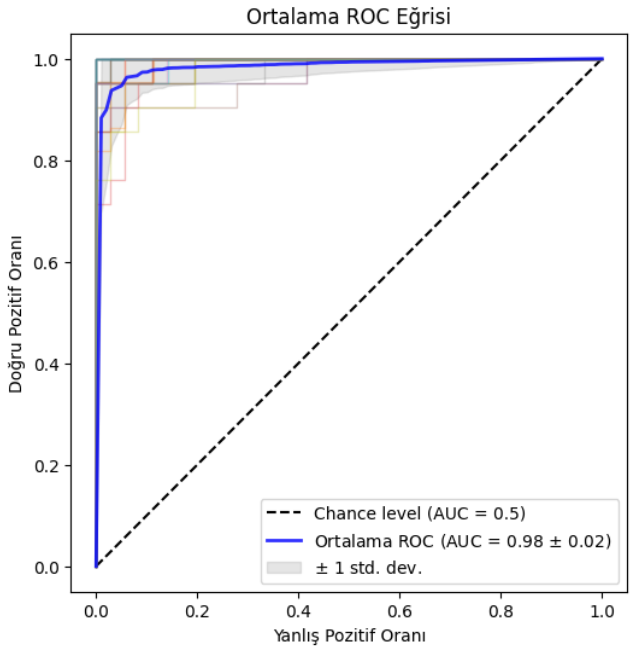
Standardizasyon, SMOTE ve hiperparametre optimizasyonu ile tekrarlı katmanlı k-kat çapraz doğrulama için 3 tekrar 10-kat kullanılarak 0.9631 doğruluk, 0.9354 duyarlılık, 0.9656 hassasiyet, 0.9489 F1-skor elde edilmiştir.

Ayarlanan parametreler: XGBClassifier(max_depth=1, min_child_weight=1, n_estimators=100)

Şekil 5.24'te %98 gerçek pozitif, %93 gerçek negatif başarıyla tahmin edilmiştir. Şekil 5.25'te katlardaki ve ortalama ROC eğrisi bulunmaktadır. 0.98 AUC elde edilmiştir.



Şekil 5.24: XGBoost Karışıklık Matrisi.



Şekil 5.25: XGBoost ROC Eğrisi AUC.

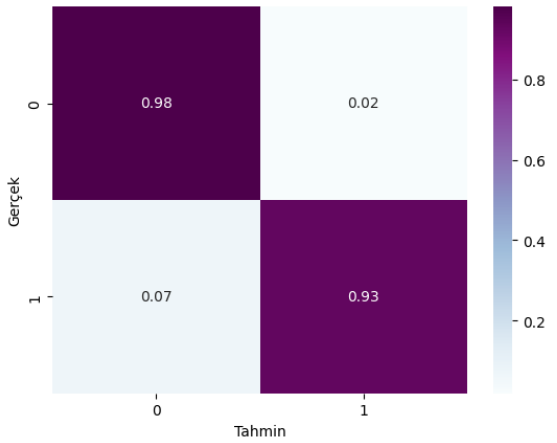
5.2.10 LightGBM

LightGBM çok sayıda parametreye sahiptir. Bu çalışmada değerlendirilen parametrelerin varsayılan değerleri: LGBMClassifier(learning_rate=0.1, num_leaves=31, max_depth=-1, n_estimators=100)

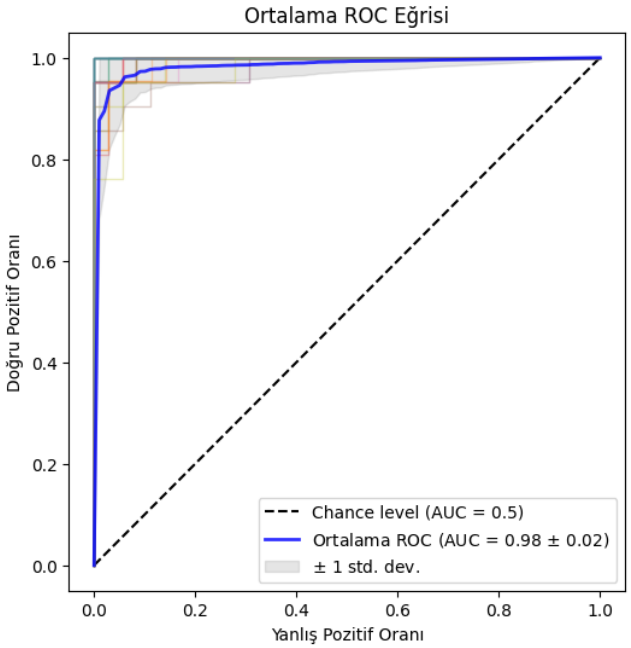
Standardizasyon, SMOTE ve hiperparametre optimizasyonu ile tekrarlı katmanlı k-kat çapraz doğrulama için 3 tekrar 10-kat kullanılarak 0.9672 doğruluk, 0.9386 duyarlılık, 0.9740 hassasiyet, 0.9543 F1-skor elde edilmiştir.

Ayarlanan parametreler: LGBMClassifier(learning_rate=0.1, max_depth=2, n_estimators=500, num_leaves=20)

Şekil 5.26'da %98 gerçek pozitif, %93 gerçek negatif başarıyla tahmin edilmiştir. Şekil 5.27'de katlardaki ve ortalama ROC eğrisi bulunmaktadır. 0.98 AUC elde edilmiştir.



Şekil 5.26: LightGBM Karışıklık Matrisi.



Şekil 5.27: LightGBM ROC Eğrisi AUC.

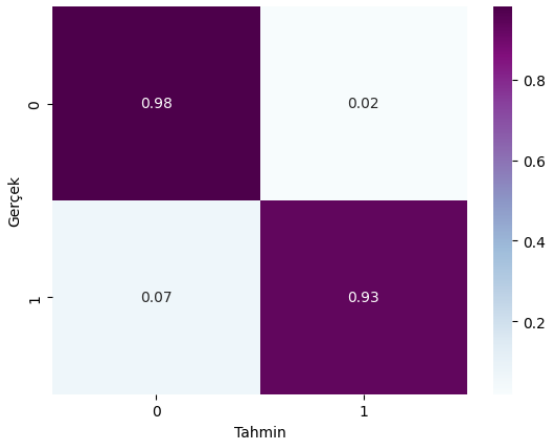
5.2.11 Yapay Sinir Ağları

Varsayılan ANN parametreleri: MLPClassifier(hidden_layer_sizes=(100,), activation='relu', *, solver='adam', alpha=0.0001, batch_size='auto', learning_rate='constant', learning_rate_init=0.001, power_t=0.5, max_iter=200, shuffle=True, random_state=None, tol=0.0001, verbose=False, warm_start=False, momentum=0.9, nesterovs_momentum=True, early_stopping=False, validation_fraction=0.1, beta_1=0.9, beta_2=0.999, epsilon=1e-08, n_iter_no_change=10, max_fun=15000)

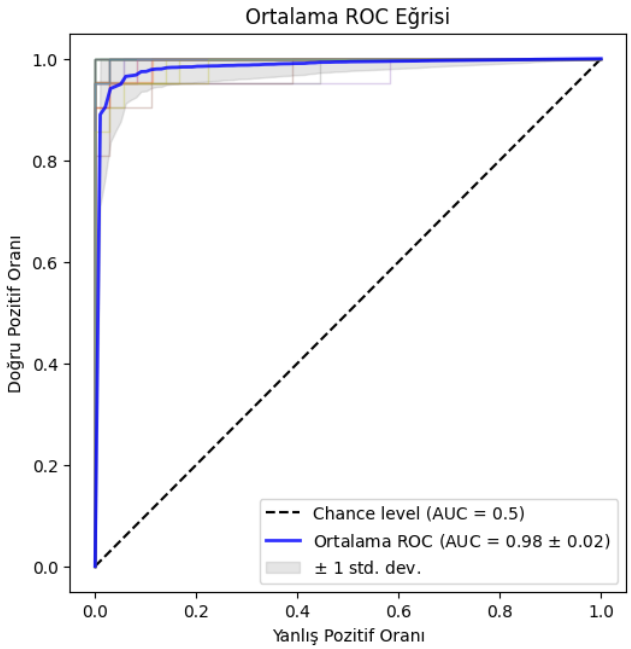
Standardizasyon, SMOTE ve hiperparametre optimizasyonu ile tekrarlı katmanlı k-kat çapraz doğrulama için 3 tekrar 10-kat kullanılarak 0.9760 doğruluk, 0.9605 duyarlılık, 0.9760 hassasiyet, 0.9673 F1-skor elde edilmiştir.

Ayarlanan parametreler: MLPClassifier(activation='relu', alpha=0.0001, hidden_layer_sizes=(20,), learning_rate='constant', solver='adam')

Şekil 5.28'de %98 gerçek pozitif, %93 gerçek negatif başarıyla tahmin edilmiştir. Şekil 5.29'da katlardaki ve ortalama ROC eğrisi bulunmaktadır. 0.98 AUC elde edilmiştir.



Şekil 5.28: ANN Karışıklık Matrisi.



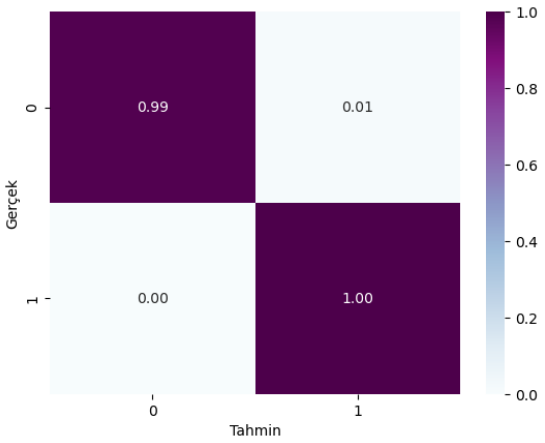
Şekil 5.29: ANN ROC Eğrisi AUC.

5.2.12 TPOT

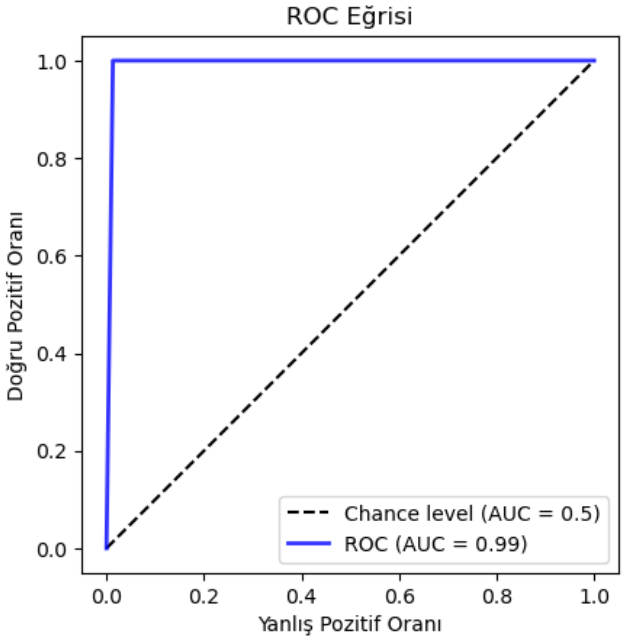
Jenerasyon sayısı 100 popülasyon büyüklüğü 100 ayarlı, herhangi bir sınırlama koyulmadan, veriseti %80 eğitim %20 test için ayrılarak TPOT ile en başarılı kombinasyon standardizasyon ve stokastik gradyan iniş algoritması seçilmiştir. 0.9912 doğruluk, 1.0 duyarlılık, 0.9744 hassasiyet, 0.9870 F1-skor elde edilmiştir.

Parametreler: SGDClassifier(alpha=0.001, eta0=0.01, fit_intercept=False, l1_ratio=0.0, learning_rate="invscaling", loss="hinge", penalty="elasticnet", power_t=0.0)

Şekil 5.30'da %99 gerçek pozitif, %100 gerçek negatif başarıyla tahmin edilmiştir. Şekil 5.31'de TPOT sonucuna ait ROC eğrisi bulunmaktadır. 0.99 AUC elde edilmiştir.



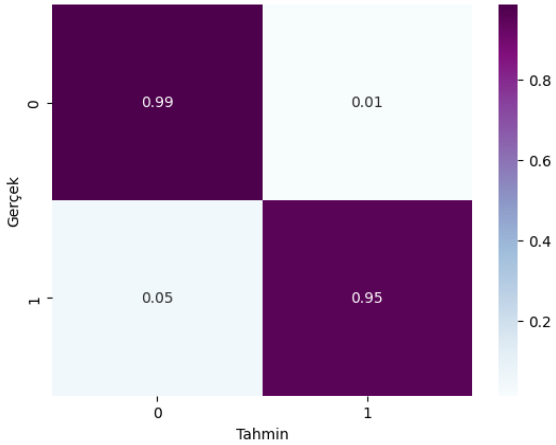
Şekil 5.30: TPOT Karışıklık Matrisi.



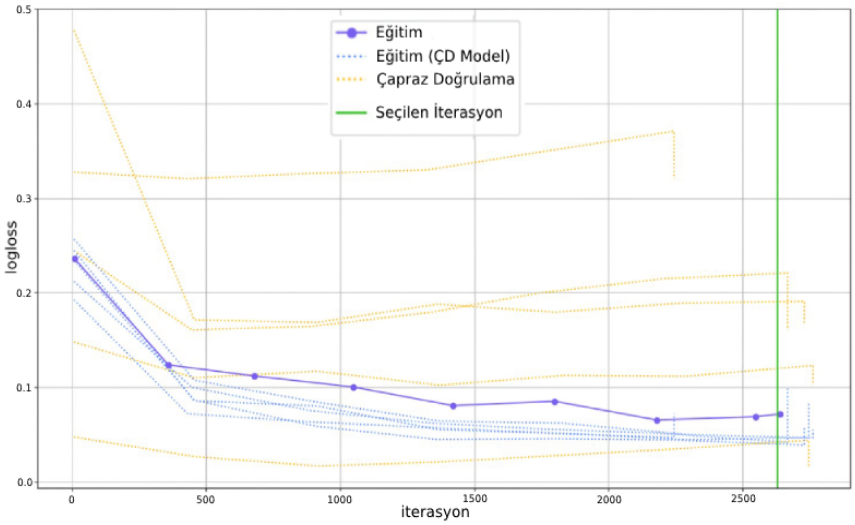
Şekil 5.31: TPOT ROC Eğrisi AUC.

5.2.13 H2O

Herhangi bir sınırlama koyulmadan veriseti %80 eğitim %20 test için ayrılarak H2O tarafından 30 girdi katmanı 20 bırakma, 100 gizli katman ReLU, 2 çıktı katman softmax mimarili derin öğrenme algoritması seçilmiştir. Derin öğrenme ile 0.9871 doğruluk, 0.9816 duyarlılık, 0.9929 hassasiyet, 0.9871 F1-skor elde edilmiştir. Şekil 5.32'de %99 gerçek pozitif, %95 gerçek negatif başarıyla tahmin edilmiştir.



Şekil 5.32: H2O Karışıklık Matrisi.



Şekil 5.33: H2O Öğrenme Eğrisi.

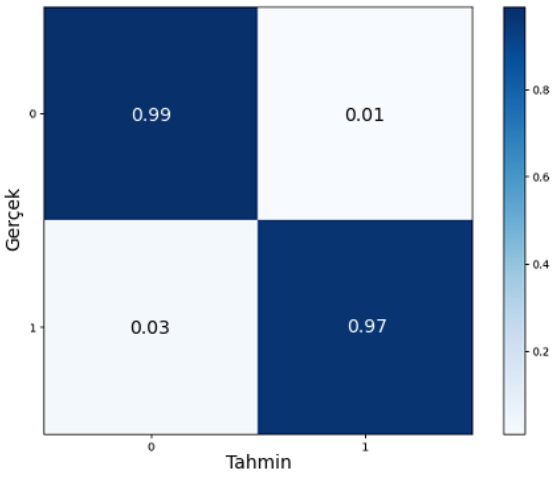
Şekil 5.33'te öğrenme eğrisi bulunmaktadır. Öğrenme eğrisi grafiği eğitim adımlarının hata metrikleri üzerindeki etkisini gösterir. Modelin aşırı uyumunu veya yetersiz uyumunu tespit etmede kullanılabilir. Olması beklenen eğitim ve doğrulama eğrilerinin yakınsamasıdır.

5.2.14 MLJAR

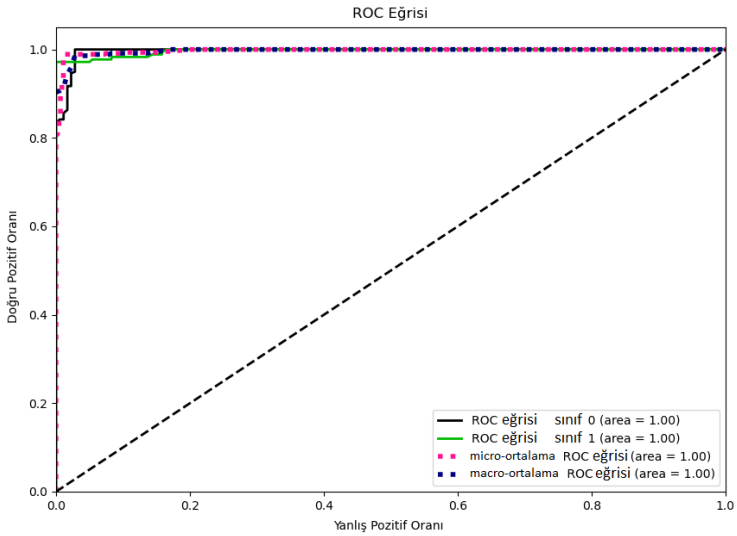
Compete modunda veriseti %80 eğitim %20 test için ayrılarak MLJAR tarafından CatBoost algoritması seçilmiştir. 0.9846 doğruluk, 0.9718 duyarlılık, 0.9885 hassasiyet, 0.9801 F1-skor elde edilmiştir.

Parametreler: CatBoostClassifier(learning_rate=0.2, depth=6, rsm=0.7, n_jobs=-1, loss_function='Logloss', eval_metric='Accuracy')

Şekil 5.34'te %99 gerçek pozitif, %97 gerçek negatif başarıyla tahmin edilmiştir. Şekil 5.35'te sınıfların ve micro, makro ortalama ROC eğrileri verilmiştir. 0.99 AUC elde edilmiştir. Macro ortalama verisetindeki her sınıf için bağımsız olarak hesaplanır ve ortalaması alınır böylelikle tüm sınıfların ağırlığı eşit değerlendirilir. Micro ortalamada tüm sınıflar toplu şekilde ele alınır, sınıflar yerine gözlemlerin ağırlığı eşit değerlendirilir.



Şekil 5.34: MLJAR Karışıklık Matrisi.



Şekil 5.35: MLJAR ROC Eğrisi AUC.

6. SONUÇ VE ÖNERİLER

Bu çalışmada UCI Makine Öğrenmesi Deposu'ndan temin edilen Wisconsin Meme Kanseri (Teşhis) veriseti üzerinde makine öğrenmesi modeli geliştirmeye yönelik iki farklı yaklaşımın karşılaştırmalı sonuçları hedeflenmiştir. Hiperparametre ayarlı ve ön işlemeli makine öğrenmesi algoritmaları ve otomatik makine öğrenmesi araçları çeşitli değerlendirme metrikleri ile karşılaştırılmıştır ve AutoML yöntemlerinin başarısı gözlemlenmiştir. Ön işleme ve hiperparametre optimizasyonu model sonuçlarını genel olarak iyileştirmiştir. Bulgulara göre AutoML yöntemleri bu çalışmada uygulanan diğer makine öğrenmesi algoritmalarının hepsinden sınıflandırma doğruluğu ve F1-skor açısından daha performanslı sonuçlar elde etmiştir. Ayrıca TPOT ile bu çalışmada kullanılan verisetiyle yapılan önceki çalışmalara kıyasla A. J. B ve Palaniswamy (2021) hariç diğerlerinden daha iyi ve Rashed ve diğerleri (2023) ile aynı sonuç elde edilmiştir. Bu da AutoML'in model seçim sürecini otomatikleştirmek ve optimize etmek için güçlü bir araç olma potansiyelinin altını çizmektedir. Tablo 6.1 ve Tablo 6.2'deki karşılaştırmada, kullanılan AutoML araçları arasından süre kısıtlaması olmadan TPOT, diğer algoritmalar arasından lojistik regresyon daha başarılı olmuştur.

Tablo 6.1: Klasik sınıflandırıcı karşılaştırması.

Sınıflandırıcı	Doğruluk	Duyarlılık	Hassasiyet	F1 Skor	Süre(sa:dk:sn)
Lojistik Regresyon	0.9801	0.9559	0.9907	0.9725	0:00:01.291
Yapay Sinir Ağı	0.9760	0.9605	0.9760	0.9673	0:00:13.713
Destek Vektör Makinesi	0.9754	0.9608	0.9745	0.9669	0:00:01.711
LightGBM	0.9672	0.9386	0.9740	0.9543	0:00:07.855
K-En Yakın Komşu	0.9672	0.9339	0.9781	0.9545	0:00:01.464
Stokastik Gradyan İniş	0.9666	0.9354	0.9748	0.9535	0:00:01.321
Adaboost	0.9655	0.9370	0.9713	0.9524	0:00:34.027
XGBoost	0.9631	0.9354	0.9656	0.9489	0:00:03.098
Rassal Orman	0.9614	0.9339	0.9624	0.9460	0:00:09.428
Naive Bayes	0.9379	0.8791	0.9532	0.9129	0:00:01.329
Karar Ağacı	0.9256	0.8680	0.9320	0.8955	0:00:01.665

Tablo 6.2: Süre kısıtlamasız AutoML karşılaştırması.

Sınıflandırıcı	Doğruluk	Duyarlılık	Hassasiyet	F1 Skor	Süre(sa:dk:sn)
Stokastik Gradyan İniş (TPOT)	0.9912	1.0000	0.9744	0.9870	2:17:06.287
Derin Öğrenme (H2O)	0.9871	0.9816	0.9929	0.9871	1:00:10.769
CatBoost (MLJAR)	0.9846	0.9718	0.9885	0.9801	1:55:55.091

Tablo 6.3'te AutoML yöntemlerinden önce farklı ön işlemeler uygulandığında süre kısıtlaması olmadan elde edilen sonuçlar bulunmaktadır. Aralarından ön işleme olarak sadece normalizasyon uygulandığında %98.25 doğrulukla TPOT ve H2O ile diğer ön işleme türlerine göre daha yüksek doğruluk elde edilmiştir. Ancak AutoML yöntemlerinden önce sadece SMOTE uygulanmasına kıyasla daha zayıf performans gözlenmektedir.

Tablo 6.3: Süre kısıtlamasız ön işlemeli AutoML karşılaştırması.

	Sınıflandırıcı	Doğruluk	Duyarlılık	Hassasiyet	F1 Skor	Süre(sa:dk:sn)
Normalizasyon	Topluluk (TPOT)	0.9825	0.9767	0.9767	0.9767	3:41:28.749
	Derin Öğrenme (H2O)	0.9825	1.0000	1.0000	0.9773	0:59:29.871
	CatBoost (MLJAR)	0.9737	0.9302	1.0000	0.9639	1:06:32.419
Standardizasyon Temel Bileşen Analizi	Çok Katmanlı Algılayıcı (TPOT)	0.9639	0.9535	0.9318	0.9425	2:46:20.439
	Derin Öğrenme (H2O)	0.9737	1.0000	1.0000	0.9647	1:17:08.778
	Topluluk (MLJAR)	0.9649	0.9535	0.9535	0.9535	1:06:12.101
Standardizasyon Temel Bileşen Analizi SMOTE	Topluluk (TPOT)	0.9649	0.9767	0.9333	0.9545	5:05:10.573
	Derin Öğrenme (H2O)	0.9474	1.0000	1.0000	0.9333	1:00:02.688
	Topluluk (MLJAR)	0.9737	0.9767	0.9545	0.9655	1:09:16.733

AutoML, meme kanseri tespiti için model seçme sürecini otomatikleştirmeye yardımcı olma konusunda büyük bir potansiyele sahiptir. Bu teknoloji doğru tahmin modelleri geliştirmenin karmaşık ve zaman alıcı kısmını kolaylaştırmaktadır. AutoML, model geliştirme için gereken süreyi ve kaynakları önemli ölçüde

azaltmaktadır ve makine öğrenmesini daha geniş bir kitle için erişilebilir hale getirmektedir. Kapsamlı makine öğrenmesi uzmanlığına sahip olmayan araştırmacılar ve uzmanlar doğru modeller geliştirmek üzere AutoML araçlarından yararlanabilirler. Bu çalışmada kullanılan yöntemlerin ve farklı AutoML araçlarının daha büyük ve çeşitli verisetlerinde genelleştirilebilirlik durumu gelecek çalışmalara bırakılmıştır.

Çalışmada karşılaşılan bir durum olan dengesiz verisetleriyle uğraşırken çapraz doğrulamadan önce veri artırma için SMOTE uygulandığında, veri sızıntısına neden olabilir ve aşırı iyimser model performansı tahminlerine yol açabilir. Örnek artırma verisetine çapraz doğrulamadan önce uygulanırsa bunun sonucunda bazı kopyalar hem eğitim hem de çapraz doğrulama katlarında bulunabilir. Bu nedenle model eğitim sırasında zaten görülmüş olan bazı örneklerle doğrulanır bu da doğrulama katının temel amacına ters düşer ve veri sızıntısı gerçekleşmiş olur. Bunun sonucunda doğrulama katlarında gerçek verileri temsil etmeyen yanlı performans ölçümleri gerçekleşir (Kuhn 2023). Bu sorunu önlemek için SMOTE, çapraz doğrulama sürecinin her bir katında uygulanmalı ve sentetik örneklerin yalnızca söz konusu kat için eğitim verileri kullanılarak oluşturulması sağlanmalıdır. Bu manuel olarak veya ardışık düzen kullanılarak gerçekleştirilebilir. Bu çalışmada verisetindeki sınıf dengesizliğinin model performansını olumsuz etkilememesi için azınlık veri artırma yöntemi SMOTE rassal az örnekleme ile birlikte kullanılmıştır. Azınlık sınıf örnekleri çoğunluk sınıf örneklerinin %95'ine eşit olacak şekilde artırılıp, çoğunluk sınıf örnekleri yeni azınlık sınıf örnek sayısına indirgenmiştir.

7. KAYNAKLAR

A. J. B, and Palaniswamy, S., "Comparison of Conventional and Automated Machine Learning approaches for Breast Cancer Prediction", *2021 Third International Conference on Inventive Research in Computing Applications (ICIRCA).* doi: 10.1109/icirca51532.2021.9544863, (2021)

Abbas, Z., and Rehman, S., "An Overview of Cancer Treatment Modalities", *Neoplasm.* doi: 10.5772/intechopen.76558, (2018)

Agarap, A. F. M., "On breast cancer detection", *Proceedings of the 2nd International Conference on Machine Learning and Soft Computing - ICMLSC '18.* doi: 10.1145/3184066.3184080, (2018)

Aghalarova, S., and Bozkurt Keser, S., "Öğrencilerin Akademik Performanslarının Tahmin Edilmesi için AutoML Tekniğinin Uygulanması", *El-Cezeri Fen ve Mühendislik Dergisi*, 9(2), 394–412. doi: 10.31202/ecjse.946505, (2021)

Agrapetidou, A., Charonyktakis, P., Gogas, P., Papadimitriou, T., and Tsamardinos, I., "An AutoML application to forecasting bank failures", *Applied Economics Letters*, 1–5. doi: 10.1080/13504851.2020.1725230, (2020)

Ahmed, U., Jerry Chun-Wei Lin, and Srivastava, G., "Multivariate time-series sensor vital sign forecasting of cardiovascular and chronic respiratory diseases", *Sustainable Computing: Informatics and Systems*, Elsevier BV, 38, 100868–100868. doi: 10.1016/j.suscom.2023.100868, (2023)

Aldrich, J. E., "Basic physics of ultrasound imaging", *Read Online: Critical Care Medicine | Society of Critical Care Medicine*, 35(5), S131. doi: 10.1097/01.CCM.0000260624.99430.22, (2007)

Anand, P., Kunnumakara, A. B., Sundaram, C., Harikumar, K. B., Tharakan, S. T., Lai, O. S., Sung, B., and Aggarwal, B. B., "Cancer is a Preventable Disease that

Requires Major Lifestyle Changes", *Pharmaceutical Research*, 25(9), 2097–2116. doi: 10.1007/s11095-008-9661-9, (2008)

Angarita-Zapata, J. S., Maestre-Góngora, G., and Calderín, J. F., "A Case Study of AutoML for Supervised Crash Severity Prediction", *Atlantis studies in uncertainty modelling*, Atlantis Press. doi: 10.2991/asum.k.210827.026, (2021)

Arnold, M., Pandeya, N., Byrnes, G., Renehan, A. G., Stevens, G. A., Ezzati, M., Ferlay, J., Miranda, J. J., Romieu, I., Dikshit, R., Forman, D., and Soerjomataram, I., "Global burden of cancer attributable to high body-mass index in 2012: a population-based study", *The Lancet Oncology*, 16(1), 36–46. doi: 10.1016/s1470-2045(14)71123-4, (2015)

Artasanchez, A., and Joshi, P., *Artificial intelligence with Python*, Packt, (2020)

Baykara, O., "Current therapies and latest developments in cancer treatment", *Horizons in Cancer Research*, 57, 105-156, (2015)

Bayo, J., Molina, R., Pérez, J., Pérez-Ruíz, E., Aparicio, J., Beato, C., Berros, J. P., Bolaños, M., Graña, B., and Santaballa, A., "SEOM clinical guidelines to primary prevention of cancer (2018)", *Clinical & Translational Oncology*, 21(1), 106–113. doi: 10.1007/s12094-018-02016-4, (2019)

Bayrak, E. A., Kirci, P., and Ensari, T., "Comparison of Machine Learning Methods for Breast Cancer Diagnosis", *2019 Scientific Meeting on Electrical-Electronics & Biomedical Engineering and Computer Science (EBBT).* doi: 10.1109/ebbt.2019.8741990, (2019)

Bharat, A., Pooja, N., and Reddy, R. A., "Using Machine Learning algorithms for breast cancer risk prediction and diagnosis", *2018 3rd International Conference on Circuits, Control, Communication and Computing (I4C).* doi: 10.1109/cimca.2018.8739696, (2018)

Breiman, L., "Bagging Predictors", *Machine Learning*, 24, 123–140. doi: 10.1023/a:1018054314350, (1996)

Breiman, L., “Random Forests”, *Machine Learning*, 45(1), 5–32. doi: 10.1023/a:1010933404324, (2001)

Calhoun, K. E., and Anderson, B. O., “Needle Biopsy for Breast Cancer Diagnosis: A Quality Metric for Breast Surgical Practice”, *Journal of Clinical Oncology*, 32(21), 2191–2192. doi: 10.1200/jco.2014.55.6324, (2014)

Chawla, N. V., Bowyer, K. W., Hall, L. O., and Kegelmeyer, W. P., “SMOTE: Synthetic Minority Over-sampling Technique”, *Journal of Artificial Intelligence Research*, 16, 321–357. doi: 10.1613/jair.953, (2002)

Chen, C., Liaw, A., Breiman, L., “Using Random Forest to Learn Imbalanced Data [online]”, (13 Nisan 2023), https://statistics.berkeley.edu/tech-reports/666, (2004)

Collaborative Group on Hormonal Factors in Breast Cancer., “Menarche, menopause, and Breast Cancer risk: Individual Participant meta-analysis, Including 118 964 Women with Breast Cancer from 117 Epidemiological Studies”, *The Lancet Oncology*, 13(11), 1141–1151. doi: 10.1016/s1470-2045(12)70425-4, (2012)

Congalton, R. G., Oderwald, R. G., and Mead, R. A., “Assessing Landsat classification accuracy using discrete multivariate analysis statistical techniques.”, *Photogrammetric Engineering and Remote Sensing, American Society for Photogrammetry and Remote Sensing*, 49(12), 1671–1678., (1983)

Cortes, C., and Vapnik, V., “Support-vector networks”, *Machine Learning*, 20(3), 273–297. doi: 10.1007/bf00994018, (1995)

Crosby, D., Bhatia, S., Brindle, K. M., Coussens, L. M., Dive, C., Emberton, M., Esener, S., Fitzgerald, R. C., Gambhir, S. S., Kuhn, P., Rebbeck, T. R., and Balasubramanian, S., “Early detection of cancer”, *Science*, 375(6586). doi: 10.1126/science.aay9040, (2022)

Darapureddy, N., Karatapu, N., and Battula, T. K., “Implementation of optimization algorithms on Wisconsin Breast cancer dataset using deep neural network”, *2019 4th International Conference on Recent Trends on Electronics, Information,*

Communication & Technology (RTEICT), 351–355. doi: 10.1109/rteict46194.2019.9016822, (2019)

Dietterich, T., "Machine learning", *ACM Computing Surveys*, 28(4es), 3-es. doi: 10.1145/242224.242229, (1996)

Divyavani, M., Kalpana, G., "An Analysis on SVM & ANN Using Breast Cancer Dataset", *Aegaeum J.*, 8(12), 0776-3808, (2020)

Dora, L., Agrawal, S., Panda, R., and Abraham, A., "Optimal breast cancer classification using Gauss–Newton representation based algorithm", *Expert Systems with Applications*, 85, 134–145. doi: 10.1016/j.eswa.2017.05.035, (2017)

Dossus, L., Boutron-Ruault, M.-C., Kaaks, R., Gram, I. T., Vilier, A., Fervers, B., Manjer, J., Tjonneland, A., Olsen, A., Overvad, K., Chang-Claude, J., Boeing, H., Steffen, A., Trichopoulou, A., Lagiou, P., Sarantopoulou, M., Palli, D., Berrino, F., Tumino, R., and Vineis, P., "Active and passive cigarette smoking and breast cancer risk: Results from the EPIC cohort", *International Journal of Cancer*, 134(8), 1871–1888. doi: 10.1002/ijc.28508, (2014)

Erickson, N., Mueller, J., Shirkov, A., Zhang, H., Larroy, P., Li, M., and Smola, A. J., "AutoGluon-Tabular: Robust and Accurate AutoML for Structured Data", (2 Mart 2023), doi: 10.48550/arxiv.2003.06505, (2020)

Ewertz, M., and Jensen, A. B., "Late effects of breast cancer treatment and potentials for rehabilitation", *Acta Oncologica*, 50(2), 187–193. doi: 10.3109/0284186x.2010.533190, (2011)

Fath, A. H., Pouranfard, A., and Foroughizadeh, P., "Development of an artificial neural network model for prediction of bubble point pressure of crude oils", *Petroleum*, 4(3), 281–291. doi: 10.1016/j.petlm.2018.03.009, (2018)

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

Fournier, A., Berrino, F., Riboli, E., Avenel, V., and Clavel-Chapelon, F., "Breast cancer risk in relation to different types of hormone replacement therapy in the E3N-EPIC cohort", *International Journal of Cancer*, 114(3), 448–454. doi: 10.1002/ijc.20710, (2005)

Freund, Y., and Schapire, R. E., "A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting", *Journal of Computer and System Sciences*, 55(1), 119–139. doi: 10.1006/jcss.1997.1504, (1997)

Gad, A. F., "Part 1: Image Classification using Features Extracted by Transfer Learning in Keras [online]", (26 Haziran 2023), https://www.alibabacloud.com/blog/part-1-image-classification-using-features-extracted-by-transfer-learning-in-keras_595289

Ganaie, M. A., Tanveer, M., Suganthan, P. N., and Snášel, V., "Oblique and rotation double random forest", *Neural Networks*, Elsevier BV, 153, 496–517. doi: 10.1016/j.neunet.2022.06.012, (2022)

Gavrilova, Y., "How to Choose a Machine Learning Technique [online]", (26 Haziran 2023), https://serokell.io/blog/how-to-choose-ml-technique

Ghiasi, M. M., and Zendehboudi, S., "Application of decision tree-based ensemble learning in the classification of breast cancer", *Computers in Biology and Medicine*, 128, 104089. doi: 10.1016/j.compbiomed.2020.104089, (2021)

Greenacre, M., Groenen, P. J. F., Hastie, T., D'Enza, A. I., Markos, A., and Tuzhilina, E., "Principal component analysis", Nature Reviews Methods Primers, 2(1), 1–21. doi: 10.1038/s43586-022-00184-w, (2022)

Greenhalgh, T. A., and Symonds, R. P., "Principles of chemotherapy and radiotherapy", *Obstetrics, Gynaecology & Reproductive Medicine*, 24(9), 259–265. doi: 10.1016/j.ogrm.2014.06.004, (2014)

Gupta, A., Tatbul, N., Marcus, R., Zhou, S., Lee, I., Gottschlich, J., "Class-weighted evaluation metrics for Imbalanced Data Classification", (11 Mart 2023) doi: 10.48550/arXiv.2010.05995, (2020)

Gupta, C., & Sharma, K., "Early Breast Cancer Detection using Various Machine Learning Techniques", *International Journal of Engineering Research & Technology (IJERT)*, 11(6). doi: 10.17577/IJERTV11IS060104, (2022)

Géron, A., *Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow*, O'Reilly Media, Inc., (2022)

H2O, "H2O AutoML: Automatic Machine Learning [online]", (6 Nisan 2023b), https://docs.h2o.ai/h2o/latest-stable/h2o-docs/automl.html

H2O, "Welcome to H2O 3 [online]", (6 Nisan 2023a), https://docs.h2o.ai/h2o/latest-stable/h2o-docs/welcome.html

Hambali, M., Saheed, Y., Oladele, T., Gbolagade, M., "ADABOOST Ensemble Algorithms for Breast Cancer Classification", *Journal of Advances in Computer Research*, 10(2), 31-52., (2019)

Hanahan, D. and Weinberg, R. A. "The Hallmarks of Cancer", *Cell*, 100(1), 57–70. doi: 10.1016/s0092-8674(00)81683-9, (2000)

Hooley, R. J., Scoutt, L. M., and Philpotts, L. E., "Breast Ultrasonography: State of the Art", *Radiology*, 268(3), 642–659. doi: 10.1148/radiol.13121606, (2013)

Iliyas, I. I., Dauda, A. B., Isa, A., Umoru, A., "Performance Analysis of Machine Learning Algorithms For Breast Cancer Detection", *Timbou-Africa Academic Publications*, 9(9), 2623-7861, (2022)

Ippolito, P. P., "Hyperparameter Tuning", Applied Data Science in Tourism: Interdisciplinary Approaches, Methodologies and Applications, 231–251. doi: 10.1007/978-3-030-88389-8_12, (2022)

Jain, A., Khalid, M., Qureshi, M. M., Georgian-Smith, D., Kaplan, J. A., Buch, K., Grinstaff, M. W., Hirsch, A. E., Hines, N. L., Anderson, S. W., Gallagher, K. M., Bates, D. D. B., and Bloch, B. N., "Stereotactic core needle breast biopsy marker migration: An analysis of factors contributing to immediate marker migration", *European Radiology*, 27(11), 4797–4803. doi: 10.1007/s00330-017-4851-7, (2017)

James, G., Witten, D., Hastie, T., and Tibshirani, R., *An Introduction to Statistical Learning: with Applications in R*, Springer, (2013)

Jin, H., "Hyperparameter Importance for Machine Learning Algorithms", (20 Mart 2023) doi: 10.48550/arXiv.2201.05132, (2022)

Jin, L., Huang, Z., Chen, L., Liu, M., Li, Y., Chou, Y., and Yi, C., "Modified single-output Chebyshev-polynomial feedforward neural network aided with subset method for classification of breast cancer", *Neurocomputing*, 350, 128–135. doi: 10.1016/j.neucom.2019.03.046, (2019)

John, S., Broggio, J., "Cancer survival in England: adult, stage at diagnosis and childhood patients followed up to 2018", *Office for National Statistics,* (2019)

Kamińska, M., Ciszewski, T., Łopacka-Szatan, K., Miotła, P., and Starosławska, E., "Breast cancer risk factors", *Menopausal Review*, 14(3), 196–202. doi: 10.5114/pm.2015.54346, (2015)

Kayl, A. E., and Meyers, C. A., "Side-effects of chemotherapy and quality of life in ovarian and breast cancer patients", *Current Opinion in Obstetrics and Gynecology*, 18(1), 24–28. doi: 10.1097/01.gco.0000192996.20040.24, (2006)

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., and Liu, T-Y., "Lightgbm: A highly efficient gradient boosting decision tree", *In Advances in Neural Information Processing Systems*, 3149–3157., (2017)

Kibria, H. B., and Matin, A., "The Severity Prediction of The Binary And Multi-Class Cardiovascular Disease -- A Machine Learning-Based Fusion Approach", (18 Nisan 2023) doi: 10.48550/arxiv.2203.04921, (2022)

Kozan, M., "Supervised and Unsupervised Learning (an Intuitive Approach) [online]", (26 Haziran 2023), https://medium.com/@metehankozan/supervised-and-unsupervised-learning-an-intuitive-approach-cd8f8f64b644, (2021)

Kuhn, M., Johnson, K., *Applied Predictive Modeling*, Springer, (2013)

Kuhn, M., “Subsampling For Class Imbalances [online]”, (24 Mart 2023), https://topepo.github.io/caret/subsampling-for-class-imbalances.html

Kulkarni, G. N., Ambesange, S., A, V., and Sahoo, A., “Comparision of diabetic prediction AutoML model with customized model”, doi: 10.1109/icais50930.2021.9395775, (2021)

Kumar, V., Mishra, B. K., Mazzara, M., Thanh, D. N. H., and Verma, A., “Prediction of Malignant and Benign Breast Cancer: A Data Mining Approach in Healthcare Applications”, *Advances in Data Science and Management*, 435–442. doi: 10.1007/978-981-15-0978-0_43, (2020)

Kumari, M., and Singh, V., “Breast Cancer Prediction system”, *Procedia Computer Science*, 132, 371–376. doi: 10.1016/j.procs.2018.05.197, (2018)

Kwon, H., Park, J., and Lee, Y., “Stacking Ensemble Technique for Classifying Breast Cancer”, *Healthcare Informatics Research*, 25(4), 283. doi: 10.4258/hir.2019.25.4.283 (2019)

Lehman, C. D., and Schnall, M. D., “Imaging in breast cancer: Magnetic resonance imaging”, *Breast Cancer Research*, 7(5). doi: 10.1186/bcr1309, (2005)

LightGBM, “Features [online]”, (15 Haziran 2023b), https://lightgbm.readthedocs.io/en/latest/Features.html

LightGBM, “Light Gradient Boosting Machine [online]”, (15 Haziran 2023a), https://github.com/microsoft/LightGBM

Lynch, H. T., Shaw, T. G., and Lynch, J. F., “Inherited predisposition to cancer: A historical overview”, *American Journal of Medical Genetics*, 129C(1), 5–22. doi: 10.1002/ajmg.c.30026, (2004)

MLJAR, “MLJAR Automated Machine Learning for Humans [online]”, (6 Nisan 2023), https://github.com/mljar/mljar-supervised

Madni, H. A., Umer, M., Ishaq, A., Abuzinadah, N., Saidani, O., Alsubai, S., Hamdi, M., and Ashraf, I., “Water-Quality Prediction Based on H2O AutoML and Explainable AI Techniques”, *Water*, 15(3), 475. doi: 10.3390/w15030475, (2023)

Maimon, O., and Rokach, L., “Data Mining and Knowledge Discovery Handbook”, Springer, New York, (2010)

Majeed, W., Aslam, B., Javed, I., Khaliq, T., Muhammad, F., Ali, A., and Raza, A., “Breast Cancer: Major Risk Factors and Recent Developments in Treatment”, *Asian Pacific Journal of Cancer Prevention*, 15(8), 3353–3358. doi: 10.7314/apjcp.2014.15.8.3353, (2014)

Mammone, A., Turchi, M., and Cristianini, N., “Support vector machines”, *Wiley Interdisciplinary Reviews: Computational Statistics*, 1(3), 283–289. doi: 10.1002/wics.49, (2009)

Mangukiya, M., “Breast Cancer Detection with Machine Learning”, *International Journal for Research in Applied Science and Engineering Technology*, 10(2), 141–145. doi: 10.22214/ijraset.2022.40204, (2022)

McCarthy, J., "What is artificial intelligence? [online]", (5 Nisan 2023), jmc.stanford.edu/artificial-intelligence/what-is-ai/index.html

McCarthy, J., Minsky, M. L., Rochester, N., and Shannon, C. E., “A Proposal for the Dartmouth Summer Research Project on Artificial Intelligence”, (1955)

McDonald, E. S., Clark, A. S., Tchou, J., Zhang, P., and Freedman, G. M., “Clinical Diagnosis and Management of Breast Cancer”, *Journal of Nuclear Medicine*, 57(Supplement_1), 9S16S. doi: 10.2967/jnumed.115.157834, (2016)

Mikulski, B., “PCA—how to choose the number of components? [online]”, (29 Haziran 2023), mikulskibartosz.name/pca-how-to-choose-the-number-of-components, (3 Haziran 2019)

Mishra, S. P., Sarkar, U., Taraphder, S., Datta, S., Swain, D. P., Saikhom, R., Panda, S., and Laishram, M., “Principal Component Analysis”, International Journal of Livestock Research, 7(5), 1. doi: 10.5455/ijlr.20170415115235, (2017)

Mitchell, T. M., *Machine Learning*, Mcgraw-Hill Education, New York, (1997)

Mohsen, F., Biswas, Md. R., Ali, H., Alam, T., Househ, M., and Shah, Z., "Customized and Automated Machine Learning-Based Models for Diabetes Type 2 Classification", Studies in health technology and informatics. doi: 10.3233/shti220779, (2022)

Mushtaq, Z., Yaqub, A., Sani, S., and Khalid, A., "Effective K-nearest neighbor classifications for Wisconsin breast cancer data sets", *Journal of the Chinese Institute of Engineers*, 43(1), 80–92. doi: 10.1080/02533839.2019.1676658, (2019)

Naji, M. A., Filali, S. E., Aarika, K., Benlahmar, E. H., Abdelouhahid, R. A., and Debauche, O., "Machine Learning Algorithms For Breast Cancer Prediction And Diagnosis", *Procedia Computer Science*, 191, 487–492. doi: 10.1016/j.procs.2021.07.062, (2021)

Nasser, F. K., and Behadili, S. F., "Breast Cancer Detection using Decision Tree and K-Nearest Neighbour Classifiers", *Iraqi Journal of Science*, 4987–5003. doi: 10.24996/ijs.2022.63.11.34, (2022)

National Cancer Institute. "What Is Cancer? [online]", (10 Mart 2023), National Cancer Institute, Cancer.gov, https://www.cancer.gov/about-cancer/understanding/what-is-cancer, (2021)

Nguyen, Q. T., and Tsien, R. Y., "Fluorescence-guided surgery with live molecular navigation — a new cutting edge", *Nature Reviews Cancer*, 13(9), 653–662. doi: 10.1038/nrc3566, (2013)

Niwas, S. I., Palanisamy, P., Sujathan, K., and Bengtsson, E., "Analysis of nuclei textures of fine needle aspirated cytology images for breast cancer diagnosis using Complex Daubechies wavelets", *Signal Processing*, 93(10), 2828–2837. doi: 10.1016/j.sigpro.2012.06.029, (2013)

Obaid, O. I., Mohammed, M. A., Ghani, M. K. A., Mostafa, S., and Al-Dhief, F. T., "Evaluating the Performance of Machine Learning Techniques in the Classification

of Wisconsin Breast Cancer", *International Journal of Engineering and Technology.* 7. 160-166. doi: 10.14419/ijet.v7i4.36.23737, (2018)

Olson, R. S., and Moore, J. H., "TPOT: A Tree-Based Pipeline Optimization Tool for Automating Machine Learning", *Automated Machine Learning*, 151–160. doi: 10.1007/978-3-030-05318-5_8, (2019)

Orlenko, A., Kofink, D., Lyytikäinen, L.-P., Nikus, K., Mishra, P., Kuukasjärvi, P., Karhunen, P. J., Kähönen, M., Laurikka, J. O., Lehtimäki, T., Asselbergs, F. W., and Moore, J. H., "Model selection for metabolomics: predicting diagnosis of coronary artery disease using automated machine learning", *Bioinformatics*, (J. Kelso., ed.), 36(6), 1772–1778. doi: 10.1093/bioinformatics/btz796, (2019)

Pais, R. J., Lopes, F., Parreira, I., Silva, M., Silva, M., and Moutinho, M. G., "Predicting Cancer Prognostics from Tumour Transcriptomics Using an Auto Machine Learning Approach", *Medical Sciences Forum*, 22(1):6 doi: 10.3390/msf2023022006, (2023)

Pandey, A., and Jain, A. "Comparative Analysis of KNN Algorithm using Various Normalization Techniques", International Journal of Computer Network and Information Security, 9(11), 36–42. doi: 10.5815/ijcnis.2017.11.04, (2017)

Panje, C. M., Glatzer, M., Sirén, C., Plasswilm, L., and Putora, P. M., "Treatment Options in Oncology", *JCO Clinical Cancer Informatics*, (2), 1–10. doi: 10.1200/cci.18.00017, (2018)

Park, H.-L., and Hong, J., "Vacuum-assisted breast biopsy for breast cancer", *Gland Surgery*, 3(2), 120–127. doi: 10.3978/j.issn.2227-684X.2014.02.03, (2014)

Park, H.-L., and Kim, L. S., "The Current Role of Vacuum Assisted Breast Biopsy System in Breast Disease", *Journal of Breast Cancer*, 14(1), 1. doi: 10.4048/jbc.2011.14.1.1, (2011)

Pearson, K., "On Lines and Planes of Closest Fit to Systems of Points in Space", The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science, 2(11), 559–572. doi: 10.1080/14786440109462720, (1901)

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, É., "Scikit-learn: Machine Learning in Python", *J. Mach. Learn. Res.*, 12, 2825–2830. doi: 10.5555/1953048.2078195, (2011)

Putrada, A. G., Laeli, E. K., Pane, S. F., Alamsyah, N., and Fauzan, M. N., "TPOT on Increasing the Performance of Credit Card Application Approval Classification", 2022 2nd International Conference on Intelligent Cybernetics Technology & Applications (ICICyTA). doi: 10.1109/icicyta57421.2022.10038063, (2022)

Radzi, S. F. M., Karim, M. K. A., Saripan, M. I., Rahman, M. A. A., Isa, I. N. C., and Ibahim, M. J., "Hyperparameter Tuning and Pipeline Optimization via Grid Search Method and Tree-Based AutoML in Breast Cancer Prediction", *Journal of Personalized Medicine*, 11(10), 978. doi: 10.3390/jpm11100978, (2021)

Rashed, A. E. E., Elmorsy, A. M., and Atwa, A. E. M., "Comparative evaluation of automated machine learning techniques for breast cancer diagnosis", *Biomedical Signal Processing and Control*, 86, 105016. doi: 10.1016/j.bspc.2023.105016, (2023)

Refaeilzadeh, P., Tang, L., and Liu, H., "Cross-Validation", *Encyclopedia of Database Systems*, 532–538. doi: 10.1007/978-0-387-39940-9_565, (2009)

Rojas, K., and Stuckey, A., "Breast Cancer Epidemiology and Risk Factors", *Clinical Obstetrics and Gynecology*, 59(4), 651–672. doi: 10.1097/GRF.0000000000000239, (2016)

Russel, S., and Norvig, P., *Artificial Intelligence: A Modern Approach*, Pearson, (2020)

Sahu, B., Mohanty, S., and Rout, S., "A Hybrid Approach for Breast Cancer Classification and Diagnosis", *ICST Transactions on Scalable Information Systems*, doi: 10.4108/eai.19-12-2018.156086, (2018)

Saini, A., "Guide on Support Vector Machine (SVM) Algorithm [online]", (11 Nisan 2023), https://www.analyticsvidhya.com/blog/2021/10/support-vector-machinessvm-a-complete-guide-for-beginners/, (2021)

Sardanelli, F., Fallenberg, E. M., Clauser, P., Trimboli, R. M., Camps-Herrero, J., Helbich, T. H., and Forrai, G., "Mammography: an update of the EUSOBI recommendations on information for women", *Insights into Imaging*, 8(1), 11–18. doi: 10.1007/s13244-016-0531-4, (2016)

Seely, J. M., and Alhassan, T., "Screening for breast cancer in 2018— what should we be doing today?", *Current Oncology*, 25(1), 115. doi: 10.3747/co.25.3770, (2018)

Seufert, E. B., "Quantitative Methods for Product Management", *Freemium Economics*, 47–82. doi: 10.1016/b978-0-12-416690-5.00003-8, (2014)

Shamrat, F. J. M., Raihan, M. A., Rahman, A. S., Mahmud, I., and Akter, R., "An Analysis on Breast Disease Prediction Using Machine Learning Approaches", *International Journal of Scientific & Technology Research*, 9(2), 2450-2455, (2020)

Sneha, N., and Gangil, T., "Analysis of diabetes mellitus for early prediction using optimal features selection", *Journal of Big Data*, 6(1). doi: 10.1186/s40537-019-0175-6, (2019)

Song, M., Vogelstein, B., Giovannucci, E. L., Willett, W. C., and Tomasetti, C., "Cancer prevention: Molecular and epidemiologic consensus", *Science*, 361(6409), 1317–1318. doi: 10.1126/science.aau3830, (2018)

Spencer, M. L., "Stem cells transplants in cancer treatment.", (2021)

Stazio, A., Victores, J. G., Estévez, D., and Balaguer, C., "A Study on Machine Vision Techniques for the Inspection of Health Personnels' Protective Suits for the Treatment of Patients in Extreme Isolation", *Electronics, Multidisciplinary Digital Publishing Institute*, 8(7), 743–743. doi: 10.3390/electronics8070743, (2019)

Stuckey, D. W., and Shah, K., "Stem cell-based therapies for cancer treatment: separating hope from hype", *Nature Reviews Cancer*, 14(10), 683–691. doi: 10.1038/nrc3798, (2014)

Sun, Y.-S., Zhao, Z., Yang, Z.-N., Xu, F., Lu, H.-J., Zhu, Z.-Y., Shi, W., Jiang, J., Yao, P.-P., and Zhu, H.-P., "Risk Factors and Preventions of Breast Cancer", *International Journal of Biological Sciences*, 13(11), 1387–1397. doi: 10.7150/ijbs.21635, (2017)

Sung, H., Ferlay, J., Siegel, R. L., Laversanne, M., Soerjomataram, I., Jemal, A., and Bray, F., "Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries", *CA: A Cancer Journal for Clinicians*, 71(3), 209–249. doi: 10.3322/caac.21660, (2021)

Sutton, R. S., and Barto, A. G., *Reinforcement Learning: An Introduction*, The Mit Press, (2018)

TPOT, "An example Machine Learning pipeline [online]", (6 Nisan 2023), https://github.com/epistasislab/tpot, (2022)

TPOT, "An example TPOT pipeline [online]", (6 Nisan 2023), https://github.com/epistasislab/tpot, (2022)

TPOT, "TPOT [online]", (6 Nisan 2023), https://automl.info/tpot

Thomas, T., Vijayaraghavan, A. P., and Emmanuel, S., Machine Learning Approaches in Cyber Security Analytics, Springer Singapore, Singapore. Doi: 10.1007/978-981-15-1706-8, (2020)

UCI Machine Learning Repository, "Breast Cancer Wisconsin (Diagnostic) [online]", (14 Mart 2023), https://archive.ics.uci.edu/dataset/17/breast+cancer+wisconsin+diagnostic

Wang, L., Wang, T., and Hu, X., "Logistic Regression Region Weighting for Weakly Supervised Object Localization", *IEEE Access*, 1–1. doi: 10.1109/access.2019.2935011, (2019)

Wang, X., Yan, L., and Zhang, Q., "Research on the Application of Gradient Descent Algorithm in Machine Learning", *IEEE Xplore.* doi: 10.1109/ICCNEA53019.2021.00014, (2021)

Williams, C. K. O., “Risk Factors for Cancer”, *Cancer and AIDS*, 115–178. doi: 10.1007/978-3-319-99235-8_5, (2018)

Woods, D., and Turchi, J. J., “Chemotherapy induced DNA damage response”, *Cancer Biology & Therapy*, 14(5), 379–389. doi: 10.4161/cbt.23761, (2013)

Yahya, E. B., and Alqadhi, A. M., “Recent trends in cancer therapy: A review on the current state of gene delivery”, *Life Sciences*, 269, 119087. doi: 10.1016/j.lfs.2021.119087, (2021)

Yu, F. R., and He, Y., “Reinforcement Learning and Deep Reinforcement Learning”, *Deep Reinforcement Learning for Wireless Networks*, 15–19. doi: 10.1007/978-3-030-10546-4_2, (2019)

Yunus, M. M., Yusof, A. K. M., Rahman, M. Z. A., Koh, X. J., Sabarudin, A., Nohuddin, P. N. E., Ng, K. H., Kechik, M. M. A., and Karim, M. K. A., “Automated Classification of Atherosclerotic Radiomics Features in Coronary Computed Tomography Angiography (CCTA)”, *Diagnostics*, 12(7), 1660. doi: 10.3390/diagnostics12071660, (2022)

Zaman, M., Kaul, S., and Ahmed, M., “Analytical Comparison Between the Information Gain and Gini Index using Historical Geographical Data”, *International Journal of Advanced Computer Science and Applications*, 11(5). doi: 10.14569/ijacsa.2020.0110557, (2020)

Zgajnar, J., “Clinical Presentation, Diagnosis and Staging of Breast Cancer”, *Breast Cancer Management for Surgeons*, 159–176. doi: 10.1007/978-3-319-56673-3_14, (2017)

Zhang, W., Xu, C., Li, R., Cui, G., Wang, M., and Wang, M., “Correlation analysis between ultrasonography and mammography with other risk factors related to breast cancer”, *Oncology Letters*. doi: 10.3892/ol.2019.10246, (2019)